JSEALS Special Publication No. 3

PAPERS FROM THE SEVENTH INTERNATIONAL CONFERENCE ON AUSTROASIATIC LINGUISTICS

Edited by:

Hiram Ring
Felix Rau

UNIVERSITY of HAWAI'I PRESS

OPEN ACCESS – Semiannual with periodic special publications
E-ISSN: 1836-6821
http://hdl.handle.net/10524/52438

Cover photo courtesy of Hiram Ring: Pnar speakers planting rice near Sohmynting, Meghalaya, North-East India.

JSEALS
Journal of the Southeast Asian Linguistics Society

The Journal of the Southeast Asian Linguistics Society publishes articles on a wide range of linguistic topics of the languages and language families of Southeast Asia and surrounding areas. JSEALS has been hosted by the UH Press since the beginning of 2017.

Contents

INTRODUCTION TO SPECIAL ISSUE: PAPERS FROM ICAAL 7

This special issue of the Journal of South-East Asian linguistics consists of a selection of papers from the 7th International Conference on Austro-Asiatic Linguistics, held in Kiel, Germany between September 29 - October 1, 2017 at the Christian Albrechts University (CAU). The conferences are held every two years and provide an opportunity for scholars working on Austroasiatic languages to present and discuss their work. At the business meeting a proceedings volume was proposed - these nine papers are the result of the proposal.

Austroasiatic languages are relatively diverse typologically and are located non-contiguously over a large geographical area stretching from eastern India to Vietnam. The relationship of the typological diversity with the geographical spread of the languages is a central issue that continued to arise at the conference. The degree of difference between various groups of languages (particularly the traditional distinction between "Munda" and "Mon-Khmer") and exactly what motivates such differences has been much debated, and a number of papers in this volume address these questions, particularly in relation to the proposed history and spread of the languages.

The first paper in the volume tackles the differences between Austroasiatic languages in terms of prosody. In their paper "On prosodic structures in Austroasiatic diachrony: 'Rhythmic holism' revisited in light of preliminary acoustic studies", Hiram Ring and Gregory D. S. Anderson provide a timely critique of the widely cited work of Donegan and Stampe (references in paper). They review some of the current cross-linguistic literature on prosody, as well as studies of Austroasiatic languages, and conduct a pilot acoustic analysis of words and phrases in Sora, Pnar, and Lawa. They suggest that claims of a single rhythmic organizing principle at the prosodic level accounting for the differences between the Munda languages and other Austroasiatic languages are difficult to maintain, and that the three languages investigated seem to share the same iambic structure at the word level. They also highlight how reference to and sharing of data is crucial to make progress in disentangling the historical relationships and development of these languages.

The second paper describes grammatical structures in Munda languages. The paper, by Gregory D. S. Anderson and Bikram Jora titled "Negation, TAM and person-indexing interdependencies in the Munda languages: a preliminary report", offers a careful analysis of interacting Munda grammatical systems, namely negation, tense/aspect/mood, and person-indexing. They suggest that alignment of various elements of these systems may allow for reconstruction, and refer to a database of transcriptions and translations, providing a large number of examples to back up their claims.

The third paper, by Jurica Polančec, also has Munda languages as its focus. Titled "Correlative-Relative Clauses In Munda Languages: An Overview", it highlights how Munda languages have both headed and headless Correlative-Relative Clauses, and that while the former are likely borrowed from neighboring Indo-Aryan languages, the latter are likely original to Munda. Evidence provided comes from neighboring languages and an appeal to cross-linguistic tendencies.

The fourth and fifth paper in this issue move east from South Asia to the Palaungic languages in eastern Myanmar. With "A phonological analysis of Riang Lang" Ellie Hall adds new data to the discussion of phonemes in Riang Lang, a Palaungic language located in Shan State, Myanmar. Her analysis indicates that the language has 12 vowels and 21 consonants, which differs slightly from previous analyses.

Rachel Weymuth's paper "Verbal affixes in Rumai, Palaung" provides an initial account of verb morphology in another Palaungic language, Rumai, spoken in northern Shan State of Myanmar and in neighboring Yunnan, China. She finds that the affixes that can be grouped into aspectual, modal, and polarity domains, as well as a single reciprocal marker. For some of the morphemes a source can be identified, while for others it cannot

The next two papers shift the focus slightly from descriptive accounts to historical. In "Proto-Nicobarese phonology" Paul Sidwell gives a reconstruction of the parent of the Nicobarese languages, notable for being the only Austroasiatic languages currently located on islands. While relatively little data exists for these languages, he scours what sources exist in order to present initial results of his ongoing reconstruction, providing an appendix of forms and links to an online dataset.

The seventh paper, “Katuic presyllables and derivational morphology in diachronic perspective” by Ryan Gehrmann is also a reconstruction, but of a different type. In this paper evidence is shown for the existence of presyllables, affixes, and morphological processes in Proto-Katuic. Each of these elements are carefully reconstructed based on data from the modern Katuic languages located in southern Laos, central Vietnam, northeastern Thailand and north-central Cambodia.

The eighth paper in this issue, “The Integration of French loanwords into Vietnamese: A corpus-based analysis of tonal, syllabic and segmental aspects” by Vera Scholvin and Judith Meinschaefer, deals with lexical borrowing in the largest Austroasiatic language, Vietnamese. Through analysis of a corpus of data made available online, the authors identify how words from French have been borrowed into the phonological system of Vietnamese with no influence from French phonology.

The ninth and final paper is somewhat more speculative. In “Waterworld: Lexical evidence for aquatic subsistence strategies in Austroasiatic”, Roger Blench gives an anthropological perspective on the groups that speak Austroasiatic languages, highlighting the diverse nature of cultural and linguistic overlaps in the region. We are reminded of the difficulties inherent in separating out older core vocabulary from borrowings at multiple historical strata and linking this with the dating of cultural patterns, particular flora/fauna, and relics, a problem that has plagued research in the area for centuries. He uses existing, publicly accessible databases of lexical data from different languages and families in the South-East Asian area to suggest potential subsistence strategies and movement patterns of Austroasiatic people groups, along with possible fauna of the areas they inhabited.

This introduction would not be complete without acknowledging the work that went into it. As all editors know, receiving and managing multiple papers, interfacing with authors and reviewers, is not an easy task, though it has its rewards. We commend each of the authors for their timely submission and thank them for the rapid revisions that have allowed for a relatively quick publication of this special issue. Each of the papers were reviewed by two separate anonymous reviewers, and we wish to thank each of these reviewers for their insightful critiques. Throughout the process we received invaluable assistance and advice from Mark Alves, the JSEALS managing editor. Other important sources of advice were Paul Sidwell and Mathias Jenny.

In conclusion, we feel it is important to note that this is the first published ICAAL proceedings since the ICAAL 4 proceedings were published in 2011. The ICAAL 4 publication, in turn, was preceded only by an Oceanic Linguistics special publication in 1976, which makes the current issue only the third published ICAAL proceedings since the inception of the conference in 1973 at the University of Hawai’i. With this special issue we return full circle to publication under the University of Hawai’i Press, and are extremely excited to be part of a new wave of Austroasiatic studies. There is much work yet to be done on these languages with all their diversity and complexity, but given the multiple perspectives and insights represented by the authors in this volume, and the increasing focus by AA researchers on making underlying data accessible, the outlook for AA studies in the coming century is incredibly positive.

Hiram Ring
Zurich, Switzerland

Felix Rau
Cologne, Germany

The Editors

HIRAM RING (PhD) is a Postdoctoral Research Fellow in the Department of Comparative Linguistics at the University of Zürich. He received his PhD from Nanyang Technological University, Singapore in 2015. He does data-driven research in phonetics, phonology, morphology, and syntax, and is interested in understanding grammaticalization and language change on multiple levels, particularly to shed light on how the Austroasiatic languages and peoples have reached their present geographically and linguistically diverse state. He has conducted fieldwork on Khasian languages in North-East India and Palaungic languages in Myanmar and Thailand.

FELIX RAU (MA) works at in the Department of Linguistics at the University of Cologne and manages the Language Archive Cologne at the Data Center for the Humanities. He has conducted extensive fieldwork on Gorum and has been working in the Koraput Area of Odisha (India) since 2002. He has worked on Gorum, Santali, and on historical phonology and historical morphology of the Munda languages.

FROM THE JSEALS EDITOR-IN-CHIEF

This is the third JSEALS special publication since JSEALS became a University of Hawai'i Press publication in January 2017. The goal of JSEALS special publications is to share collections of linguistics articles, such as select papers from conferences or other special research agendas, as well as to offer a way for linguistic researchers in the greater Southeast Asian region to publish monograph-length works.

This collection of nine articles were developed from papers given at ICAAL 7 (the Seventh International Conference on Austroasiatic Linguistics) by top researchers in Austroasiatic linguistics. Some of these papers constitute significant advances in the field by making new claims in individual language groups (e.g. Nicobarese and Katuic) and by reevaluating previous theories of the history the entire language family. Other papers advance understanding of aspects of modern languages (e.g. Palaung) and branches of Austroasiatic (e.g. Munda). Thus, the tradition of the ICAAL continues to make significant contributions in linguistics and Austroasiatic language history 45 years after its first conference in 1973.

We are proud to be able to publish studies of such high caliber, we are are very pleased that JSEALS is able to contribute to the sharing of quality linguistic research in both Austroasiatic linguistics and the Greater Southeast Asia region.

Mark J. Alves

December 1st, 2018

Rockville, Maryland

ON PROSODIC STRUCTURES IN AUSTROASIATIC DIACHRONY: 'RHYTHMIC HOLISM' REVISITED IN LIGHT OF PRELIMINARY ACOUSTIC STUDIES

Hiram Ring and Gregory D. S. Anderson
University of Zurich, Living Tongues Institute for Endangered Languages/University of South Africa (UNISA)
hiram.ring@uzh.ch, livingtongues@gmail.com

Abstract
This paper revisits claims regarding the division between Mon-Khmer and Munda languages on prosodic grounds (Donegan and Stampe 1983, 2002, 2004; Donegan 1993). Specifically, we attempt to re-evaluate their claims by investigating pitch at the level of the word in three languages from different families within the Austroasiatic phylum. First, we critique Donegan and Stampe's work, presenting data on Sora and other Munda languages showing a similar prosodic pattern across the whole family that does not conform to claims of a rhythmic holistic shift in prosody to the degree previously suggested. Second, we present a pilot acoustic study of Sora phrasal prosody in comparison with prosodic structures in both Pnar, a language in the Khasian group (the Munda languages' geographically nearest relatives), and prosody in Lawa, a Palaungic language. We find that Khasian word/phrase prosodic structures are quite similar to those found in many Munda languages, which has interesting implications for our understanding of the development of Austroasiatic languages.[1]

Keywords: Austroasiatic; word; phrase; prosody; pitch; Sora; Pnar; Lawa; Khasi

ISO 639-3 codes: srb, pbv, lwl, kha[2]

1 Introduction

Among the concerns of linguists working on Austroasiatic (AA) languages has been how to reconcile the typological differences between "(a) the largely co-ordinating (*sic*) and analytic Khmer-Nicobar languages, and (b) the largely subordinating and synthetic Munda languages." (Pinnow 1963:145) In particular, as Jenny, Sidwell, and Alves (Forthcoming) note, "Claims have been made that the major change that occurred in Munda languages was from rising to falling intonation patterns (Donegan and Stampe 2004), and that all other changes naturally followed from this." Despite indications that there are many different intonation contours throughout Munda languages (cf. Anderson 2015, Forthcoming), there are few investigations on

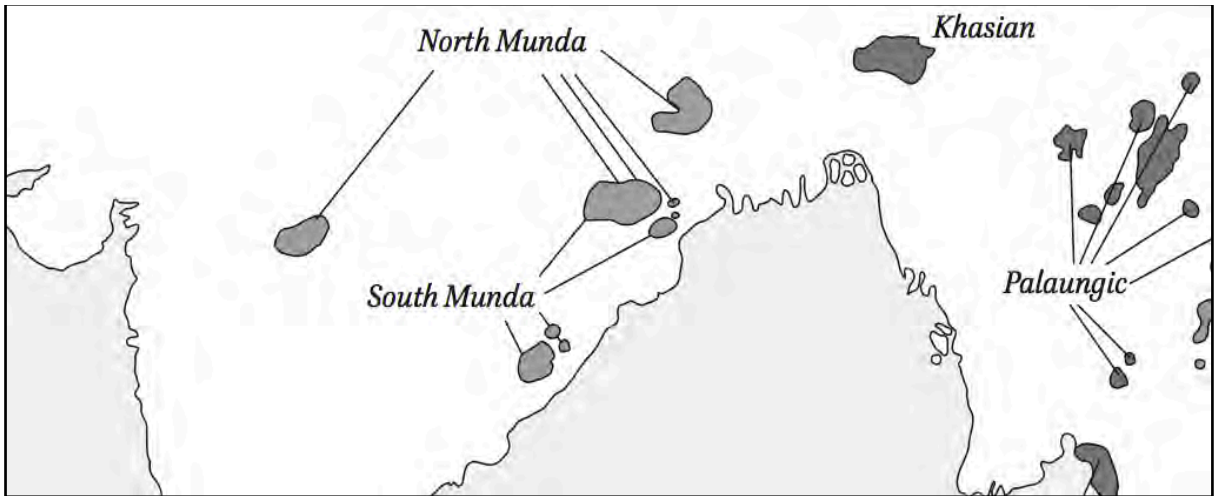
[1] We wish to acknowledge the following speakers, consultants, and language experts who have assisted us in gathering and analysing the data: Opino Gomango and Srinivas Gomango (Sora), Fidell War (Pnar), and Greg Blok and Ta Saai (Lawa). We thank attendees of ICAAL 7 for their comments, and two anonymous reviewers for their insightful critiques which helped greatly to focus this paper. We also acknowledge support from the Swiss National Science Foundation and a Singapore Ministry of Education grant MOE2012-1-100 (for work on Khasian and Palaungic), National Science Foundation Awards 1500092 (for work on Gutob) and 0853877 (for work on Remo), National Endowment for the Humanities Award PD-50025-13 (for work on Gta'), a Genographic Legacy Fund award (for work on Ho), several National Geographic Society awards (for work on Sora, Santali, Mundari, Kera', Gorum, Juray, Korku, Nihali, Juang and Kharia) and an award from the Zegar Family Foundation (for work on Birhor). Also thanks to Dr. Bikram Jora and Dr. Luke Horo for valuable assistance in data recording sessions during fieldwork and fruitful discussion thereon.

[2] Abbreviations used for examples are: ACT Actor, AUX Auxiliary, CAUS Causative, DES Desiderative, IPFV Imperfective, NEG Negative, NPST Non-past, PRON Pronominal, PST Past, RDPL Reduplication, RECIP Reciprocal, TAM Tense-Aspect-Mood, UND Undergoer, 1PL First person plural, 1SG First person singular, 2SG Second person singular, 3 Third person.

AA languages that would clarify prosodic differences and similarities between the Munda languages and their eastern cousins.

In particular, the existing literature on prosody in AA languages and its relation to the historical development of AA is limited to four papers (Donegan and Stampe 1983, 2002, 2004; Donegan 1993) that have been widely cited by scholars but which have various problems of evidence and methodology. In order to suggest a clear direction for prosodic research in AA and put the field on firmer scientific footing, we critique the previous research and present data on Munda that reveals similar stress and pitch patterns across the whole family, hard to square against claims of a rhythmic holistic shift in prosody (Donegan 1993). We also conduct an initial instrumental investigation of pitch in three languages of the AA phylum, submitting this data and our analysis in order to contribute to discussions about the development of AA languages.

Figure 1: *Munda, Khasian and Palaungic languages (adapted from Jenny and Sidwell 2015)*



The languages we investigate in our pilot study are Sora (Munda), Pnar (Khasian), and Lawa (Palaungic, Waic), all of which are reported to have no contrastive lexical tone. The language families that these individual languages belong to are geographically located in non-contiguous areas, but on a relatively straightforward West-East plane (Figure 1) stretching from eastern India to northwest Thailand. Munda languages are located in and around Orissa in Eastern India, while the Khasian group are in North-East India north of Bangladesh, and the Palaungic languages are found in North-Eastern Myanmar and North-Western Thailand. This makes the Khasian languages the geographically nearest AA relatives of Munda. Scholars have for a long time considered the Khasian and Palaungic families (and somewhat more distantly the Munda languages) to be closely related within the AA tree based on lexical similarity (see Diffloth 2005; Sidwell 2015), and thus these languages provide an interesting comparison, likely sharing a common parent.

The paper is structured as follows. In the remainder of this introduction we focus on describing current issues in the growing field of prosodic typology, introduce the claims that have previously been made regarding AA prosody as well as the prosodic structure of surrounding languages more generally, and define the terms that we will use throughout the rest of the paper. In section 2 we provide a detailed critique of D&S's claims, involving review of recent work among Sora and other Munda languages, information about Munda word-hood, observations regarding areal and genetic homogeneity, and phrasal prosodic features that may be common to Munda and Khasian languages. In section 3 we then describe the instrumental study we conducted to compare the languages. Section 4 concludes with a discussion of the findings and implications for further research.

1.1 Prosodic typology

Before looking at the specifics of the data and detailing the issue within the field of comparative Austroasiatic linguistics, we first turn to a (brief and by no means comprehensive) synopsis of the meta-theoretical and methodological issues in the description of prosodic domains and word/phrasal intonation. Such issues are of particular concern for researchers attempting to analyze these features of Austroasiatic (and other) languages.

There is currently a lively, ongoing debate about prosodic analysis, and points of discussion include: 1) what the basic units of analysis should be, 2) what the properties that define these prosodic units are and how they contrast with each other, and 3) what the interface between the abstract phonological structures and their instrumentally trackable acoustic manifestations or cues might be. Among the key scholars contributing to this debate, Hyman, Gordon, van der Hulst, and Jun all have fairly distinct approaches partly overlapping in assumptions and terminology. All, however, assume that different phenomena belong to and operate at different domains or levels in the generation of well-formed utterances.

Such a position minimally recognizes a lexical layer and a post-lexical or post-grammatical layer with inherent hierarchical relations.[3] What this means is that such hierarchical levels could logically reference different interface options with other domains (lexicon, syntax, etc.), not strictly the phonetic/phonological domains alone. This is important because many of the distinctions made by these theorists in this area reference such different analytical strata and assume *a priori* that they exist. However, not all approaches to syntax are multi-stratal or derivational/generative, and therefore it is not fully clear how such terms can be reconciled with other (e.g., functionalist) approaches to 'grammar'.

Given the diffuse and varied nature of the acoustic or phonetic cues associated with perceived prominence, i.e., 'stress' or 'accent', it is well known that duration, intensity, fundamental frequency, spectral tilt and various other features sometimes referred to as 'hyper-articulation' can all conspire to serve as acoustic correlates to 'stress'. Complicating this further is that there are virtually no studies on the prosodic relationship of 'word' vs. domains of strings of sounds at a higher hierarchical level within a language. That is, we only know that words recorded in isolation are different prosodically than those that occur in natural or focused speech due to physical constraints on phonation and speaking (e.g. terminal drop), phrasal or utterance-level intonational features/characteristics of the language, or discourse-sensitive prominence marking through pitch or intonational perturbations, whether anchored in information *structure* (e.g. focus vs. topic) or in information *status* (given vs. new).[4] This makes prosodic research a fascinating and yet difficult topic, in which researchers attempt to find patterns while at the same time trying to control for effects at multiple levels.

Generalizing for the sake of synopsis and at the possible risk of oversimplifying, there is a major difference between the views expressed by Hyman and that of van der Hulst, Gordon and Jun: Hyman considers only 'tone' and 'stress' to be the relevant and active parameters, and that so-called 'pitch accent' systems are nothing more than intermediate systems between canonical stress systems and canonical tone systems, whereby pitch-denoted prominence is restricted to syllables that are potentially bearers of primary stress. 'Tone' in a language is defined by a system where "an indication of pitch enters into the lexical realization of at least some morphemes" (Hyman 2006:229). 'Stress' is defined as having to do with "metrical prominence" (Hyman 2006:231) in a system whereby stress accent is defined by two features, "obligatoriness" and "culminativity" such that "every lexical word must have *at least* one syllable marked for the highest degree of metrical prominence" (obligatoriness) and "every lexical word must have *at most* one syllable marked for the highest degree of metrical prominence" (culminativity). In other words, every word must have a head syllable in the Autosegmental-Metrical framework (sometimes recast as a requirement for a head foot). Hyman (2006:231) further identifies a salient functional distinction between 'stress' and 'tone' in language systems: tone is paradigmatic and distinctive, stress syntagmatic and contrastive.

Some prosodic systems seem to behave very differently, in what researchers have called 'pitch accent' languages. For Hyman, pitch accent is a system partly tonal and partly 'stress'-based. Other researchers have very different definitions of what they consider to be pitch accent. Thus, according to Hayes (1995:49-50)

---

[3] With respect to the hierarchical structuring of sound and prosodic units, while not every language recognizes all possible levels, a maximal extension of such domains might include the following levels smallest to largest: mora > syllable > foot > prosodic word > accentual phrase > intonational phrase > utterance. This leaves aside for now the non-trivial issue of how these combine and how to conceptualize the structure that models their hierarchical organization.

[4] As Roettger and Gordon (2017: 7) remind us "spontaneous speech is expected to yield more variable data due to the larger number of confounding factors (e.g., higher level prosodic structure, intonation, syntax, etc.)" and also that in words recorded in frames (2017: 4), even if shielded from edge, invariably the frame introduces new information and the words in the different slots are implicitly contrasted with each other.

"pitch-accent languages must satisfy the criterion of having invariant tonal contours on accented syllables, since tone is a lexical property". Bybee et al. (1998:227) put it this way: "a pitch accent system is one in which pitch is the primary correlate of prominence and there are significant constraints on the pitch patterns for words". Jun (2014) defines pitch accent, incredibly, as a system in which a certain, but not every, syllable of the word has lexical specification of pitch, showing syntagmatic contrast, but not 'stress' in the sense of Beckman (1986).

Harry van der Hulst and Matthew Gordon call into question Hyman's two conditions of "obligatoriness" and "culminativity" as necessary and sufficient to characterize stress to the exclusion of other features. As all researchers reviewed here point out, while one can identify the most prominent syllable in a string of connected syllables within a domain (the 'prosodic word'), and this is a psychologically real concept to speakers, the acoustic correlates of such perceived prominence can be a conspiracy of different phonetic features such as duration, intensity, and pitch. So 'stress' per se is a diffuse concept or a kind of epiphenomenon of perceived salience, however acoustically cued in particular languages. Gordon and van der Hulst (to appear) suggest that duration is the most reliable and consistent cue of 'stress' cross-linguistically, while Gordon and Roettger (2017:7) suggest that fundamental frequency played a greater predictability role in 6/11 languages in a sample, while duration did for only 5/11 of these (three of which have lexically-distinguishing tone as well).

van der Hulst (2014:5) comments that "when stressed syllables are measured in out-of-focus positions they do often not include pitch as a significant factor", downplaying the role of pitch in word-level prominence distinctions. He notes further (2014:28) that in many cases, pitch properties associated with stressed syllables are actually those linked to an intonational pitch accent. So pitch prominence in a speech stream may turn out to reflect other intonational/prosodic considerations as natural declination in pitch towards the end of utterances or discourse-grounded focalization or topicalization of the prominent element. As Hyman (2006:246) reminds us, for particular languages, "if word stress is so hard to find, perhaps it is not there at all". This is worth keeping in mind for the discussion on the varied and opposing views of 'stress' in the Munda languages below.

Turning to 'stress' proper, van der Hulst (2012:1495) breaks the meta-category of 'stress' into several components that interact or 'conspire' in its surface manifestations, specifically he enumerates 'stress' relating to four distinct phenomena: accent, edge prominence, rhythm and weight, each of which may have a 'stress correlate', viz., stress^{A} stress^{EP} stress^{R} and stress^{W}. The first three correlates involve strengthening of articulation with effects on duration, intensity and pitch; stress^{A} is lexical and may also correlate to greater phonotactic complexity, while stress^{W} may be nothing more than the perceptual effect of the intrinsic properties of heavy syllables. On the reasons to distinguish 'stress' from 'rhythm', van der Hulst (2012:1496) comments that "primary 'stress' location is often subject to morphological information and lexical irregularity, (but) the distribution of rhythmic beats appears to always be fully regular and automatic", or in his framework, primary stress belongs to the lexical level, full rhythmic pattern to the post-lexical level in a multi-stratal, generative grammar. On the distinction between 'accent' and 'rhythm', van der Hulst (2012:1497) comments that accent is the formal representation of primary 'stress' and rhythm pertains to rhythmic/secondary stress, but that the 'window' of 'accent' is restricted to word peripheries, specifically to building one foot at the Left or Right periphery of a word, while in 'rhythm', beats are assigned to syllables post-grammatically, with an interface condition that an accented syllable must have a rhythmic beat too; in the hierarchical ordering of elements this belongs to the utterance-level. Subsequent to that, a second factor operating at the utterance level is edge prominence, which serves to strengthen syllables on the edge opposite to the accent.

In this model, A, EP and R are all abstract notions with no inherent phonetic content, while 'stress' only means the phonetic correlates of these three (little more than prominence) and instantiates articulatory force that exaggerates or hyper-articulates inherent properties of the speech signal along dimensions of time (duration), fundamental frequency (pitch), intensity, etc. Thus, A, EP, and R are properties of syllables but syllables come in different forms: they may have different weight, for example, and heavy syllables however defined often have a special status in systems. As van der Hulst (2012:1497) puts it, heavy syllables can influence the location of A, EP and R beats, but in the absence of these extrinsic features, heavy syllables can intrinsically be perceived as prominent. In effect, 'stress' really is nothing more than prominence and the phonetic properties usually called 'stress' basically serve to provide the head with greater perceptual salience than all non-heads in the hierarchical organization mentioned above (van der Hulst 2012:1513).

The work of Jun is primarily concerned with units above the word level. Jun (2014) distinguishes at least three possible levels relevant to prosody and intonation above the word, but says some languages only use one. The level immediately above the word she calls the Accentual Phrase [AP] and the highest level the Intonational Phrase [IP], with an intermediate node called "ip". Prosodic units in Jun's approach (2014:433) can be defined by two different types of domains and their associated phenomena: by the degree of juncture and by intonation pattern. An AP may be mora-, syllable- or stress-timed (Jun 2014:432). Generally, a prominent word comes at the beginning or end of a prosodic unit (as for van der Hulst) and there is a phrasal tone demarcating the edge of such a unit. If a sentence is short, it will be one IP, if it is long it will be broken into two or more IPs. The length of an IP can vary considerably language to language, e.g., 7-15 syllables, but typically they are 1.5 seconds in length on average (when controlled for discourse style and genre), while APs tend to include one morphosyntactic content 'word' (plus optionally other elements) and are typically 3-5 syllables cross-linguistically.

Jun (2014:440), following Cruttenden (1997:8-9), define tone, as Hyman does, by a system that has "prescribed pitches for syllables or sequences of pitches for morphemes or words" with a paradigmatic contrast and that stress accent systems can be identified if a certain syllable in a word is more prominent than others in duration and/or amplitude, thus showing syntagmatic contrast. In Jun's terms 'rhythm' and 'prosody' pertain at different levels. Thus (Jun 2014:441) states that the rhythmic pattern refers to a timing unit below the level of the word, while the prosodic pattern refers to a prosodic unit above the level of word. So for Jun, 'word stress' would be a rhythmic pattern, while 'phrase stress' would be a prosodic pattern.

There is general agreement on certain prosodic patterns cross-linguistically, such that according to Gordon et al. (2010:133) the "unmarked intonational tune in most languages consists of a final pitch trough... whereas phrasal stress is typically associated with raised pitch"; some languages resolve this by pushing stress to the penultimate syllable. This also aligns with Ladd's (2001:1381) comments that overall pitch trends within an Intonational Phrase show a decline from beginning to end, and new ones can be marked by a local sharp rise 'reset'. IPs in utterance-final position typically have a drop of pitch at the end. This can be moderated, suspended or even reversed in non-final position and in questions.

We can see from this review of the field of prosodic typology that there is some diversity with regard to how prosodic contour relates with specific word or phrasal effects, uses, and phonetic realizations. However, all the authors we reviewed here agree that prominence plays a role in disambiguating units of speech, though exactly which units are being disambiguated (word, phrase, etc.), their exact relationship to morphological/grammatical/discourse features, and the exact correlates of such prominence, is a language-specific question. For our purposes, this requires that we specify the precise correlates of the prominence being investigated for all the languages under investigation, to ensure that we are comparing similar things.

Moreover, for polysynthetic languages such as Sora, specific issues arise in the analysis of the word vs. phrase, and indeed the difficulties inherent in analyzing prosodic domains in general are magnified. As in many languages that have been described as polysynthetic (Bickel and Zúñiga 2017), the notion of word in a unitary sense is largely elusive in the languages of the Munda family. Approaching the issue from the perspective of prosodic typology, Munda languages offer conflicting information as to what one would want to call a 'word' both within individual languages and across the family as a whole, with different (morpho)phonological processes defining different prosodic domains equal to, as well as smaller and larger than, what a 'traditional' understanding of 'word' would entail, and thus, projecting back into earlier stages in the development of the languages, what the prosodic, etc., structure of a 'word' in proto-Munda might have been like. We revisit this in our critique of the claims of Donegan and Stampe below.

1.2 Donegan and Stampe on prosody in AA

The two scholars who have most seriously broached the subject of prosodic structures in AA languages are Patricia Donegan and David Stampe (henceforth 'D&S' except where specific papers are cited), whose major articles (Donegan and Stampe 1983; Donegan 1993; Donegan and Stampe 2002; Donegan and Stampe 2004) have been widely cited by scholars on this issue. These articles consecutively update many of the same arguments and claims regarding the prosodic nature of Munda vs. Mon-Khmer languages – the term

"Mon-Khmer" here refers to all non-Munda AA languages, following a traditional branching-tree reconstruction of AA.[5]

The central claim of these papers is summed up in the statement by Donegan and Stampe (1983:1) that "Munda and Mon-Khmer are typologically opposite at every level" and their most recent tabulation of differences (from Donegan and Stampe 2004:3) is partially reproduced here as Table 1 below. It seems clear that the differences tabled there are major, and such has been the accepted view of linguistic scholars for many years. The general observation is that there is a clear typological distinction between the AA languages located in eastern India (Munda languages) and all other AA languages, particularly when it comes to what is counted as a 'word'. The reason for such a difference has been attributed to the influence of neighboring Indo-Aryan and Dravidian languages (see Emeneau 1954, 1956; Pinnow 1963, 1966), which are described as largely agglutinating and polysynthetic. South-East Asian AA languages, on the other hand, are considered largely isolating and only mildly synthetic (see above references, also Enfield 2005).

Table 1: *Differences between Munda and Mon-Khmer (as per Donegan and Stampe 2004)*

	Munda	**Mon-Khmer**
Grammar:	Synthetic	Analytic
Word Order:	Head-last: OV, Postpos.	Head-first: VO, Prepos.
Phrases:	Falling (initial)	Rising (final)
Words:	Falling (trochaic)	Rising (iambic/monosyllabic)
Affixation:	Pre/infixing, Suffixing	Pre/infixing or Isolating
Timing:	Isosyllabic, Isomoraic	Isoaccentual
Syllable Canon:	(C) V (C)	(C(u)) + (C) V (/) (C)
Etc..		

To explain the differences that they tabulated, Donegan and Stampe (2004:5) sought "a linguistic opposition which might pervade and organize every level from syntax to phonetics." Despite the fact that such linguistic oppositions do not really exist, in their analysis they seem to have discovered that "the only plausible candidate is initial vs final accent in phrases and in words" (2004:5). This is a slightly different formulation from Donegan (1993), where she frames the discussion of prosody in terms of *falling* (2004: initial) vs. *rising* (2004: final) accent. As she states in this paper: "Regarding the phonetic manifestation of accent, I will mention only stress and pitch. Stress accent seems to be a combination of greater effort and greater length." (1993:10) She goes on to say that:

> "Falling-accented languages are typically mora-timed, and in that case there can be no lengthening of accented syllables, and so they mark accent, if at all, with pitch. But rising-accented languages, if they are stress-timed, are free to lengthen accented syllables, and so they mark accent with stress." (1993:11)

Besides using somewhat confusing terminology (from what we can tell, 'accent' is used by D&S somewhat interchangeably with 'stress' and also has the sense of 'prominence' as used in more recent phonological literature), D&S seem to state that in languages described as being 'mora-timed', accent (or stress) is primarily marked by pitch (F_0). Languages described as being 'stress-timed,' on the other hand, mark accent (or stress) by a combination of length/duration and intensity. This sets up a dichotomy that separates languages into those that use pitch for stress and those that do not (though stress is usually a bundle of features, of which pitch is often a component, see Gordon and Roettger 2017). It is not fully clear whether her use of 'mora-timed' is intended to align with the 'heavy' and 'light' syllables identified by Henderson (1952) for Khmer. Most AA languages, are syllable(mora)-timed (according to D&S's definition) and it is

[5] Sidwell (2015) suggests that a more strongly-branching radial tree with "spokes" from a single origin might have more explanatory power. Independently but similarly Anderson (2015, 2016) has suggested that within Munda only North Munda with sub-branches of Kherwarian and Korku, Sora-Juray-Gorum and Gutob-Remo, consisting of the named languages, are valid intermediate taxa, all other languages being isolated groups coordinate with these.

thus unclear why Donegan argues for 'stress-timed' languages being equated with 'rising accent' when most Mon-Khmer languages (in her view) are not 'stress-timed' but do show 'rising accent'.

Donegan (1993:3-5) uses the observations in her paper to make historical claims, stating that:

> "Proto-Austroasiatic had rising accent and head-dependent word order, like Mon-Khmer. Munda languages reversed the structure to falling accent and dependent-head order, but preserved the old word order in the morpheme order of complex words… Proto-Austroasiatic, like Mon-Khmer (and other mainland Southeast Asian languages) had rising accent not only in phrases but also in words. Munda shifted to falling accent not only in phrases but also in words."

These claims nicely account for the perceived differences between "Munda" languages on the one hand, and "Mon-Khmer" languages on the other, but unfortunately they gloss over some of the discrepancies between this generalization and the actual prosodic realization of individual languages. We address this issue in sections 2 and 3 by presenting data on some of these languages.

1.3 The terms defined

Before venturing further into the discussion, and based on the issues raised in sections 1.1 and 1.2 above, we need to define the terms we will use for the remaining sections of this study so that it is at least clear to the reader what we intend when we use the terms. Unfortunately, due to the lack of clear correlates and definitions of terms in D&S, this requires us to make some assumptions.[6]

We assume that the use of "accent" by D&S corresponds to "stress", in part because the words are used somewhat interchangeably in their work. Below we use the term "stress" to refer to a language-specific indication of prominence at the syllable level (within a word or phrase). We also identify three phonetic correlates of stress (pitch/F_0, length/duration, intensity) and their relation to its realization in each of the languages we discuss. This means that for each language, we can identify a phonetic cue (or cues) that indicate(s) prominence of a particular syllable within a word or a phrase.

We take the use of "rising" in D&S to refer to an increase in some phonetic correlate of stress across the time-span of the domain of investigation in question (word, phrase) and the term "falling" to refer to a decrease in such a correlate. We can then use this correspondence in terms to compare our findings with theirs. So if increased pitch is a primary correlate of syllable prominence for a particular language (as it is for the languages in our study), and given a two-syllable word in that language where pitch increases to the second syllable, the word can be described as having a "rising" stress pattern. Whether the language consistently shows this pattern in two-syllable words, and whether such a pattern aligns with D&S's claims for the particular language can then be assessed.

2. Detailed Critique of D&S's treatment of Munda and 'Mon-Khmer'

As noted above, D&S make strong claims about rhythmic holism based primarily on observations and examples drawn from Sora, in comparison with sweeping generalizations regarding the non-Munda Austroasiatic languages. Below, we critique several specific issues regarding these claims: Sora correlates of

---

[6] In the full-scale study called for by our preliminary study here, we must rigorously define all cross-linguistically quasi-generalizable terms that we might need in the prosodic analysis of the Austroasiatic languages diachronically and synchronically. Thus, to use one scale of increasing prosodic domains (Hildebrandt and Bickel 2007), one might seek to determine the existence and defining parameters of such units as *mora, syllable, foot, 'word'* (*p-word, m-word*), *'phrase'* (*phonological phrase, intonational phrase, accentual phrase*), and *utterance*. Here the notoriously problematic 'word' and 'phrase' may have several non-overlapping or partly overlapping domains they encompass, which may or may not have unique relationships with other terms as widely understood or used in typology, or with independently validated syntactic units in languages under investigation. However, as there are a wide range of features that might help to determine whether such units have any analytic validity for any given language, each language will likely have different and specific manifestations of prosodically sensitive or otherwise phonologically active processes whose patterning would uniquely determine the nature and extent of the various analytic atoms that one might seek to compare. In addition, we must have a rigorous definition of each of the following: *stress, tone, intonation, accent,* and *rhythm*, so that we know that we are truly comparing like with like when we approach the comparative study of prosodic phonology in Austroasiatic.

stress (section 2.1), Munda word-hood (section 2.2), areal and genetic over-generalizations regarding "Mon-Khmer" and South Asian phonologies/prosody (section 2.3), and phrasal prosodic features of Khasian languages that were ignored by their analysis (section 2.4). Before we move to these specific critiques, however, we have several general critiques of D&S.

D&S claim that there is a 'rhythmic holism' in language that drives word formation, and while we are sympathetic to the idea, it risks oversimplifying a highly complex issue. The various concerns to be addressed here are: 1) the descriptive accuracy of prosodic correlates in the individual languages that such a claim is based on, 2) the role of prosody at the sentence and the word level in individual languages, 3) the historical development of these languages, and 4) the lack of data to back up the claims of the authors.

The last issue is related to the others, and so we address it first. We take the position that scientific linguistic research should be evidence-based. Strong claims should be based on strong evidence. The claims of D&S, while repeated in multiple publications, are supported by few Sora examples, with only one of the sentences annotated for prosodic rhythm, and no quantitative information about the analysis they undertook.[7] We acknowledge the difficulty of making and analyzing recordings during the 1980s-90s, and the fact that linguists during this period had a different standard for acceptable evidence to back up their claims. In the 21st century, however, overarching claims of phonological (prosodic) patterns that are not backed up by diagrams, statistics, and (or at the very least) reference to a corpus of data, if only rudimentary, are highly problematic, particularly when programs/technology such as Praat (Boersma and Weenik 2018) and handheld recorders are widely and easily available, not to mention many means for sharing data online. This is the main issue that we attempt to remedy (albeit incompletely) with our pilot instrumental study.

Regarding the first issue noted above, there is concern regarding the descriptive accuracy (perhaps 'descriptive completeness' is a better term) of prosody, stress, and their correlates in existing descriptions of the Austroasiatic languages in question. To our knowledge, existing descriptions of these languages do not describe such features to a great degree. Therefore, it is unclear what data besides their own underlies such claims by Donegan and Stampe about overarching "Munda" or "Mon-Khmer" prosodic patterns. To clarify: many of the AA language families in question have only a few languages for which some description exists, and these descriptions often either contain no prosodic information or contain only limited and impressionistic accounts of prosody. We discuss this further in section 2.2 below.

Regarding the second issue, D&S present the Munda languages as a monolithic entity with no internal variation that aligns 'rhythmic' structures 'holistically' with purported South Asian norms. However, what do we know about the role of prosody at the sentence and word level in individual Munda and "Mon-Khmer" languages? For that matter, beyond impressionistic observations (see Khan 2016), what do we know about prosody in South Asian languages? It seems clear from research on prosody that different languages have different uses for prosody at the word and sentence level for (word/sentence) boundaries, supra-lexical information and potentially other features (besides the references reviewed above, see Kawaguchi et al. 2006; DePaolisa et al. 2008; O'Brien et al. 2014; Xu 2011, 2012). Further, and as noted above, prosody researchers do not always agree on terminology, though recent work in this area is beginning to bring clarity. In section 2.3 we discuss this in more detail by providing data showing that the word and prosodic picture is rather diverse for Austroasiatic and South Asian languages.

Regarding the third issue, the claims of D&S are not just a claim about the existence of certain phonological features/structures in the modern languages, but rather a hypothesis about the state of the parent

[7] A scan of Austroasiatic examples in the work of D&S gave the following counts. In Donegan and Stampe (1983): 14 Sora sentences illustrating syntax, 7 Khmer sentences illustrating syntax, 7 Sora words illustrating morphemes, 2 glossed Sora examples, 2 annotated glossed Sora examples, 2 Sora examples from poetry/verse, and 1 glossed Khmer example. In Donegan (1993): 3 Sora genitive sentences, 3 Sora genitive words, 1 Sora glossed word, 1 Sora glossed clause, 11 glossed Sora words with stress/accent marking. In Donegan and Stampe (2002): 3 disyllabic Khmer words with accent marking, 3 disyllabic Sora words with accent marking, 4 glossed Sora sentences, 4 glossed Khmer sentences, 1 Sre sentence, 4 Sora words/phrases of 3 or 4 syllables, and set of word comparisons in Kharia, Sora and Mundari illustrating sound changes. Donegan and Stampe (2004): 3 Sora sentences, 1 Khmer sentence, 3 Khmer words, 3 Sora words, 1 Sora word with rhythm annotated, 3 Khmer sentences with rhythm annotated. Many of these examples are re-used across publications and there is no indication regarding how representative such examples are of their data.

language, Proto-Austroasiatic. Germane to this concern is the observation mentioned above in section 1.1 regarding the problem of 'word' vs 'phrase' in polysynthetic languages, significant not only from the perspective of an adequate synchronic analysis of the contemporary Munda languages, but their history as well. D&S treat the Munda languages as a monolithic entity in opposition to "Mon Khmer" languages with respect to effectively all typological features, allegedly mediated and triggered by a fundamental shift from rising to falling rhythm. Such a total resetting of typological parameters would have to have occurred (likely gradually) during the emergence of proto-Munda from late-proto-Austroasiatic/pre-proto-Munda for it to uniformly apply to all the modern Munda languages. As it turns out (and as we discuss below), there is both significant and revealing synchronic variation among the Munda languages that demonstrate that a one-time resetting of parameters is untenable, and for some of the languages has not occurred.

Below we review some recent work on Sora, address specific claims of Munda word-hood based on examples used by D&S, relate this to what is known about stress correlates in Munda languages generally, and highlight the diversity of phonological realizations found in Austroasiatic languages of Mainland South-East Asia as well as in South Asian languages neighboring the Munda languages.

2.1 Sora stress correlates (Horo and Sarmah)

With a series of phonetic studies, Horo and Sarmah (2014, 2015) worked on the variety of Sora spoken in the Assam tea gardens. Speakers of Sora in Assam are recent migrants in the last century, and while D&S primarily refer to Orissa Sora, Horo's (2017b) PhD thesis includes an acoustic study of Orissa Sora, which shows that the two varieties do not differ significantly. As a result, findings in these studies serve to dismantle several claims about the structure of words and the vowel system of Sora made by D&S.

In Sora (whether Assam or Orissa as a whole), correlates of stress include pitch (F_0), whereby strong stress correlates primarily with higher relative pitch, duration (longer = stronger stress), and intensity (increase = stronger stress).[8] For Sora, Horo and Sarmah (2015:78) determined that "vowels (in Assam Sora) in the first syllables are more centralized" while "vowels in the second syllable are more representative of the canonical vowel space". This is exactly counter to what a 'falling' word rhyme would predict, where initial vowels are more canonical.[9]

Their analysis also examined Sora words where one might expect to get a different tendency, such as in an open syllable followed by a closed syllable. Even in these forms the data is counter what one might expect, such that "in V.CVC words, even though the vowel in the second syllable is in a closed syllable, the vowel in the first syllable is still significantly shorter than the vowel in the second syllable" (Horo and Sarmah 2015:79). In sum:

> "the second syllable is stressed in a disyllabic word in Assam Sora, characterized by greater pitch, longer duration, and by change in vowel quality… [and] the second syllable displays higher F_0 and duration of the vowel… [all of which] suggest [that it has] greater prominence" (Horo and Sarmah 2015:82).

Acoustically speaking, then, the phonetic details of Sora do not support previous assertions about falling word prosody of Sora disyllabic words and, by extension, Munda as a whole – which is a core/central assertion of Donegan and Stampe's (2004) thesis. Rather, the acoustic findings suggest that iambic words (right-headed) are the norm. These may combine in trochaic (right-headed) phrases (see below), but with sequences of iambic words. In other words, Sora (and other Munda languages like Remo, see below) appear to conform to a word prosody more in line with other AA languages (and likely an "old" inherited structure).

---

[8] Horo (2017a, 2017b) demonstrates that there is dialectal variation in Assam Sora in some details, such that while Lamabari Assam Sora does not clearly use F_0 differences to differentiate word stress in disyllabic words, Koilamari Assam Sora and Raiguda Orissa Sora do use F_0. However, in terms of the major findings regarding placement of stress in disyllables, all varieties are shown to be similar, and we discuss these studies below.

[9] They go on to demonstrate that "(t)he first syllable has statistically significant lower F_0 and maximum F_0 than the second syllable" (Horo and Sarmah 2015: 80). They also state that "(t)he vowel space in initial syllables is reduced. …the average F_0 and maximum F_0 of the second syllables is higher" (Horo and Sarmah 2015: 82). Note however that low pitch may signal prominence in other Munda languages like Kharia, mentioned further below.

2.2. Munda word-hood

The primary examples used by D&S are of attested polysynthetic words in Sora, which they state show a "falling" contour. There are inherent issues with the use of this data, since it is not known how the words were recorded. If in isolation, phrasal prosody and/or utterance intonational contours may be at play in addition to (or exclusive) to word-level prosodic features. Even if in a frame removed from edge effects, e.g., "I ___ said", this is an inherently contrastive position and cannot be considered to be fully independent from potential information structure effects of intonation.

Moreover, their analysis does not align with what speakers of the language seem to conceive of as a 'word'. In a test of word-hood in Sora conducted by Anderson and other researchers, native Sora speakers with knowledge of transcription of their mother tongue (in different orthographies) were asked to listen to sequences recorded from other speakers and to transcribe the words in the recording. The speakers consistently wrote combinations of characters that decomposed morphological words into smaller units. While not fully conclusive, this suggests a strong tendency to correlate iambic structure to the unit 'word'.[10]

Put differently, large morphological words are often conceived of by Sora speakers as sequences of iambic phonological words, with certain prosodically weak elements perceived as permissible (but unstressed) in words as well. Sora has long been recognized linguistically for its large morphological words with lots of internal complexity. However, the constructs that many linguists consider words are recognized by Sora speakers as phrases. Where a linguist might transcribe them as a single unit, speakers break them into two- or three-syllable sequences of words with a rising contour. To illustrate, the following two examples (1-2) were given by Donegan and Stampe (2004:4) as particularly idiomatic renderings, but neither were considered single words when tested with native speakers (Anderson, field notes).

(1) *ədməltijdariŋdae*
əd-məl-tij-dar-iŋ-da-e
NEG-DES-give-rice-1.UND-AUX:TAM-3.ACT
'he does not want to give me rice'

(2) *əədnəlgəbrɔɟlaj*
ə-ədn-əl-gə/b/rɔɟ-l-aj
1PL-NEG-RECIP-shame/CAUS/shame-PST-1.ACT
'we did not shame each other'

The first form (1) was rejected when given without a subject pronoun, further underscoring its perception as a phrase/sentence and not a single word by native Sora speakers. It was repeated as follows (3).

(3) *anin* *{[əb-mə́l]$_{pw}$+[tijg-dar-íŋ]$_{pw}$=[dá-j]$_{pw}$}$_{mw}$*
3.PRON NEG-DES=give-rice-1.UND=AUX:TAM-3.ACT
'he does not want to give me rice'

The second word (2) was also rejected as one phonological word, despite being a morphological word conceptually. In this case the sequence of '1pl' and 'neg' marker at the beginning were reduced to a single element, suppressing the subject marker (4). In Gajapati Sora, there appears to be one prefix slot now shared

[10] A nearly identical pattern is observed in Kherwarian languages, specifically Mundari, Ho, Birhor, Bhumij, Kera' and Santali, where linguistically trained native speakers took part in a series of transcription exercises. Long morphological complexes were almost invariably written as sequences of disyllabic or maximally trisyllabic units. Given that Sora and Kherwarian languages have the most complex morphology of all Munda language subgroups and in theory could produce the longest morphological words, this is highly suggestive that a disyllabic or trisyllabic unit 'feels' most like a 'word' in Munda languages as a whole. While neither test has been adequately pursued nor quantitatively assessed, anecdotal skewing towards disyllabic words (with some trisyllabic sequences) in both of these contexts suggest that a more systematic implementation of such a test would support native speaker intuitions of di- or tri-syllabic wordhood.

between the plural subject markers and negative marker, with the negative taking precedence when both are allowed for semantically. This may be a change in the language since data collection in the 1930s-1960s.

(4) *{[(ə)-ədn-ál]$_{pW}$ [gə/b/rɔ́j]$_{pW}$ [l-á.j]$_{pW}$}$_{mW}$*
(1PL)-NEG-RECIP shame/CAUS/shame PST-1.ACT
'we did not shame each other'

Sora is not alone in showing this tendency in Munda: iambic words dominate and typify Gtaʔ (Anderson 2008, in preparation-a) and Gorum (Anderson and Rau 2008). Mundari (Osada 2008) and Kharia (Peterson 2011) are reported as having iambic words, and Santali (Ghosh 2008:30) and Juang (according to Patnaik 2008) are reported to have fixed second-position stress. In Remo, which has second/final position stress, a two-syllable word has final stress (5a), while a three-syllable morphological word may have second syllable stress and an optionally extra-metrical grammatical index in final position (5b) or the final syllable may be stressed; Gutob has a similar system (Anderson in preparation-b). Four-syllable Remo morphological words ($_{m}$w) first are assigned to phonological prosodic words ($_{p}$w) in an iambic pattern with primary stress on the second syllable of the first word, and secondary stress on the fourth syllable (5c-5d):

(5) a. *sum-óʔ*
eat-PST.TR/ACT
'she ate'

[sum-óʔ] $_{p}$w
- *
- **

b. *sum-óʔ=niŋ*
eat-PST.TR/ACT=1SG
'I ate'

[sum-óʔ=niŋ]$_{p}$w
- *=ø
-**=ø

c. *susúm=ɖen-t-ìŋ*
RDPL~eat=IPFV-NPST-1SG
'I am eating'

{[*susúm*]$_{p}$w=[*ɖen-t-ìŋ*]$_{p}$w}$_{m}$w 'I am eating'
- * - *
- ** - *

d. *a-goi=tə-no*
NEG-die-NPST-2SG
'you do not die'

{[a-gói]$_{p}$w=[tə-nò]$_{p}$w}$_{m}$w 'you do not die'
- * - *
- ** - *

{[Right headed]$_{p}$w +[Left headed]}$_{m}$w

These facts run counter to D&S's claims regarding stress in Munda languages. Having said this, a review of literature on the Munda languages reveals mixed results when it comes to what has been reported regarding prominence or stress and its acoustic correlates in individual languages of the family, such that D&S are not entirely to blame for their conclusions. We summarize this diversity in Table 2 by language and source, including what is reported for the placement of prominence or prosodic contour in the language and its reported acoustic correlate. While this data and the analyses cited here deserve fuller treatment, we limit ourselves to making some general observations.

One observation is that not all the Munda languages are represented in this table, and for several languages in the table accounts and analyses differ or are couched in different terms. For Gutob, Judith Voß

(p.c.) identifies a LH prosodic contour with a primary correlate of pitch, while Griffiths (2008:639-40) cites Zide (1965:44) as stating that heavy syllables of the word are stressed, presumably based on quantity of morae. For Ho, Pucilowski (2013) identifies a trochaic stress pattern, while Anderson, Osada and Harrison (2008:204) state that the language exhibits initial stress – neither source identifies an acoustic correlate of this measure. For Juang there are two accounts (Patnaik 2008:513; Dasgupta 1978:20) that seem to contradict each other – one that identifies main stress on the second syllable (with a main correlate of pitch) and one on the first syllable. For Sora, the sources seem to identify the same correlates, but D&S claim that the language shows trochaic prominence, and Horo and Sarmah identify iambic stress.

Another observation is that for most of the languages listed here, it is not clear what the acoustic correlate of prominence is. So while for Remo, both Anderson and Harrison (2008:565) and Bhattacharya (1968:xxii) identify main stress as occurring on the second syllable of disyllabic words, neither explicitly state what the acoustic correlate of this stress pattern is (though for the former, given the section heading, we can guess that 'pitch' is intended). The lack of clearly identified correlates in the majority of sources is problematic for broad generalizations about stress or accent patterns in these languages.

Table 2: *Munda languages, stress patterns, and acoustic correlates*

Language	Reference/Source	Prominence placement	Correlate
Gorum	Anderson and Rau (2008:386)	final (closed syllables)	unclear
Gtaʔ	Anderson (2008:686)	final stress	pitch
Gutob	Judith Voß (p.c.)	LH prosody	pitch
Gutob	Griffiths (2008:639-40)	heavy syllables stressed	unclear
Ho	Pucilowski (2013)	trochaic	unclear
Ho	Anderson, Osada, and Harrison (2008:204)	initial stress	unclear
Koɽowa	Barker (1953?/nd:31)	final stress (on verb stems)	unclear
Juang	Patnaik (2008:513)	stress on syll 2	pitch
Juang	Das Gupta (1978:20)	stress on syll 1	unclear
Keraʔ Mundari	Kobayashi and Murmu (2008:169)	LH in disyllables	pitch
Kharia	Biligiri (1965:19-20)	initial stress	unclear
Kharia	Rehberg (2003:23-28)	initial accent	low pitch
Kharia	Peterson (2011:35)	LH prosody	pitch
Korku	Zide (2008:260)	final stress	unclear
Mundari	Osada (2008:104)	normally final for 2-syll	pitch (accent)
Mundari	Langendoen (1963:14-15)	final if open	unclear
Mundari	Cook (1965:100)	final (if closed), else initial	pitch (accent)
Mundari	Sinha (1975:39)	second syllable (if heavy)	unclear
Remo	Anderson and Harrison (2008:565)	stress on syll 2	unclear
Remo	Bhattacharya (1968:xxii)	final stress	unclear
Santali	Ghosh (2008:30)	stress on syll 2	unclear
Santali	Neukom (2001:8)	initial unless syll 2 is heavy	unclear
Sora	Donegan and Stampe (2004)	trochaic	pitch, ints, dur
Sora	Horo and Sarmah (2015), Horo (2017a, b)	iambic	pitch, ints, dur

A third observation is that some languages in the table do not have easily comparable correlates of stress or accent – for Kharia, two sources identify a similar pitch pattern (LH) across words, but Rehberg (2003:23-28) states that the initial accented syllable is identified by its low pitch. If low pitch is the main

correlate of prominence in one language, and high pitch is the main correlate in another (such as Sora), this complicates the picture and adds another layer of historical development that must be accounted for.
From this brief summary we can see that the systems of stress assignment in Munda are both varied and subject to considerable analytical debate, with some seemingly incompatible analyses being offered for the same language by different researchers. But whatever the actual phonetic and phonological details are, rhythmic holism as a one-time parametric reset affecting all Munda languages is untenable, given 1) the diversity of prosodic patterns reported and 2) that some languages do not appear to have ever undergone phonological restructuring away from final or second-syllable stress at the word level (Gorum, Gtaʔ). It is clear that a systematic family wide survey of prosodic domains will be necessary to fully resolve this situation within Munda, and that Munda researchers need to work toward being able to compare the same kinds of things, possibly by identifying acoustic correlates of the features they describe.

2.3 Areal and genetic over-generalizations

Among the more obvious over-simplifications presented in the theory of rhythmic holism *à la* D&S are the lack of monolithic featural/structural complexes defining each of the groups. In other words, the alleged areal split is far from as clear-cut as implied in the work of D&S. As demonstrated above, Munda is not a monolithic entity in terms of the features they describe, and thus the distributional data at minimum require a diachronic periodization whereby different languages and sub-groups have accommodated to *local* norms or have undergone unique (simultaneous) developments through independently attested processes of borrowing/copying or metatypic shift. The diversity identified above cannot simply be explained via vague and semi-mysterious macro-areal processes of drift, nor a one time parametric re-setting of 'rhythm' showing pattern-copying at the proto-Munda level. Moreover, the non-Munda Austroasiatic languages (section 2.3.1) and the South Asian languages neighboring Munda (section 2.3.2) likewise do not represent a single monolithic, unvarying entity in terms of various phonological features.

2.3.1 Diversity in Austroasiatic

In terms of phonological structures and prosodic features, Mainland South-East Asian (MSEA) languages tend to a similar profile (c.f. Enfield 2005) that likewise typifies the AA languages spoken there: sesquisyllabic word structures with significant distributional restrictions on the nature and type of elements permitted in the minor syllable, and final stress on the major syllable. However, while in broad strokes these appear to be similar or identical, there is significant variation as to the nature of the restrictions on the segments and even on the structure of the minor syllables. For example, Schiering et al. (2007) identified different kinds of minor (and major) syllable templates in the AA languages they surveyed – to this we can add data from other Austroasiatic branches that they did not survey (Table 3).

Along with these clear differences in syllable template, AA languages also differ regarding which consonants can occupy C_1, C_2 or C_3 positions in either the minor or major syllable. Not all core MSEA languages belonging to AA require reduced vowels in minor syllables, but some rather restrict the set of full vowels that can occur in the initial syllables of words – a situation that is also found in some Munda languages. Thus, as in Munda, weak+strong word profiles endure even if the specific restrictions on the 'weak' syllables are not as pronounced or as strict as they are in many of the AA languages found in the core regions of MSEA.

Table 3: *Major and minor syllable templates in some AA languages*

Language	AA Branch	Minor syllable	Major syllable
Schiering et al. (2007):			
Khasi	Khasian	CC./Cə	(C)CCVVC
Khmu	Khmuic	CC. (Cə)	CCVVC
Semelai	Aslian	CəC/CuC	CVC
Car	Nicobarese	[NO]	CVC
Mon	Monic	Cə	C(C)V(C)
Vietnamese	Vietic	[NO]	C(w)V(V)(C)
Pacoh	Katuic	CV(C)	(C)CV(V)(C)
Cambodian	Khmeric	Cə	C(C)(C)V(V)(C)
Chrau	Bahnaric	CV	(C)(C)CV(C)
Other sources – Premsrirat and Rojanakul (2015); Li and Luo (2015); Deepadung, Rattanapitak and Buakaw (2015):			
Chong	Pearic	C(C)ə(C)	C(C)V(C)
Bugan	Mangic	[NO]	CV(V)(C)T
Dara'ang Palaung	Palaungic	CV/N	C(C)V(C)

Prosodic words in eastern AA languages tend to be maximally disyllabic in structure, which is also true of several Munda languages, regardless of how many syllables a morphological 'word' might entail. However, as in some Munda languages, there are several non-Munda AA languages which have words longer than two syllables. So for example, Car Nicobarese permits four-syllable words, as does Aslian Semelai, while Aslian Jahai can have four-syllable words of the structure C.C.C(V)(C).CVC, e.g., *tbtadɔʔ* [REL-PROG-wait] 'waiting' (Kruspe, Burenhult, and Wnuk 2015:423).

In terms of prosodic systems, Schiering et al. (2007:5) find word stress and phrasal stress to be final in Khasi and in Mon. However, Schiering and van der Hulst (2010:592) state that evidence for word stress is weak in a number of Austroasiatic languages included in the original StressTyp database:[11] they describe the status of word stress in Sedang as unclear and in Khmu' and Khmer as debatable, and consider the status of word stress in Khasi as highly debatable. Thus, we can conclude that non-Munda Austroasiatic languages (D&S's 'Mon-Khmer') are not a monolithic entity with regards to syllable structure, nor word (and possibly phrasal) stress. Given that Munda shows analogs to many of these features, the strict areal dichotomy of Munda vs. non-Munda in Austroasiatic, as asserted by D&S, cannot be maintained.

2.3.2 Diversity in South Asian languages

Languages in other language families of South Asia also do not present a monolithic areal profile prosodically. Indo-Aryan and Dravidian are neither identical nor uniform in their prosodic systems, as cogently demonstrated for Indo-Aryan by Khan (2016). In South Asian languages there are conflicting tendencies that further draw into question the notion of 'rhythmic holism'. South Asian languages are supposed to have "falling" rhythm (initial stress and a fall across the utterance), but Khan (2016:23-24) speaks of the fact that South Asian Languages [SALs] "are generally considered to have no lexical contrast in prominence ("stress") placement, and there are in fact no clear signs that stress is even a phonetic property of SALs at all" but are characterized by "repeating rising contours (RRCs) built from L tones on the left edge and H tones on the right edge of each content word, followed by a final boundary tone marking the edge of the intonation phrase (IP)". This supports Ladd's (1996) assertions of "Bengali (and probably most of the languages of India)" as representing "non-stress accent" with "postlexical pitch only".

[11] http://fonetiek-6.leidenuniv.nl/pil/stresstyp/stresstyp.html , see also http://st2.ullet.net/

There is clear evidence in the major South Asian language families (Indo-Aryan and Dravidian) for the prominence of the first syllable in disyllabic words. In Bengali (Dasgupta 2003; Khan 2008) the first syllable shows more vowel contrasts, while short [i, u, a] are centralized in non-initial position in Tamil (Keane 2004, 2014).[12] Intonational phenomena, on the other hand, can be quite varied across the major South Asian languages. Thus with respect to Indo-Aryan vs. Dravidian features, Khan (2016:30-31) states:

> "...the H target is typically AP-final in the Indic SALs studied (Assamese, Bengali, Nepali, Hindi), in line with the majority of previous work on SAL intonation; this can be considered the "typical" pattern for SALs. For Telugu and Tamil, however, the peak of the H target is (also) typically reached on the second syllable (Tamil) or third vocalic mora (Telugu), suggesting a complex pitch accent (L*+H) with a language-specific alignment specification for the trailing tone. In fact, I propose that Tamil and Telugu have more complex tonal templates available than for the other SALs studied, with the option of having two H targets per RRC, one closely following the prominence and another near or at the phrase boundary"

Within Indic and Dravidian there is non-trivial intra-family variation with respect to the realizations of the areally common repeated rising contours intonational patterns. Khan (2016:35) summarizes his study as follows:

> "The A[ccentual] P[hrase]'s L tone marks the prominent syllable, which can be non-initial in Hindi; Assamese shows more variation. Similarly, the AP's H tone can mark the right edge (Assamese, Hindi), the long vowel closest to the right edge (Telugu), the tail of the prominent syllable, or some combination of these, in alternation (Bengali) or simultaneously (Tamil)."

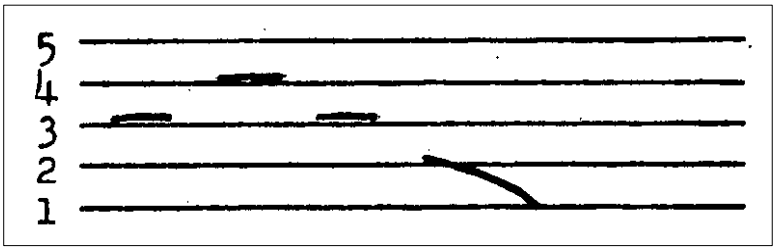
Data like this runs counter to a narrative of rhythmic holism as an organizing feature that defines South Asia vs. MSEA. The identification of 'falling' vs. 'rising' rhythmic patterns in these areas that account for a wide range of phonetic, phonological, morphological and syntactic features, is similarly problematic.

2.4 Phrasal prosody in relation to D&S's claims

Phrasal prosody generally in the Munda and non-Munda Austroasiatic languages deserves further investigation. As noted above and described further below in our pilot study, in two- and three-syllable units (words), pitch contour in Sora tends to rise to the end of the unit, which corresponds to the rising rhythm of D&S (or the 'Mon-Khmer' model). With longer units of four, five, and six syllables, however (described briefly below), there is often a peak at the second or third syllable, and then a fall to the end of the unit. If we simply consider pitch contour of the unit (which possibly aligns with grammatical phrases), Munda languages seem to show a "falling" pitch pattern after an initial rise, based on impressionistic observations.

This pitch contour in phrases is not simply a Munda phenomenon within Austroasiatic, however, as it also occurs in Khasian languages which belong to the "Mon-Khmer" group that D&S place in opposition to Munda. For Khasi, this was reported in the middle of the last century by Rabel (1961:32), who states that phrases "are characterized by a special pitch contour, which differs from the word pitch contour, which is basically 3:4:3:2-1" (Figure 2).

Figure 2: *Khasi pitch contour (reproduced from Rabel 1961)*



[12] As reported above, this is the mirror image of what Horo (2017a,b) found for Sora, where the second syllable in disyllables is more canonical and shows less centralized vowels. As we reviewed for Austroasiatic more generally, this situation precisely reflects the minor syllable : major syllable distributional properties that typify non-Munda Austroasiatic languages (Schiering and van der Hulst 2010).

The fact that this observation (albeit impressionistic) was missed or ignored by D&S is concerning, since by their terminology Khasi shows the same falling phrasal contour as Munda. Based on their argumentation, this contradicts their claim that "Mon-Khmer" languages (which include Khasi) show a rising contour in the phrase. The analysis of pitch is what we turn to in the following section, starting with pitch contour of words.

3. Instrumental analysis of word pitch

The carefully designed study by Horo (2017b) samples multiple varieties of Sora and conducts statistical analysis on a large number of recorded samples from multiple speakers of each variety. His study shows that in Sora the concept of 'word' aligns with units that are primarily disyllabic and have iambic stress patterns, with primary acoustic realizations of stress being pitch, duration, and intensity. Our study offers some initial data on three geographically disparate Austroasiatic languages. The data presented here is not intended to be conclusive, but rather to show that more research needs to be done in order to clarify the similarities and differences in word and phrasal prosody in Austroasiatic languages. We believe that this data necessitates significant refinement to the approach of (contact-triggered) restructuring that has occurred and is still ongoing among minority languages of India belonging to the Austroasiatic family.

We must also acknowledge that there are significant shortcomings with the quality of the data to be analyzed here. We have yet to design a controlled, laboratory-appropriate context for recordings of the languages involved that recognizes the shortcomings inherent in many previous phonetic and phonological analyses, as cogently pointed out by Roettger and Gordon (2017), where phrasal intonation is likely being measured. We also note that while Horo and Sarmah (2015) and Horo (2017a, b) demonstrate a clear iambic structure for disyllabic stems/words in Sora, there may be confounds. The recordings underpinning their analyses are of words in a standard frame, potentially a focus position where phrasal and information structure intonational dynamics may be in play. Our own data on connected speech suggests that there is a rise and then a fall in Sora units of 5-7 syllables. The relationship of intonation with information structure is still not well-explored in Sora, so it is unclear what effect such features have had in the existing research.

We lack recordings in frames where a different word has been focused, as suggested by Gordon and van der Hulst, that eliminates this potential problem. A project to fully document the prosodic domains and units in Sora is being planned at present, so for now we must be limited to some preliminary observations. Thus, in a sense, this paper should be viewed as justification for a future research agenda.

Given these caveats, below we give example pitch traces of Sora words in comparison with Pnar and Lawa. Our goal in this section is to offer visible acoustic evidence regarding the correlation of pitch with what researchers on these languages say about the realization of stress. We will then see what generalizations can be made about word prosody in these languages, and whether any of these generalizations match the statements made by D&S. All the data presented in this pilot study is available for download on GitHub.[13]

We have discussed the correlates of stress for Sora (and Munda generally), but a few words about word stress in Pnar and Lawa are in order, to justify our acoustic examination of pitch. In Pnar (Ring 2015a, b), as in Khasian languages generally (Rabel 1961; Nagaraja 1985), word stress has been reported as iambic, where strong stress falls on the final syllable of a word. The primary correlate of word stress in Pnar is pitch (F_0), such that strong stress is marked primarily with higher relative pitch. Other correlates of stress include duration (longer = stronger stress) and intensity (increase = stronger stress).

In Lawa stress is also tied to the syllable, whereby final syllables of words receive the strongest stress (Mitani 1978; Blok 2013). The primary correlate of word stress in Lawa is reported as pitch, such that strong stress corresponds to higher pitch. Falling pitch is reported in Lawa on words in isolation, but generally not in phrases or words uttered in context – this observation led Mitani (1978) to conclude that lexical pitch was non-phonemic in Lawa, and that falling pitch contours were an effect of list intonation applied to lexical roots that were not considered phrases.

Our data sources are three recordings from fieldwork. The recordings in Sora and Pnar are stories of 5 minutes in length, while the Lawa story was about 2.5 minutes long. Each story was by a single male speaker, transcribed, translated, and interlinearized for the purpose of grammatical analysis. For this initial pilot study we re-annotated each of the stories, using Praat TextGrids to segment word and phrase utterances

[13] https://github.com/lingdoc/data_AA_prosody_paper

by syllable number. Data from a single speaker is not sufficient to generalize to all speakers of these languages, but we present it here due to the complete lack of such annotations in other AA research.

Below we present pitch traces in each language, focusing on words between 2-5 syllables. Unfortunately, due to constraints of time and our existing data, we could not control for word classes or constructional features. Instead, we present the pitch of one word of *n*-syllables for each language that is representative of the pitch of all such items generally, as well as a normalized pitch trace for words of *n*-syllables using a Praat script for extraction, normalization, and plotting of F_0 (Ring 2017), with automatic detection of F_0 to between 75 and 300 Hz. Table 4 presents the number of words by syllable length for each.

Table 4: *Number of words by syllable for Sora, Pnar, Lawa*

Syllables:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tokens**	**Syllables**
Sora:	54	198	168	111	44	23	2	1	601	1778
Pnar:	591	386	76	36	7	1	0	0	1097	1776
Lawa:	223	69	1	0	0	0	0	0	293	364

As can be seen from Table 4, the distribution of syllable types is not equal in the samples we investigated. While total number of syllables was comparable between Sora and Pnar, the Lawa sample contained fewer tokens and syllables overall, despite only being two minutes shorter. The samples also differ in terms of the relative number of syllables in a word – the Sora sample contains many more words with large syllable counts, while the Lawa sample contains mostly monosyllabic words and some disyllabic words.

3.1 Sora word pitch

Sora words in the single text we annotated include words between 1 and 8 syllables in length. Simply based on the number of words in each category (Table 2), we can see that two- and three-syllable words dominate, and that there is a significant downward trend (in terms of numbers of instances per category) for words of more than 5 syllables. Since our focus is to assess the claims of D&S in relation to the pitch contour of Munda words, below we highlight the pitch patterns of 2- to 5-syllable words within our Sora text.

3.1.1 Sora 2-syllable words

Two-syllable words in the Sora text we annotated make up the largest category of words, with 198 instances. While there is some variation, the primary pitch pattern of these words is of increasing pitch to the second syllable. In Figure 3 we see this pattern exemplified by the numeral *bagun* 'two'.

Figure 3: *Sora word pitch in 2-syllable word* bagun *'two'*

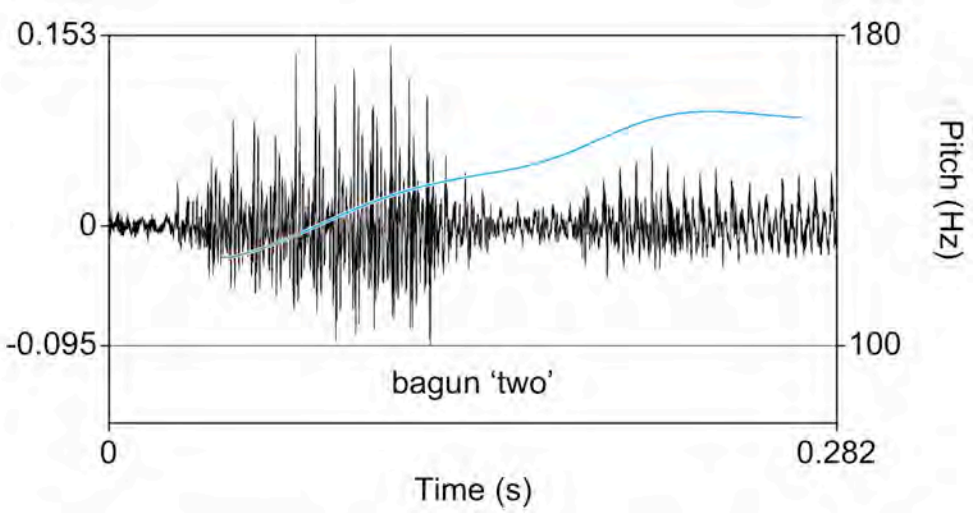
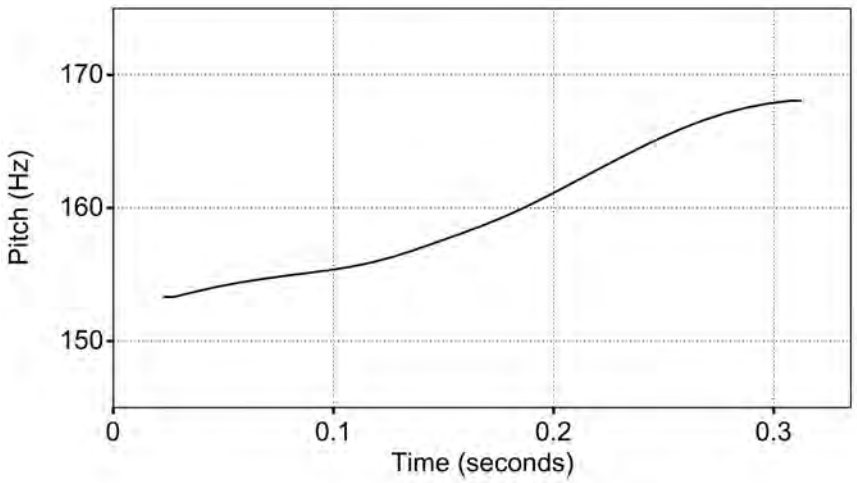


Figure 4 below displays the normalized pitch contour of 2-syllable words in our sample, which also rises across the duration of the two-syllable words, with an average length of 300 milliseconds.

Figure 4: *Sora normalized word pitch in 2-syllable words (198 instances)*

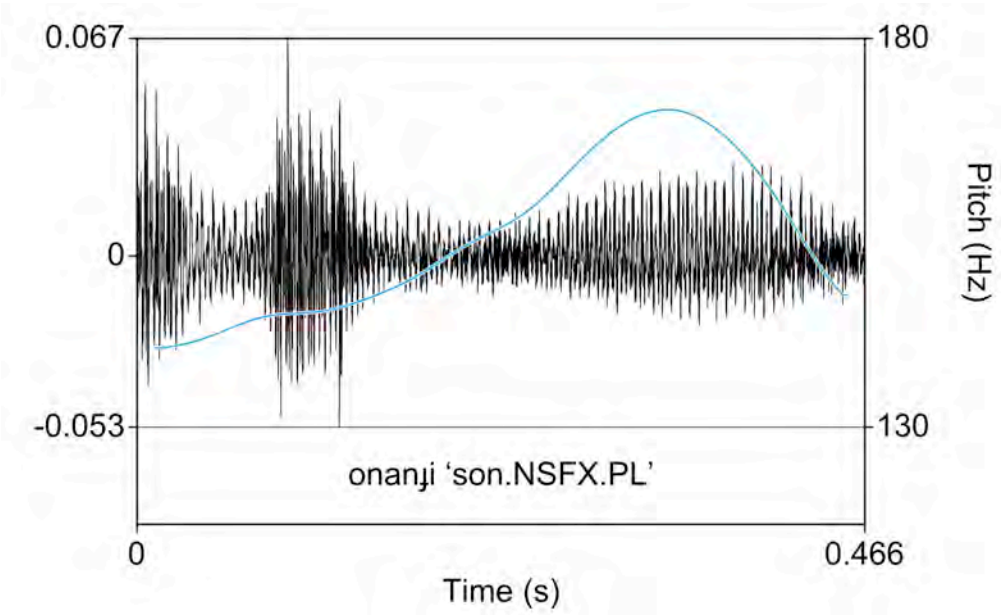


We acknowledge that a normalized pitch trace does not do justice to the range of variation in the data. However, this observation of the pitch in Sora two-syllable words aligns with Horo's (2017a, b) observations, and seems to indicate a clear tendency toward iambic stress in Sora disyllables.

3.1.2 Sora 3-syllable words

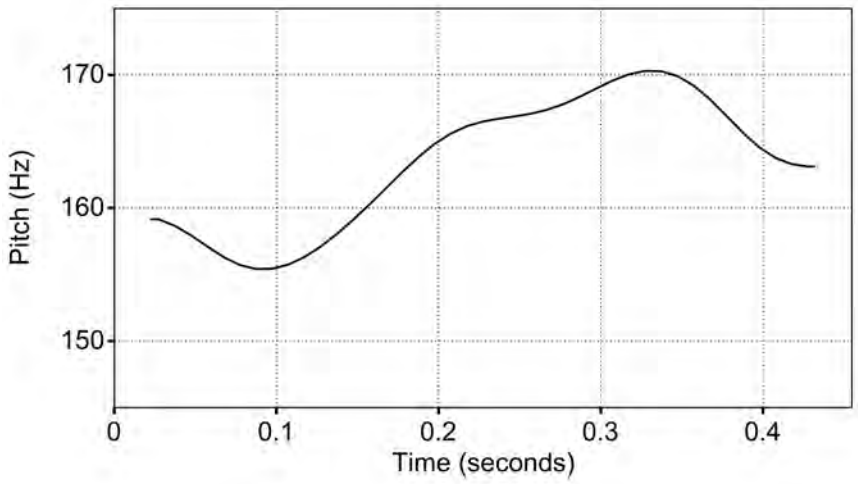
Three-syllable words are similarly well-represented in our Sora text, with 166 items. The majority of these words show a rise in pitch to the final syllable, as in the word *onaŋi* 'son.NSFX.PL' in Figure 5 below.

Figure 5: *Sora word pitch in 3-syllable word* onaŋi *'son.NSFX.PL'*



Creating a normalized pitch trace of all the words with three syllables in our text (Figure 6) illustrates that these words tend to have a rising pitch contour to the final syllable before a fall that corresponds with cessation of sound as the speaker prepares for the next word.

Figure 6: *Sora normalized word pitch in 3-syllable words (166 instances)*



3.1.3 Sora 4-syllable words

There are 113 instances in our text of Sora words with four syllables. Here we found more variation in pitch patterns. In Figure 7 below, of the word *adaŋgaɽa* 'PFX-young', we can see a high initial pitch followed by low pitch in the second and third syllable, rising to end with mid-level pitch on the final syllable. In Figure 8, however, the word *dakuliŋi* 'be.like.that' shows an initial mid-level pitch, high pitch on the second syllable, a sharp fall in the third syllable, and then a slight rise to the final syllable.

Figure 7: *Sora word pitch in 4-syllable word* adaŋgaɽa *'PFX-young'*

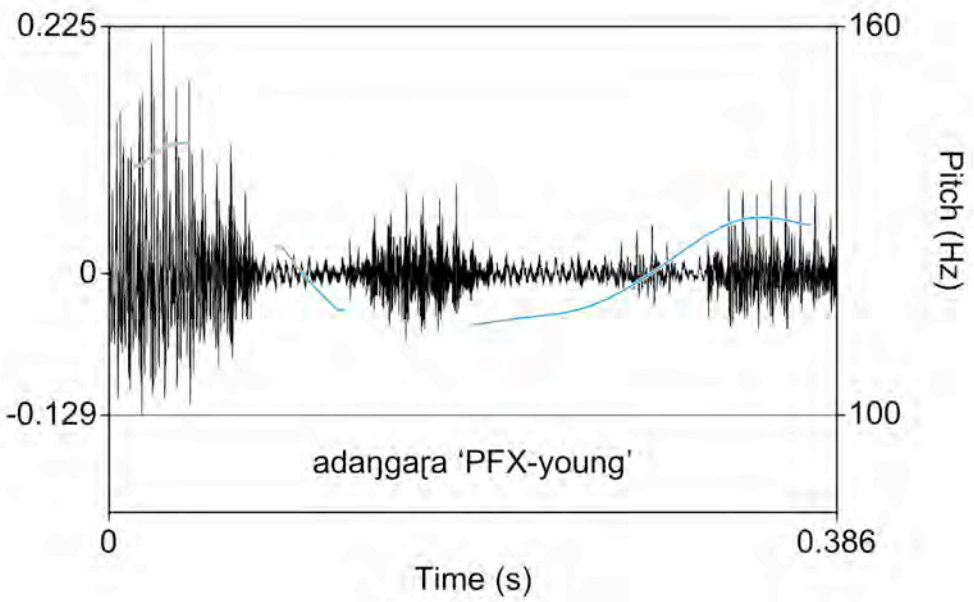
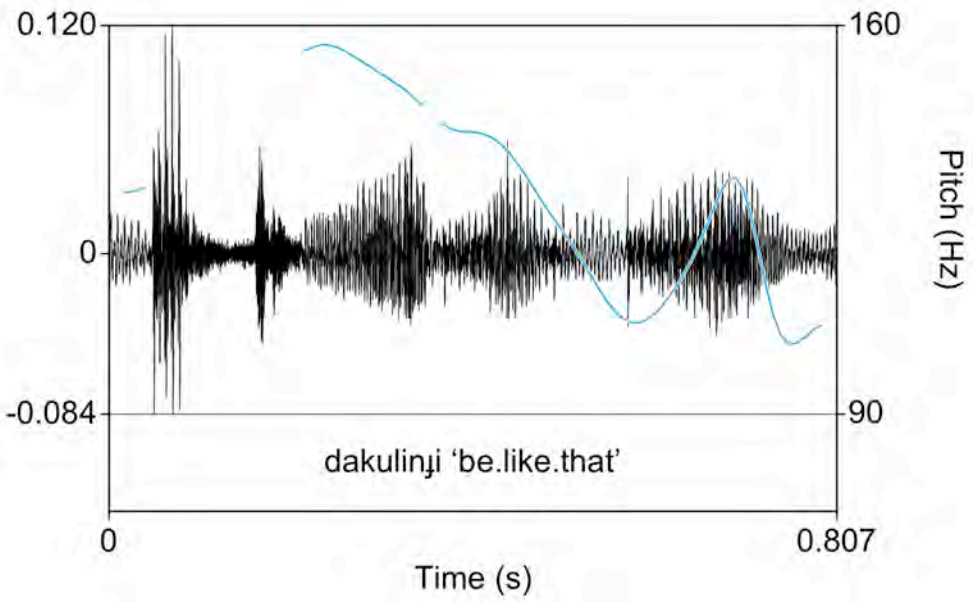
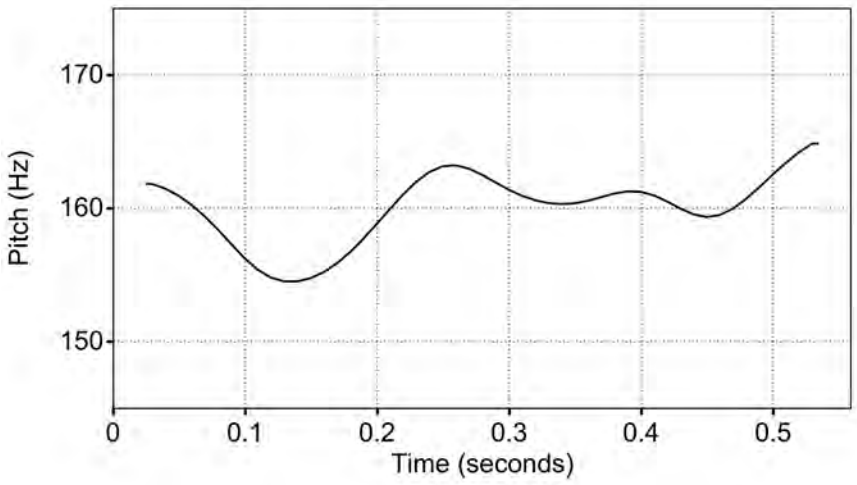


Figure 8: *Sora word pitch in 4-syllable word* dakulinɟi *'be.like.that'*



This is reflected in our normalized pitch trace (Figure 9), where we can see that the variability in four-syllable words, when normalized, leaves us with a pitch contour that shows no clear rises or falls and in relation to the previous normalizations of pitch is rather flat. We discuss this pattern below in our summary of Sora word pitch patterns.

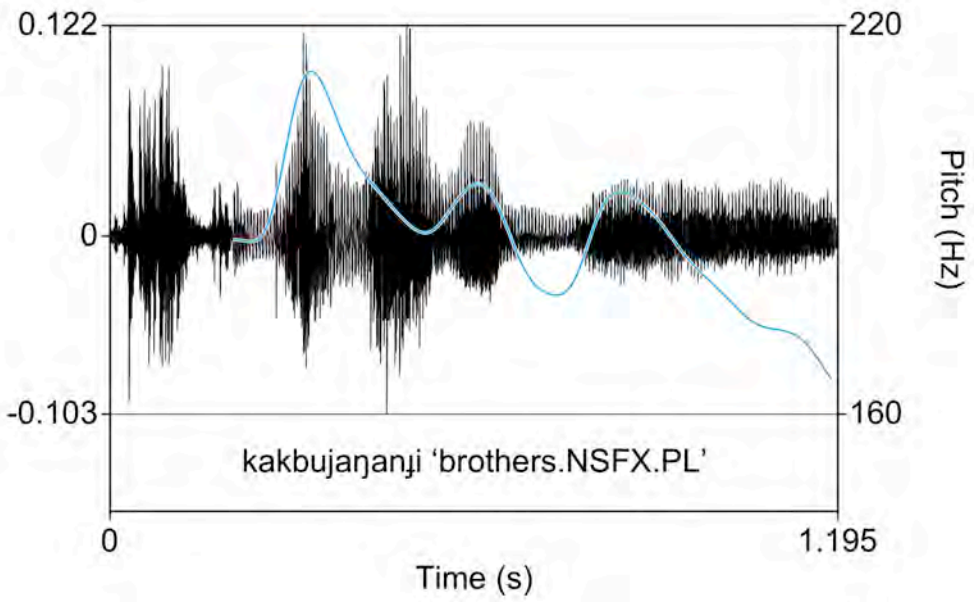
Figure 9: *Sora normalized word pitch in 4-syllable words (113 instances)*



3.1.4 Sora 5-syllable words

Interestingly, Sora words in our sample with five syllables do not show the same degree of variability as four-syllable words. Here we see a falling pitch contour, illustrated in Figure 10 by the Sora word *kakbujaŋanɟi* 'brothers.NSFX.PL'. Even more interesting, the highest pitch is generally recorded on the second or third syllable, and there is then a fall (punctuated by several height adjustments) from the highest point to the end of the word.

Figure 10: *Sora word pitch in 5-syllable word* kakbujaŋanɟi *'brothers.NSFX.PL'*

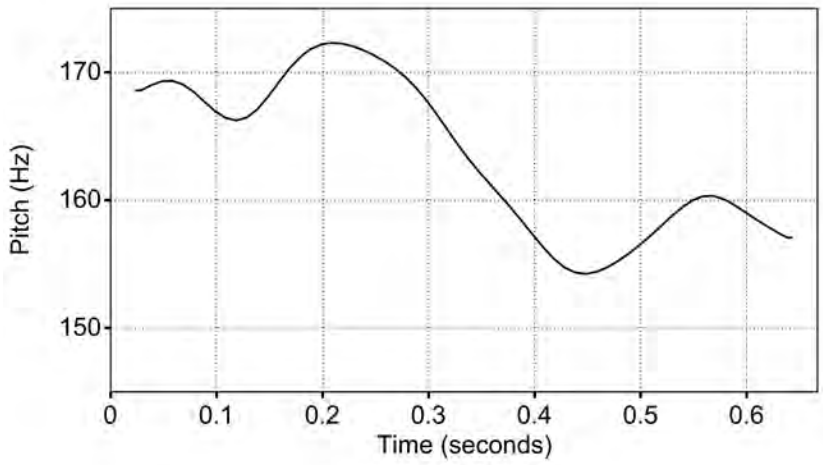


With only 44 instances of words with five syllables in our sample, the normalized pitch trace of these words (Figure 11) is not entirely conclusive, but illustrates an interesting trend. Here we see a slight rise to the second (or third) syllable, and then a fall to the end of the word, punctuated by slight peaks in pitch.

3.1.5 Sora word pitch summary

We can draw the following generalizations from our Sora data. First, a relatively stable pitch pattern seems to be present in words of two and three syllables, where pitch rises to the second or third syllable. Four- and five-syllable words, however, show no clear pitch prominence in four-syllable words, and for five-syllable words an initial rise in pitch to the second syllable before a fall in pitch to the end of the word.

Figure 11: *Sora normalized word pitch in 5-syllable words (44 instances)*



There are three possible reasons for the difference in pitch pattern for Sora words of four and five syllables. The first is that there are not enough tokens for pitch normalization to clearly reflect a pattern. The second is that longer words in Sora are more likely to have multiple affixes, with both prefixes and suffixes. This highlights a need to investigate the effect of affixation on realizations of word pitch/stress in Sora.

The third possibility is that such words are more likely to be composed of several phonological words, each with a prominent syllable. Longer words may thus show pitch prominence differences from other words of the same length, making generalizations more difficult. Words of fewer syllables are more likely to be

composed of single phonological words and show a dominant/primary pitch pattern. This also has implications for pitch in Sora phrases, discussed after we describe word pitch in the other languages.

3.2 Pnar word pitch

For Pnar, the majority of words in our sample were monosyllables. There were a large number of disyllables, and enough three- and four-syllable words to make observations. However, there were only seven five-syllable words, and a single six-syllable word, so below we describe words between 2 and 4 syllables.

3.1.1 Pnar 2-syllable words

The Pnar sample contains 386 words of two syllables. Pitch tends to increase in these words to the second syllable. In Figure 12 we see this in the word *ka=tʰaw* 'place'.

Figure 12: *Pnar word pitch in 2-syllable word* ka=tʰaw *'place'*

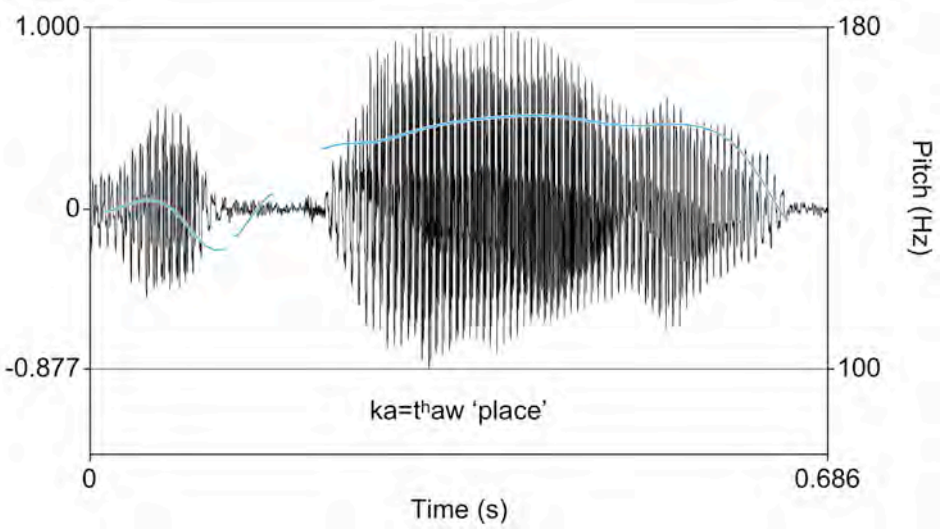
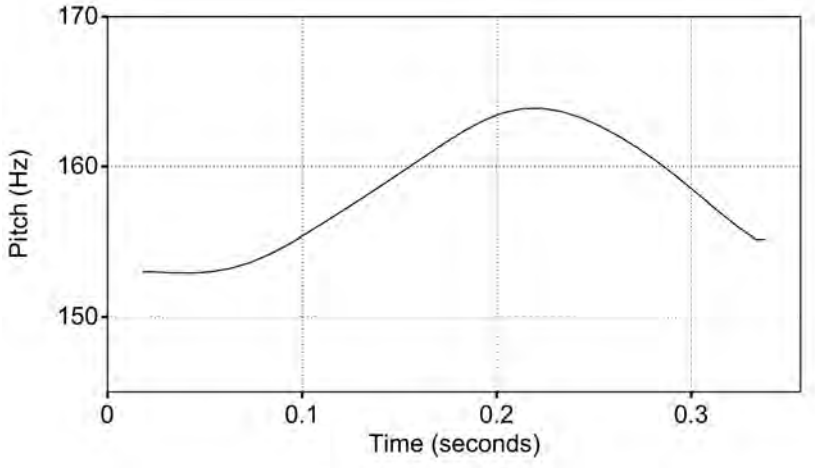


Figure 13 displays the normalized pitch contour of these words, which rise to the second syllable.

Figure 13: *Pnar normalized word pitch in 2-syllable words (386 instances)*



3.1.2 Pnar 3-syllable words

Three-syllable words are less well-represented in the Pnar text, with 76 items. The majority of these words also show a rise in pitch to the final syllable, as illustrated by the word *nɔŋhikaj* 'teacher' in Figure 14 below. Creating a normalized pitch trace of all the words with three syllables in our text (Figure 15) illustrates that these words tend to have a rising pitch contour to the onset of the final syllable before a fall to the end.

Figure 14: *Pnar word pitch in 3-syllable word* nɔŋhikaj *'teacher'*

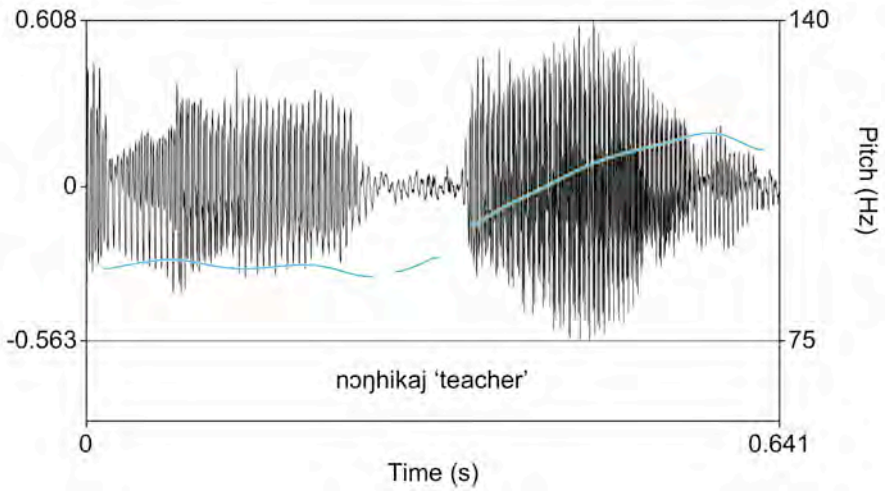
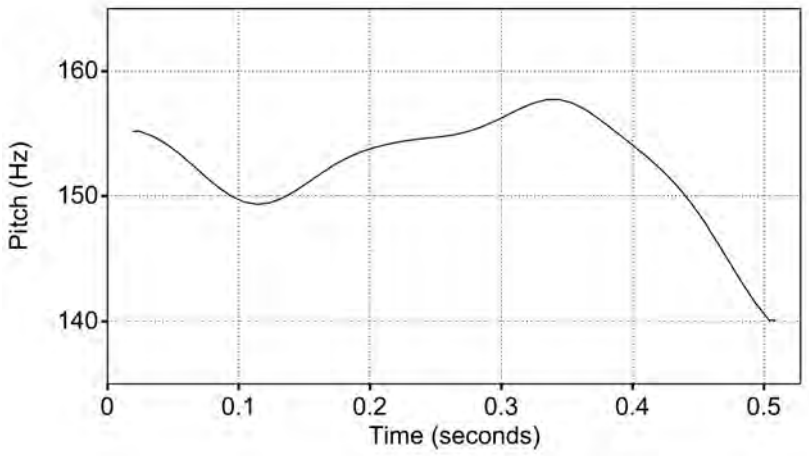


Figure 15: *Pnar normalized word pitch in 3-syllable words (76 instances)*



3.1.3 Pnar 4-syllable words

In Pnar words with four syllables we found more variation in pitch patterns. In Figure 16, *ka=ʤiŋjarap* 'the help' shows a rising pitch contour, while in Figure 17 *ka=ʤiŋjasen* 'the gathering' shows two peaks.

Figure 16: *Pnar word pitch in 4-syllable word* ka=ʤiŋjarap *'the help'*

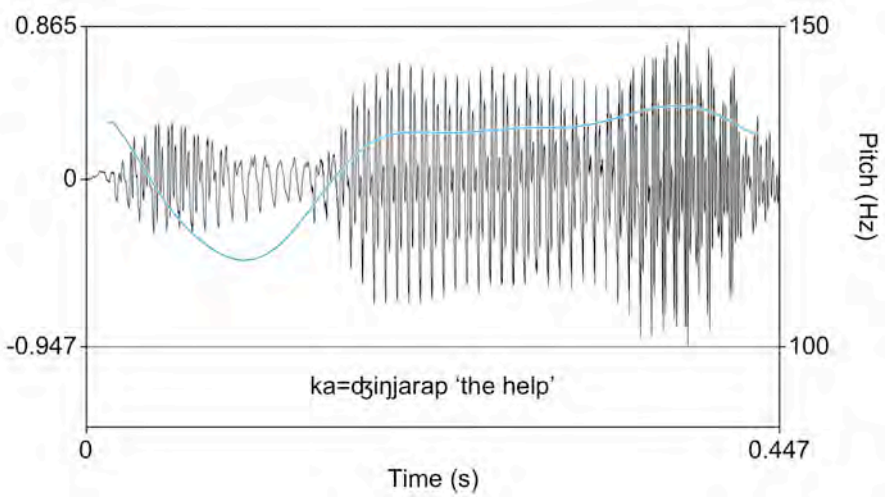


Figure 17: *Pnar word pitch in 4-syllable word* ka=ʤiŋjasen *'the gathering'*

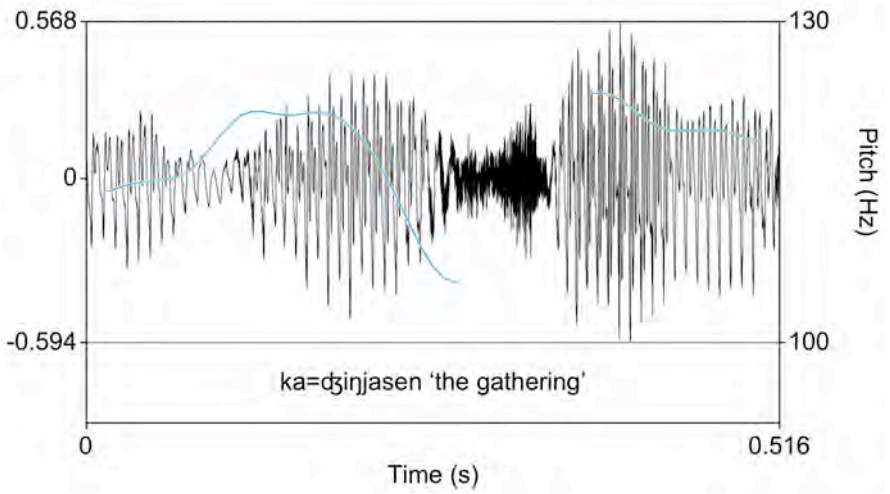
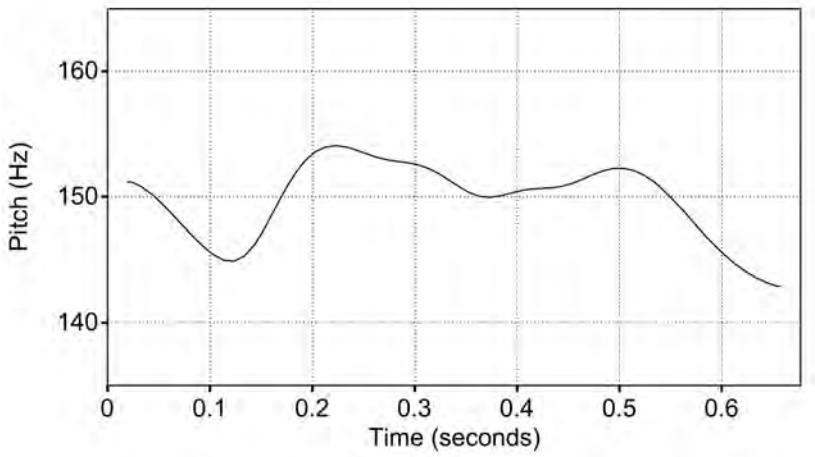


Figure 18: *Pnar normalized word pitch in 4-syllable words (36 instances)*

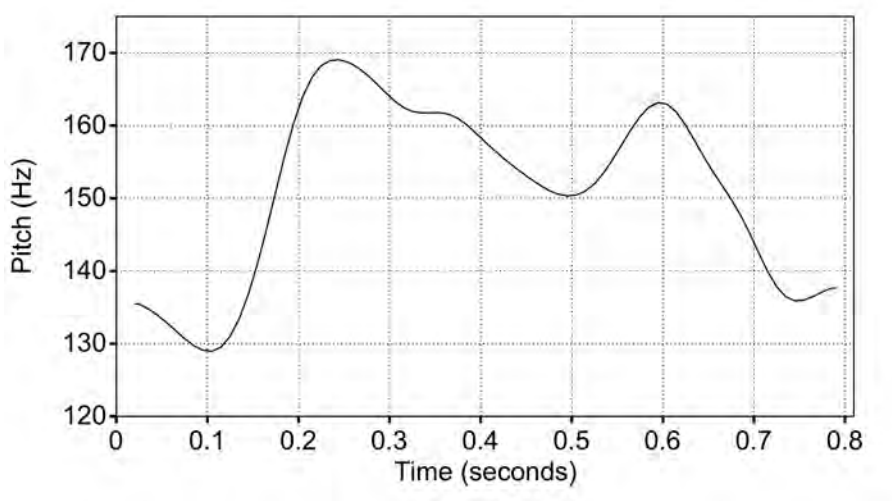


The normalized pitch trace of these four-syllable words (Figure 18), shows a pitch contour with two peaks. Occurrence of the two peaks where one would expect a second and fourth syllable onset may indicate that Pnar words of four syllables are composed of two (or more) phonological words, with pitch indicating relative prominence of their respective syllables.

3.1.4 Pnar 5-syllable words

Pnar words with five syllables in our sample are represented by only seven instances. In nearly all cases these are borrowed words (primarily from English). Exceptions to this rule are Pnar place names, of which there are two five-syllable examples in our data. In Figure 19 we show the normalized pitch pattern of Pnar five-syllable words. Here we see two pitch peaks, similar to those in four-syllable words, but there are too few tokens to offer meaningful information.

Figure 19: *Pnar normalized word pitch in 5-syllable words (7 instances)*



3.1.5 Pnar word pitch summary

What we can say for the pitch of Pnar words based on this data is somewhat similar to what can be said regarding Sora. In syllables of two and three syllables, there is a general rise in pitch to the final syllable. With four- and five-syllable words there is much more variation. While this may be clearer with more data, there are two possibilities that arise from our current observations in terms of variation in words with more syllables: the interaction of affixes and the alignment of phonological words with syllable structure.

Pnar is primarily a prefixing and procliticizing language (Ring 2015a,b). The majority of the four-syllable words in our sample have clitics and/or prefixes. The examples in Figures 16 and 17 are both nominalizations formed by a clitic and two affixes prefixed to a verb stem (Clitic=Pref-Pref-Stem). The variation in their pitch realizations may be due to whether the complex noun is treated by the speaker as a single element with final-syllable stress (*ka=ʤiŋ.ja.**rap***), or whether it is broken up into two elements, each with final-syllable stress (*ka=**ʤiŋ**-ja.**sen***). Further research is necessary to clarify this potential interaction.

3.3 Lawa word pitch

Our Lawa text shows a more drastic difference from Sora in terms of syllable numbers. Bearing in mind the shorter length of this text, the majority of words were of one syllable (223 instances), with fewer two-syllable words (69) and a single three-syllable word. Below we illustrate the pitch pattern of Lawa two-syllable words and the single three-syllable word.

3.1.1 Lawa 2-syllable words

In the Lawa words composed of two syllables in our dataset pitch increases to the second syllable. In Figure 20 we see this pattern exemplified by the word *kamtʰɔ* ‘after that’. Figure 21 below displays the normalized pitch contour of these words, which rises to the second syllable before a fall to the end of the word.

Figure 20: *Lawa word pitch in 2-syllable word* kamtʰɔ *'after that'*

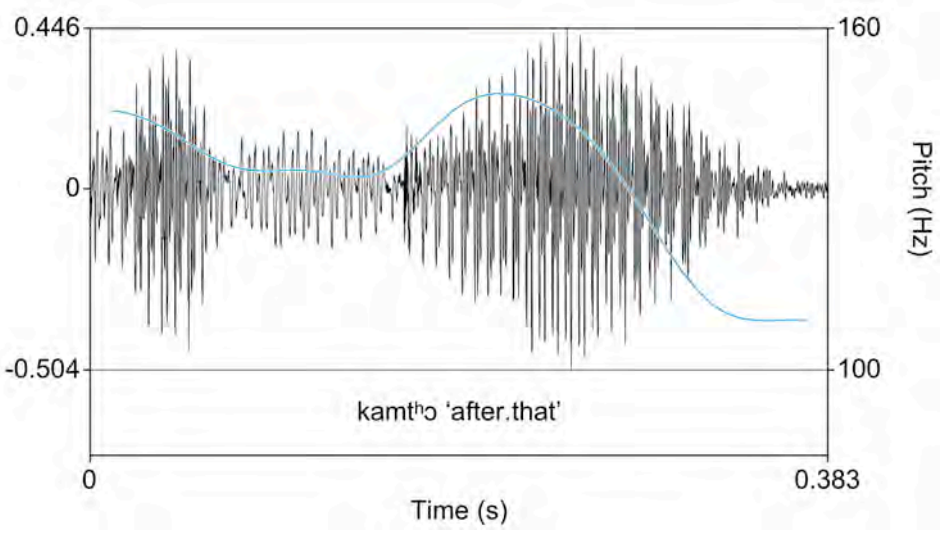
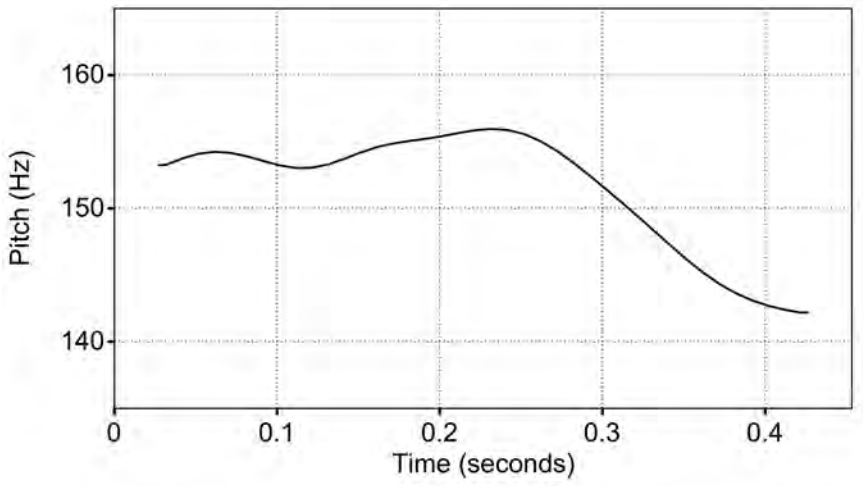


Figure 21: *Lawa normalized word pitch in 2-syllable words (69 instances)*



While we can see a peak at the beginning of the second syllable in the normalization, this is not an extremely steep increase from the beginning of the word. Unlike in Sora and Pnar, which both show an increase in pitch of 10-15 Hz across the two syllables on average, pitch in Lawa starts relatively high and increases by less than 5 Hz. However, given the dearth of two-syllable words in our Lawa sample it is difficult to treat this as a useful generalization.

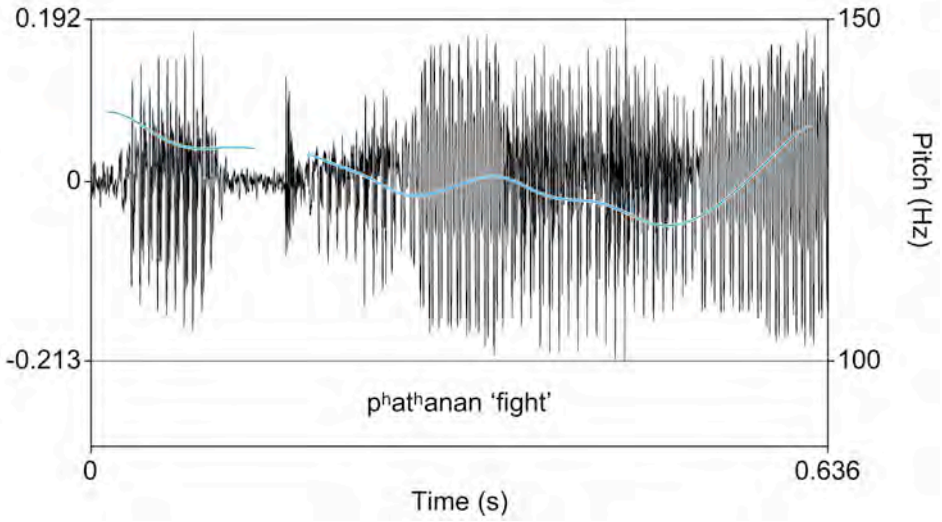
3.1.3 Lawa 3-syllable words

There is only one example of a three-syllable Lawa word in our sample, *pʰatʰanan* 'fight' (Figure 22), which our consultant said is a loanword from Northern Thai.[14] In this word, the pitch starts high and then falls, but

[14] Scholars we spoke to who are more well-versed in Thai (both historical and modern dialectal variation) are unsure exactly where this word may have been borrowed from, as a perusal of dictionaries does not reveal its presence. If a lexification, it is somewhat unusual in its use of aspirated /pʰ/ and /tʰ/ in relation to potential Thai source words, but we cannot pursue the phonology and exact source of this word at length here.

swings up at the end. Due to a lack of three-syllable words, it is difficult to draw any conclusions about this pattern, and since it is identified as a loanword it is entirely possible that the pitch pattern of the word was also borrowed from the source language.

Figure 22: *Lawa word pitch for 3-syllable word* pʰatʰanan *'fight'*



3.1.4 Lawa word pitch summary

As we see in the case of Lawa, our sample has too few two-syllable words to make good generalizations about the pitch pattern of these words. From individual analysis and normalization we only have some indication that Lawa words tend to start with high pitch and that there is a small (but possibly non-significant) increase in pitch to the start of the second syllable – probably best viewed as maintenance of the pitch target to the second syllable. While it is possible that more data will give clearer results, it is not likely to result in words with many more syllables. This makes observation of word pitch across syllables in Lawa difficult to compare with languages like Sora and Pnar, which have words with many more syllables. It is possible that comparing the pitch of phrases may overcome this challenge, but our data limits this kind of study, which we turn to briefly in the following subsection.

3.4 Instrumental analysis of phrase pitch

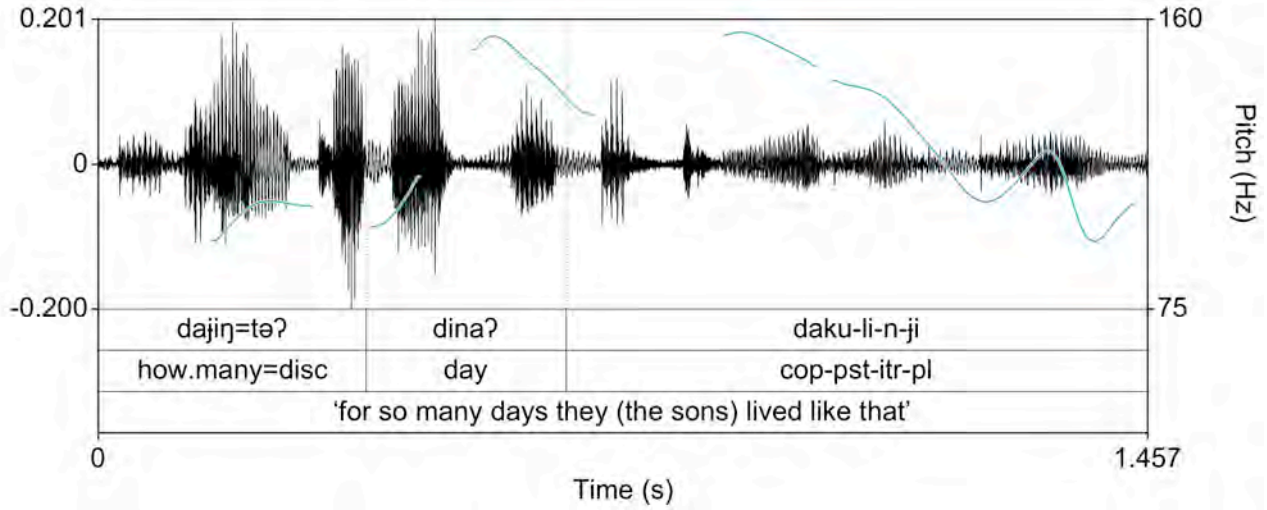
Phrasal pitch of these languages is beyond the ability of this paper to full deal with, though research in this area is a natural next step, particularly given the overarching claims of D&S. Here we present a very brief illustration of phrasal prosody in each language from the same data as above. Below, we annotate the prosodic pattern in a single sentence of each of the languages under investigation, with the acknowledgement that this is highly preliminary and deserves much more attention than we can give here.

Our criteria for choosing a sentence for display was that it be a complete, short sentence with a clear pitch trace. Due to the amount of variation in phrasal pitch between sentences, the difference in quality of the recordings, differences between our speakers, and the variation in phrase length, it was difficult to find sentences that were easily comparable between the texts in our sample. Some of these speaker differences are highlighted below, but as a result of this variation we provide the single pitch traces for illustrative purposes and use them to discuss potential rather than to make particular claims.

3.4.1 Sora phrase pitch

The Sora speaker in our recording produced sentences of the most variable length, with some short sentences like the one displayed here in Figure 23, and with long sentences of 10-12 words between 3-8 syllables long.

Figure 23: *Sample Sora pitch in phrase* daɟiŋ=təʔ dinaʔ dakuliŋi

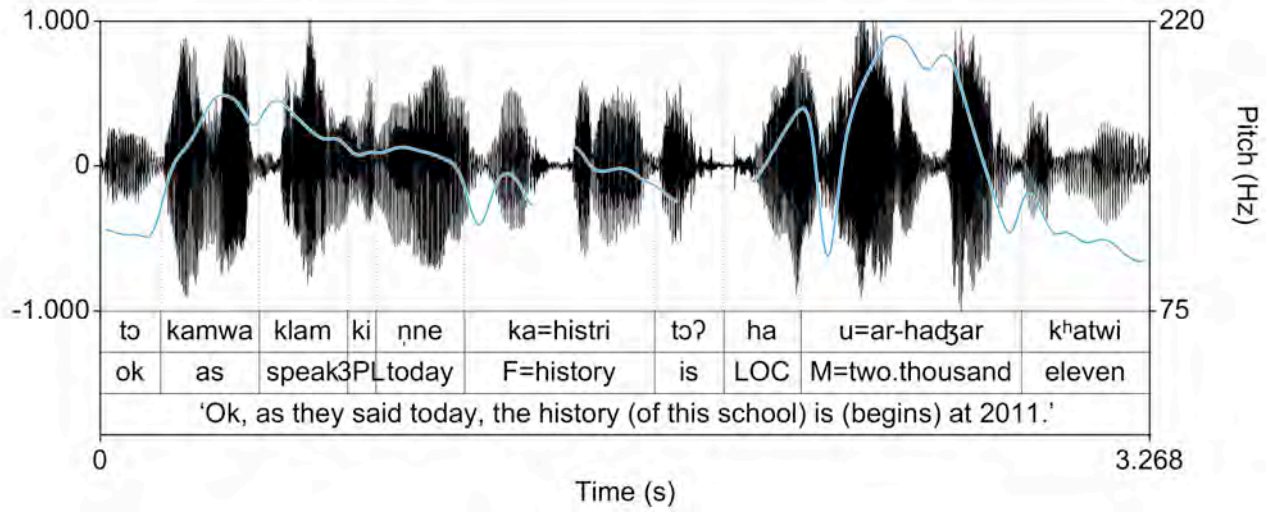


Pitch across this sentence is rather variable, with two high pitch realizations that occur in the relative middle of the clause. At the word level this figure does, however, illustrate the general trend in Sora for three-syllable words like *daɟiŋ=teʔ* and two-syllable words like *dinaʔ* to have rising intonation on the final syllable, though clearly the relative pitch of the two words is rather different.

3.4.2 Pnar phrase pitch

The Pnar speaker in the text we annotated spoke with more lengthy sentences than the Sora speaker, making it difficult to find a single short sentence that corresponded clearly to our Sora example. Figure 24 is a pitch trace of one of the shortest sentences he produced.

Figure 24: *Sample Pnar pitch in phrase* tɔ kamwa klam ki ŋne ka=histri tɔʔ ha u=arhadʒar kʰatwi



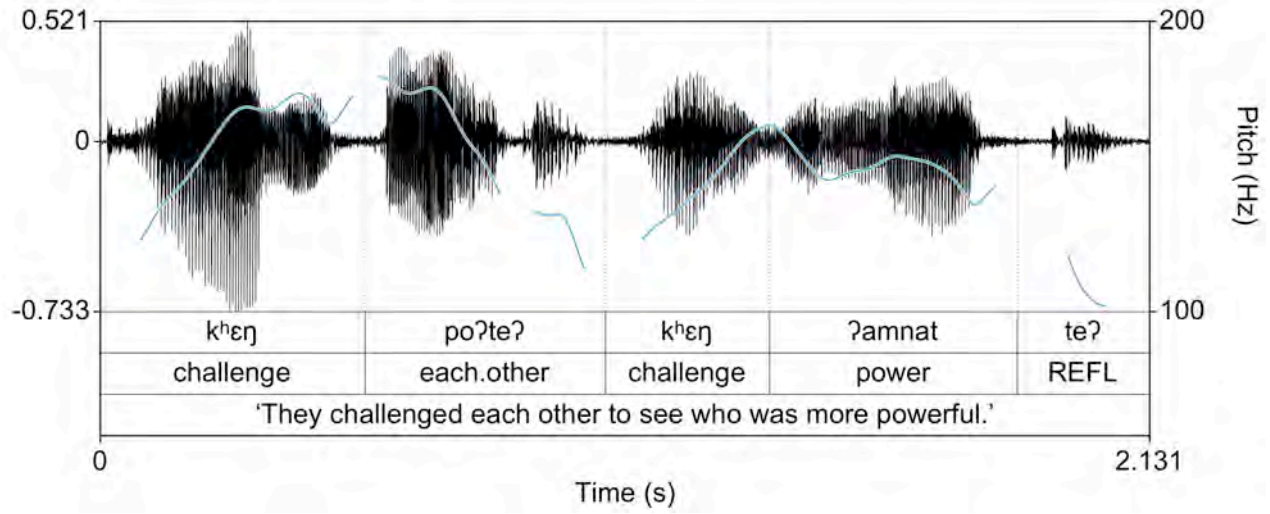
We can see that this relatively short sentence is 18 syllables long and shows multiple rising and falling pitch patterns, though with word-pitch rise and fall within these. The first large rise-fall corresponds to the adjunct phrase used to introduce the sentence (*tɔ, kamwa klam ki ŋne* 'Ok, as they said today'). The second rise-fall corresponds to the copula subject and the copular verb (*ka=histri tɔʔ* 'the history BE'), and the third rise-fall corresponds to the locative copula complement (*ha u=arhadʒar kʰatwi* 'LOC two thousand eleven').

It is not clear based on this single sentence whether rise-fall patterns in Pnar correspond to grammatical phrases, though given Rabel's observation for Khasi (section 2.3 above) this is a distinct possibility. Still, if we look more closely at the words within the sentences, we can see that apart from the end of the sentence and single-syllable words (which are more variable), high pitch in a word is associated more closely with the onset of the second syllable.

3.4.3 Lawa phrase pitch

The Lawa speaker spoke with many short sentences, one of which is given in Figure 25 below. Most sentences were between 3 and 6 words in length, with words being one or two syllables. As in Sora and Pnar, several peaks are observed in these sentences.

Figure 25: *Sample Lawa pitch in phrase* kʰɛŋ poʔteʔ kʰɛŋ ʔamnat teʔ



In Figure 25 the sentence is composed of five words, with 7 syllables in total. The larger rise-fall pitch contours correspond to two verbal clauses, *kʰɛŋ poʔteʔ* '(they) challenged each other' and *kʰɛŋ ʔamnat teʔ* '(they) challenged each others' power'. Both clauses could be considered grammatical phrases, though their proximity to each other means they are interpreted as conjoined. Unlike in the Pnar example, relative high pitch in words is consistently associated with the coda of single syllable words and with the second syllable of only one of the two-syllable words.

3.4.4 Summary of initial phrase pitch observations

While these three sentences by no means provide a comprehensive analysis of phrasal pitch patterns in the three languages, and we make no claims about the representative nature of these pitch traces, it is worth summarizing some observations. First, on the basis of these few sentences a rise and fall of pitch does seem to be tied to slightly different intonation units in each of these languages. In the Sora sentence, we see localized pitch rise relative to the word, with potential re-setting of pitch between words. In the Pnar example we see rise and fall of pitch largely corresponding to boundaries of grammatical units. In the Lawa example we also see rise and fall of pitch with grammatical units, but tied to different grammatical units than in the Pnar example. At the same time, in all three languages word-level pitch in running speech seems to follow a general LH tendency in two- and three-syllable words. Although more comprehensive analysis of sentence-level pitch needs to be made before the specific claims of D&S regarding phrasal prosody can accurately be assessed, already we can see that there are some differences in phrase pitch between Sora, Pnar, and Lawa that do not fit a Munda: Mon-Khmer division.

4. Discussion and conclusions

This paper has attempted several things. We have critiqued the work of Donegan and Stampe on the following fronts: their claims of rhythmic holism within Munda and "Mon-Khmer", their treatment of Sora prosody/stress, and their claims regarding "rising" and "falling" stress patterns, as reflected in pitch.

Regarding the final critique, we attempted to subject some of their claims to acoustic analysis in several Austroasiatic languages. In particular, we annotated words in three stories (in Sora, Pnar, and Lawa) and analyzed their pitch to see if this matched D&S's claims of a "falling" pitch pattern in Sora words and a "rising" pitch pattern in other Austroasiatic languages.

In acoustic analysis of our data we found that Sora and Pnar both showed an increase in pitch to the final syllable (or at least the onset of the final syllable) in two- and three-syllable words, and that Lawa shows a very slight rise to the final syllable of two syllable words. As noted above, D&S explicitly claim that Sora is a "falling-accented" language, and as Donegan (1993:10) states, "Falling-accented languages... mark accent, if at all, with pitch." Our acoustic findings directly contradict this claim, showing that Sora two- and three-syllable words have *rising* (not "falling") pitch, a feature the language shares with Pnar and possibly other Austroasiatic languages such as Lawa.

Sora and Pnar words with more syllables show greater variation in this pattern, with multiple peaks in pitch across the word. Sora and Pnar both show two peaks in some four-syllable words, possibly corresponding to the second and fourth syllables. This may indicate that these longer words are composed of multiple phonological words. In five-syllable words there is again variation in syllable peaks, and while there is too little data for a Pnar generalization (and possibly too little for Sora), in Sora we can observe a general fall from the second syllable of the word, with a small peak before the end. Words in Sora with more syllables, then, seem to correspond to the "falling" pitch contour of D&S, while words with fewer do not.

The single example sentence that we presented in each of the languages we investigated show multiple rises and falls of pitch which correspond to various units depending on the language. Other sentences that we could have presented for each language show somewhat different pitch patterns. That multiple pitch patterns exist in these languages does not easily square with statements by D&S regarding a single, holistic pattern of phrase pitch for the languages in question, much less a unified phrase and word pitch pattern in Munda vs Mon-Khmer. Generalizations may be possible with more data and more carefully controlled data, but so far this has not been done for Austroasiatic languages.

Given the results of our acoustic investigation, we tentatively suggest that iambic stress for (at least) two- and three-syllable words is a feature of Austroasiatic languages generally, including Sora within Munda and possibly all Munda languages. To claim otherwise creates a false dichotomy between South Asian Austroasiatic languages on the one hand, and all other Austroasiatic languages. Indeed, upon more careful examination the major claims D&S made can be simplified to:

1) an observation that the Western Austroasiatic languages (in eastern India) are different prosodically from those in the East (mainland SEA), such that pitch falls to the end of the unit in Munda languages while pitch rises to the end of the unit in 'Mon-Khmer' languages.
2) that this is a result of prosodic restructuring, leading to agglutination in Munda languages.

Our data suggests that claim #1 is largely false for words in these languages of 2-3 syllables, which show a primary rise in pitch across the West/East divide. There are also indications from observation of phrasal pitch that units composed of 4 or more syllables in Sora, Pnar, and Lawa, have consistent rise-fall patterns that in passing seem to be determined by function, though whether such patterns have similar functions in these languages is yet to be ascertained.

Claim #2 is a more specific claim about the role of prosody in encouraging certain kinds of word formation. However, given that there is little evidence for claim #1, it is difficult to see how the second claim could be assessed in regard to Munda languages. It is also not fully clear from the work of D&S how exactly a change in prosodic pitch could/would condition syllables to join more closely into agglutinating word structures in these languages. We would expect if claim #2 held true for languages descended from Proto-Austroasiatic that Pnar would also show these agglutinating word types (given its similarity to Sora in our analysis above), but in fact the number of syllables in a typical Pnar word is generally fewer than in a Sora word. This suggests that a refinement of the claim is in order.

Further, in its broad strokes the first claim is rather similar to the "Indosphere" and "Sinosphere" distinction proposed by James Matisoff, a general West/East geographical divide in terms of the linguistic typology of South and South-East Asia. Matisoff (1991:485-486) suggested that:

> "[It] is convenient to refer to the Chinese and Indian spheres of influence as the 'Sinosphere' and the 'Indosphere'... Some languages are firmly in one or the other... the Munda and Khasi branches of Austroasiatic and the Kamarupan [*sic*] branch of TB are Indospheric; while... the Loloish branch of TB and the Viet-Muong branch of Mon-Khmer are Sinospheric… Whatever their genetic affiliations, the languages

> of the ST area have undergone massive convergence in all areas of their structure – phonological, grammatical, and semantic."

By separating languages into the two Indospheric and Sinospheric camps, Matisoff makes a similar claim to Donegan and Stampe, albeit not in terms of a single organizing principle for the changes found in two groups of related languages. Instead, the claim seems to be that language contact across wide geographical areas is responsible for the changes by which languages from different families appear similar in many ways. While the observation is related to observations regarding 'spread' and 'accretion' zones around the globe (Nichols 1992), it seems a bit of an oversimplification here – a convenient generalization that does not necessarily help to illuminate the actual process of change and development of the individual languages in question.
Post (2011) makes a similar point, returning to the question of prosody as a potential source of similarity for languages in North-East India (where Khasian languages are also located). He notes (p. 218) that:

> "Diffusion of structural features requires more than simply contact: it requires learning and understanding: bilingualism and interaction... By contrast, prosodic diffusion requires little more than contact; contact, that is, followed by... imitation; not understanding... Through imitation of the observable behavior of others, prosodic features can, from a particular area of concentration, spread over vast geographical distances, bringing languages into close alignment with respect to some aspects of their linguistic profiles, despite their speakers never in fact having come into contact with one another."

While we agree with other parts of Post's paper,[15] the problem with this particular statement is that it does not clearly align with tendencies in the prosody of L2 speakers. In fact, non-native L2 speakers are easily identified by their accent, a large portion of which is prosody. This is enough of a concern that a growing area of second-language acquisition research is devoted to teaching correct stress and other prosodic patterns of a language (see Jung et al. 2017; Liu 2017; see also Xu 2011, 2012). The transfer of word and sentence prosody from a second language (L2) into a first language (L1), via what Post calls 'imitation', seems to actually require sustained bilingualism and multiple generations, just like any other feature of language claimed to spread via contact (though it may spread more easily than other aspects of language). This is an area that has not been well-studied; an edited volume by Delais-Roussarie et al. (2015) offers some insight regarding prosodic features that can spread due to contact. Many authors in this volume show that prosody spreads from the substrate (L1) to the superstrate (L2) rather than from the L2 into the L1 via 'imitation'.

This direction of spread suggests that speakers tend to maintain the prosodic system they grew up with rather than 'imitating' the prosodic system of an L2, such that an appeal to 'imitation' does not provide an explanation for why Munda and Khasian speakers might show similar prosodic patterns as neighboring languages. Other research shows that phonological features such as 'focus prosody' (see Wang et al. 2011 and papers in the same volume) are also unlikely to spread via contact, and that genealogy is a significant predictor for the kind of phonological domains found in a language (Bickel et al. 2009:72). If Munda and Khasian languages indeed present a similar prosodic profile as their Dravidian and Indo-Aryan neighbors in some aspects, this may actually be evidence for sustained bilingualism and contact at some point in history.

In any case, such sweeping statements as these need better evidence than we have seen thus far, and while the work of D&S is commendable for its attempt to say something meaningful about prosody in Austroasiatic and its relation to the history of the family, their main claims do not hold up well to scrutiny. We have attempted to provide actual speech data with our pilot study (a major shortcoming of D&S's work), but acknowledge that we do not have conclusive proof regarding word and phrase prosody in the three languages in question. What is needed are more language-specific studies within Austroasiatic that properly control for effects of word and phrase type, grammatical effects, and focus effects, as well as identifying actual acoustic correlates of stress at the word and the phrase levels. This will do the most toward advancing the study of stress (and prosody more generally) within Austroasiatic, so that we can begin to compare prosodic features of these languages with those features of their linguistic neighbors and relatives to tease out possible contact influence, historical inheritance, and historical development.

---

[15] We agree, for example, with the main point of his paper, that the terms 'Indosphere' and 'Sinosphere' are problematic, "not only because of the possibly incorrect characterization of the proximal cause of typological alignment that they provide, but because of the pre-historical dominant/subordinate population relationships that they imply, for which – in several cases at least – no evidence whatsoever is available." (Post 2011: 219)

References

Anderson, Gregory D. S. 2008. Gtaʔ. In The Munda Languages, ed. by Gregory D. S. Anderson, 682–763. Routledge Language Family Series. London: Routledge.

Anderson, Gregory D. S. 2015. Overview of the Munda languages. In The Handbook of Austroasiatic Languages, Volume 1, ed. by Mathias Jenny and Paul Sidwell, 364–414. Leiden: Brill.

Anderson, Gregory D. S. 2016. Do Koraput Munda, Lower Munda, and even South Munda really exist? Once more on the still unresolved classification of the Munda languages. In Multilingualism and Multiculturalism: Perceptions Practices and Policy, ed. by Supriya Pattanayak, Chandrabhanu Pattanayak and Jennifer Bayer, Delhi: Orient Blackswan.

Anderson, Gregory D. S. Forthcoming. Proto-Munda in Austroasiatic comparative and South Asian areal perspectives. In Proto-Austroasiatic Syntax, ed. by Mark Alves, Mathias Jenny, and Paul Sidwell, Leiden: Brill.

Anderson, Gregory D. S. In preparation-a. The Gtaʔ Language. Plains Gtaʔ and Hill Gtaʔ Texts, Lexicon and Grammar.

Anderson, Gregory D. S. in preparation-b. The Gutob and Remo Languages. Comparative lexicon, phonology and grammar with texts.

Anderson, Gregory D. S. and K. David Harrison. 2008. Remo. In The Munda Languages, ed. by Gregory D. S. Anderson, 557–632. Routledge Language Family Series. London: Routledge.

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

Anderson, Gregory D. S. and Felix Rau. 2008. Gorum. In The Munda Languages, ed. by Gregory D. S. Anderson, 381–433. Routledge Language Family Series. London: Routledge.

Barker, Philip R. 1953?/no date-ms. The phonemes of Korowa. Unpublished manuscript. Seattle, Washington.

Beckman, Mary. 1986. Stress and non-stress accent. Dordrecht: Foris.

Bhattacharya, Sudhibhushan. 1968. A Bonda Dictionary. Building Centenary and Silver Jubilee Series 18. Poona [Pune]: Deccan College.

Bickel, Balthasar and Fernando Zúñiga. 2017. The 'word' in polysynthetic languages: phonological and syntactic challenges. In The Oxford handbook of Polysynthesis, ed. by Michael Fortescue, Marian Mithun, and Nicholas Evans, 158–85. Oxford: Oxford University Press.

Bickel, Balthasar, Kristine Hildebrandt, and René Schiering. 2009. The distribution of phonological word domains: a probabilistic typology. In Phonological domains: universals and deviations, ed. by Janet Grijzenhout and Bariş Kabak, 47–75. Berlin: Mouton De Gruyter.

Biligiri, Hemmige S. 1965. Kharia: phonology, grammar, vocabulary. Poona: Deccan College.

Blok, Greg. 2013. A descriptive grammar of Eastern Lawa. MA thesis, Payap University, Thailand.

Boersma, Paul and David Weenink. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.30, http://www.praat.org/

Bybee, Joan L., Paromita Chakraborti, Dagmar Jung, and Joanne Scheibman. 1998. Prosody and segmental effect: Some paths of evolution for word stress. Studies in Language, 22 (2):267–314.

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. PhD dissertation, Georgetown University.

Cruttenden, Alan. 1997. Intonation. 2nd edition. New York: Cambridge University Press.

Dasgupta, Dipankar. 1978. Linguistic Studies in Juang, Kharia Thar, Lodha Mal-Pahariya, Ghatoali Pahariya. Calcutta: Anthropological Survey.

Dasgupta, Probal. 2003. Bangla. In The Indo-Aryan Languages, ed. by George Cardona and Dhanesh Jain 351–90. London: Routledge.

Deepadung, Sujartilak, Ampika Rattanapitak and Supakit Buakaw. 2015. Dara'ang Palaung. In The Handbook of Austroasiatic Languages, Volume 2, ed. by Mathias Jenny and Paul Sidwell, 1065–103. Leiden: Brill.

Delais-Roussarie, Elisabeth, Mathieu Avanzi, and Sophie Herment (eds.). 2015. Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations. Berlin: Springer-Verlag.

DePaolisa, Rory A., Marilyn M. Vihmann, and Sari Kunnaric. 2008. Prosody in production at the onset of word use: A cross-linguistic study. Journal of Phonetics, 36:406–22.

Diffloth, Gérard. 2005. The contribution of linguistic palaeontology to the homeland of Austro-asiatic. In The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics, ed. by Laurent Sagart, Roger Blench, and Alicia Sanchez-Mazas, 79–82. London: Routledge.

Donegan, Patricia J. 1993. Rhythm and vocalic drift in Munda and Mon-Khmer. Linguistics of the Tibeto-Burman Area 16(1):1–43.

Donegan, Patricia J. and David Stampe. 1983. Rhythm and the holistic organization of language structure. In Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax, ed. by John F. Richardson, Mitchell Marks, and Amy Chukerman, 337–53. Chicago: Chicago Linguistic Society.

Donegan, Patricia J. and David Stampe. 2002. South-East Asian features in the Munda languages: Evidence for the analytic-to-synthetic drift of Munda. Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society: Special Session on Tibeto-Burman and Southeast Asian Linguistics 28(2):111–120.

Donegan, Patricia J. and David Stampe. 2004. Rhythm and the synthetic drift of Munda. In Yearbook of South Asian Languages and Linguistics, ed. by Rajendra Singh, 3–36. Berlin: De Gruyter Mouton.

Emeneau, Murray B. 1954. Linguistic prehistory of India. American Philosophical Society, 98:282–92.

Emeneau, Murray B. 1956. India as a linguistic area. Language, 32(1):3–16.

Enfield, Nick. 2005. Areal linguistics and mainland Southeast Asia. Annual Review of Anthropology, 34:181–206.

Fernandez, Frank. 1968. A Grammatical Sketch of Remo. PhD dissertation, University of North Carolina.

Ghosh, Arun. 2008. Santali. In The Munda Languages, ed. by Gregory D. S. Anderson, 11–98. Routledge Language Family Series. London: Routledge.

Gordon, Matthew and Harry van der Hulst. (to appear). Word Stress. In The Handbook of Prosody, ed. by Carlos Gussenhoven and Aoju Chen. Oxford: Oxford University Press.

Gordon, Matthew, Carmen Jany, Carlos Nash and Nobutaka Takara. 2010. Syllable structure and extrametricality: A typological and phonetic study. Studies in Language, 34:131–66.

Gordon, Matthew and Timo Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. Linguistics Vanguard, 3(1). doi:10.1515/lingvan-2017-0007.

Griffiths, Arlo. 2008. Gutob. In The Munda Languages, ed. by Gregory D. S. Anderson, 633–81. Routledge Language Family Series. London: Routledge.

Hayes, Bruce. 1995. Metrical Stress Theory: Principles and Case Studies. Chicago: University of Chicago Press.

Henderson, Eugénie J. A. 1952. The main features of Cambodian pronunciation. Bulletin of the School of Oriental and African Studies, 14:159–74.

Horo, Luke. 2017a (ms). Phonetic comparison of Orissa Sora and Assam Sora. Presented at the 1st International Conference on Munda Languages and Linguistics, Deccan College Pune, March 2017.

Horo, Luke. 2017b. A Phonetic Description of Assam Sora. PhD dissertation, Indian Institute of Technology, Guwahati.

Horo, Luke and Priyankoo Sarmah. 2014. An acoustic analysis of vowel system size in Assam Sora. Presented at Himalayan Languages Symposium, Singapore. 16–18 July 2014.

Horo, Luke and Priyankoo Sarmah. 2015. Acoustic analysis of vowels in Assam Sora. In Konnerth, Linda, Stephen Morey, Priyankoo Sarmah, and Amos Teo (eds.), North East Indian Linguistics, Vol. 7, 69–86. Australian National University, Canberra: Asia-Pacific Linguistics.

van der Hulst, Harry G. 2012. Deconstructing stress. Lingua, 122:1494–521.

van der Hulst, Harry G. 2014. Word Stress: past, present and future. In Word Stress: Theoretical and typological issues, ed. by Harry G. van der Hulst, 3–55. Cambridge: Cambridge University Press.

Hyman, Larry M. 2006. Word-prosodic typology. Phonology, 23(2):225–57.

Jenny, Mathias, Paul Sidwell, and Mark Alves. Forthcoming. Position paper: Austroasiatic syntax in diachronic and areal perspective. In Austrosasiatic syntax in diachronic and areal perspective, ed. by : Mark Alves, Mathias Jenny, and Paul Sidwell. Leiden: Brill.

Jenny, Mathias and Paul Sidwell (eds.). 2015. The Handbook of Austroasiatic languages (2 Volumes). Leiden: Brill.

Jun, Sun-Ah. 2014. Prosodic typology: by prominence type, word prosody, and macro-rhythm. In Prosodic Typology II: The Phonology of Intonation and Phrasing, ed. by Sun-Ah Jun, 520–39. Oxford: Oxford University Press.

Jung, YeonJoo, YouJin Kim, and John Murphy. 2017. The role of task repetition in learning word-stress patterns through auditory priming tasks. Studies in Second Language Acquisition, 39(2):319–46.

Kawaguchi, Yuji, Ivan Fonagy, and Tsunekazu Moriguchi (eds.). 2006. Prosody and Syntax: Cross-linguistic perspectives. New York: John Benjamins.

Keane, Elinor. 2004. Tamil. Journal of the International Phonetic Association 34(1):111–16.

Keane, Elinor. 2014. The intonational phonology of Bangladeshi Standard Bengali. In Prosodic Typology II: the Phonology of Intonation and Phrasing, ed. by Sun-Ah Jun, 81–117. Oxford: Oxford University Press.

Khan, Sameer ud Dowla. 2016. The intonation of South Asian languages: towards a comparative analysis. Formal Approaches to South Asian Languages 6 Proceedings, 23–36.

Khan, Sameer ud Dowla. 2008. Intonational phonology and focus prosody of Bengali. PhD dissertation, UCLA.

Kobayashi, Masato and Ganesh Murmu. 2008. Kera? Mundari. In The Munda Languages, ed. by Gregory D. S. Anderson, 165–94. Routledge Language Family Series. London: Routledge.

Kruspe, Nicole, Niclas Burenhult and Ewelina Wnuk. 2015. Northern Aslian. In The Handbook of Austroasiatic languages, Volume 2, ed. by Mathias Jenny and Paul Sidwell, 419–74. Leiden: Brill.

Ladd, D. Robert. 1996. Intonational Phonology. Cambridge: Cambridge University Press

Ladd, D. Robert. 2001. Intonation. In Language Typology and Language Universals: An International Handbook, ed. by Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, 1380–90. Berlin: Mouton De Gruyter.

Langendoen, Terence. 1963. Mundari Phonology. Unpublished manuscript. Cambridge, Massachusetts.

Li, Jinfang and Yongxian Luo. 2015. Bugan. In The Handbook of Austroasiatic languages, Volume 2, ed. by Mathias Jenny and Paul Sidwell, 1033–62. Leiden: Brill.

Liu, Dan. 2017. The acquisition of English word stress by Mandarin EFL learners. English Language Teaching, 10(12):196–201.

Matisoff, James A. 1991. Sino-Tibetan linguistics: Present state and future prospects. Annual Review of Anthropology, 20:469–504.

Mitani, Yasuyuki. 1978. Phonological studies of Lawa: description and comparison. PhD dissertation, Cornell University.

Nagaraja, Keralapura S. 1985. Khasi, A Descriptive Analysis. PhD dissertation, Deccan College, Pune.

Neukom, Lukas. 2001. Santali. Languages of the World/Materials, 323. München: Lincom.

Nichols, Johanna. 1992. Linguistic diversity in space and time. Chicago: University of Chicago Press.

O'Brien, Mary G., Carrie N. Jackson, and Christine E. Gardner. 2014. Cross-linguistic differences in prosodic cues to syntactic disambiguation in German and English. Applied Psycholinguistics, 35 (1):27–70.

Osada, Toshiki. 2008. Mundari. In The Munda Languages, ed. by Gregory D. S. Anderson, 99–164. Routledge Language Family Series. London: Routledge.

Patnaik, Manideepa. 2008. Juang. In The Munda Languages, ed. by Gregory D. S. Anderson, 508–56. Routledge Language Family Series. London: Routledge.

Peterson, John M. 2011. Grammar of Kharia. Leiden: Brill.

Pinnow, Heinz-Jürgen. 1963. The position of the Munda languages within the Austroasiatic language family. In Linguistic Comparison in Southeast Asia and the Pacific, ed. by Harry L. Shorto, 140–52. London: SOAS.

Pinnow, Heinz-Jürgen. 1966. A comparative study of the verb in the Munda languages. In Studies in Comparative Austroasiatic Linguistics, ed. by Norman H. Zide, 96–193. Indo-Iranian Monographs, V, The Hague: Mouton.

Post, Mark W. 2011. Prosody and typological drift in Austroasiatic and Tibeto-Burman: Against "Sinosphere" and "Indosphere". In Austroasiatic Studies: papers from ICAAL4, ed. by Sophana Srichampa, Paul Sidwell, and Kenneth Gregerson, 198–221. Mon-Khmer Studies Journal Special Issue 3, SIL International.

Presmsrirat, Suwilai and Nattamon Rojankul. 2015. Chong. In The Handbook of Austroasiatic languages, Volume 2, ed. by Mathias Jenny and Paul Sidwell, 603–40. Leiden: Brill.

Pucilowski, Anna. 2013. Ho Morphology and Morphosyntax. PhD dissertation, University of Oregon.

Rabel, Lili. 1961. Khasi, a language of Assam. Baton Rouge: Louisiana State University Press.

Rehberg, Kerstin. 2003. Phonologie des Kharia–Prosodische Strukturen und segmentales Inventar. MA Thesis, Universität Osnabrück.

Ring, Hiram. 2015a. Pnar. In The Austroasiatic Languages, Volume 2, ed. by Mathias Jenny and Paul Sidwell, 1186–1226. Leiden: Brill.

Ring, Hiram. 2015b. A Grammar of Pnar. PhD dissertation, NTU, Singapore.

Ring, Hiram. 2017. Analyze Tone and Plot: A Praat script for tonal exploration and analysis. URL: http://github.com/lingdoc/praatscripts/.

Roettger, Timo and Matthew Gordon. 2017. Methodological issues in the study of word stress correlates. Linguistics Vanguard 3(1). doi:10.1515/lingvan-2017-0006.

Schiering, René, Franziska Crell and Thomas Goldammer. 2007. Phonological Word Domains in Austroasiatic Languages (ms). Workshop on Austro-Asiatic languages, Max Planck Institute for Evolutionary Anthropology, University of Leipzig, April 22, 2007.

Schiering, René and Harry van der Hulst. 2010. Word Accent Systems in the languages of Asia. In A survey of word accentual systems in the language of the world, ed. by Harry van der Hulst, Rob Goedemans and Ellen van Zanten, 509–613. Berlin: Mouton de Gruyter.

Sidwell, Paul. 2015. Austroasiatic classification. In The Austroasiatic Languages, Volume 1, ed. by Mathias Jenny and Paul Sidwell, 144–220. Leiden: Brill.

Sinha, N. K. 1975. Mundari Grammar. Mysore: Central Institute of Indian Languages.

Wang, Bei, Ling Wang, and Tursun Qadir. 2011. Prosodic realization of focus in six languages/dialects in China. In Proceedings of the International Congress of Phonetic Sciences 17, 144–147. Hong Kong: International Phonetic Association.

Xu, Yi. 2011. Speech prosody: A methodological review. Journal of Speech Sciences 1:85–115.

Xu, Yi. 2012. Function vs. form in speech prosody – Lessons from experimental research and potential implications for teaching. In Pragmatics, Prosody and English Language Teaching, ed. by Jesús Romero-Trillo, 61–76. Springer, New York.

Zide, Norman H. 1965. Gutob-Remo vocalism and glottalized vowels in Proto-Munda. In Indo Pacific Linguistic Studies, ed. by George B. Milner and Eugénie J. A. Henderson. Lingua 14:43–53.

Zide, Norman H. 2008. Korku. In The Munda Languages, ed. by Gregory D. S. Anderson, 256–98. Routledge Language Family Series. London: Routledge.

NEGATION, TAM AND PERSON-INDEXING INTERDEPENDENCIES IN THE MUNDA LANGUAGES: A PRELIMINARY REPORT[1]

Gregory D. S. Anderson and Bikram Jora
Living Tongues Institute for Endangered Languages
livingtongues@gmail.com

Abstract
This paper explores a range of interdependencies seen between negative marking and various other categories, specifically TAM-marking and person indexing across the languages of the Munda family. Data mainly comes from a large database of texts as well as lexical and grammatical elicitation collected by the authors and their colleagues over the past dozen years. Some preliminary historical reconstructions are offered for both various intermediate proto-languages as well as proto-Munda itself, where and when possible.

Keywords: Negation, typology, Munda languages, reconstruction

ISO 639-3 codes: sat, unr, hoc, biy, kfq, srb, juy, gbj, bfw, gaq, jun, khr[2]

1 Introduction and Overview

The present study represents the first step in attempting to unravel the synchronic complexities and diachronic origins of the systems of negation seen in the Munda languages based on a large data set collected by the authors and their colleagues between 2005-2017 under the auspices of their research institute's major scholarly undertaking, the Munda Languages Initiative.[3] It begins with a discussion of complex

[1] Thanks to National Endowment for the Humanities for grant PD50025-13 "Documentation of Hill Gtaʔ, an endangered Munda language of India", the National Science Foundation for award 1500092 "Documentation ofGutob, an endangered Munda language of India" and award 0853877 "Documentation of Remo (Bonda)", the Genographic Legacy Fund grant for the "Ho Talking Dictionary", and to Ironbound Films for in part making work possible on Sora, Remo, Juang, Santali, and Ho during filming of *The Linguists*. Other work on the following Munda languages was made possible under occasional funding to Living Tongues' Munda Languages Initiative: Bhumij, Birhor, Gorum, Juray, Sora, Korku, Santali, Kharia, Juang, Keraʔ Mundari and Tamaɽia Mundari. Particular thanks must be offered to Mr. Opino Gomango for assistance in the Munda Languages Initiative as field worker extraordinaire and to Dr. Anna Pucilowksi for assistance on Ho. Key consultants and language teachers for the languages include non-exhaustively Budra Raspeda, Loikong Raspeda, Lachmu Raspeda, Angra Raspeda and Parboti Raspeda (Gtaʔ), Tankhadhar Sisa, Kamla Sisa and Bondu Kirsani (Gutob), Sania Dangada-Maji and Sukari Dangada-Maji (Remo), Kameshwar Birhor and Madhuri Birhor (Birhor), Palo Purty, Rinky Purty, KC Naik Biruli, Chandra Mohan Haibru (Ho) and Kartal Sardar and Gaytri Sardar (Bhumij), just to name a few for a selection of the languages. Without their patient assistance, none of this work would have been possible.

[2] Abbreviations in examples represent the following: ABL ablative, ACC accusative, ACT active, ADJVZR adjectivalizer, ADS adessive, ALL allative, APPL applicative, ASP aspect, AUX auxiliary, BEN benefactive, CAP capabilitive, CAUS causative, CLSSFR classifier, COND conditional, COP copula, DAT dative, DECL declarative, DEF definite, DESID desiderative, DIR directional, DL dual, DS differentsubject, EMPH emphatic, EVID evidential, EXCL exclusive, OBL oblique, PFV perfective, PHB prohibitive, PL plural, PROG progressive, PRON pronoun, PRS present, PST past, PSV passive, PURP purposive, QUOT quotative, RDPL reduplication, OBJ object[ive], RECIP reciprocal, REF referential, RFLXV reflexive, RLS realis, SG singular, SUBJ subject, TAM tense-aspect-mood, TR transitive, 1 1stperson, 2 2ndperson, 3 3rdperson.

[3] Depending on the locale and language, languages used in the elicitation process by various field researchers include Hindi, Oḍia, Sadri/Sadani, Desia, and even English, as well as specific Munda languages as well, e.g., Remo, Gtaʔ, Gutob, Sora, Santali, Mundari and Ho. Some of the data presented here is in archival deposits at PARADISEC for Gutob (http://catalog.paradisec.org.au/collections/GA2) and Gtaʔ (http://catalog.paradisec.org.au/collections/GA1). We have been preparing the other materials for archival deposit, but this is very time consuming and it is also very

interdependencies of negation with TAM marking and person indexing attested in various conjugations and constructions, as seen in the languages of the Kherwarian group of North Munda, and their parallels in Korku, with an eye to determining the characteristics of the system of negation and its interaction with other verbal inflectional domains, and offers preliminary reconstructions of these systems and interdependent dynamics in the Proto-Kherwarian and Proto-North Munda languages, refining and adding to some proposals by Pinnow (1966).[4] We then turn to presenting some data on constructional vs. combinatorial semantics in negative conjugations in various Munda languages of Odisha not belonging to North Munda, all also reflecting complex interdependencies.

Based on our comparative Kherwarian data set, some intriguing features that we can likely project back into the Proto-Kherwarian or Proto-North Munda stages have come to light, in particular, complex interdependencies between negation, TAM-marking and person indexing. In a few cases, such interdependencies even appear to project back to the Proto-Munda stage. Section 2 presents an overview of the structure of positive and negative conjugations in these languages. Section 3 examines interdependencies of negation with subject and object indexing in North Munda languages and how these interactions are further impacted by tense-aspect-mood indexing, offering some thoughts on what aspects of the synchronic variation attested in these languages can be projected back to the various historical reconstructed proto-languages, viz., Kherwarian and Proto-North Munda. Section 4 discusses some interesting interdependencies of this sort in copular formations in possessive functions, specifically how possessa are variably encoded as subjects or objects in different tense formations under negation. Section 5 presents some negation/TAM interactions that do not also involve person marking.

We then turn in Section 6 to data from the other subgroups of Munda, traditionally known as South Munda, but as of yet lacking any defining innovations that justify such a classification, here simply referred to as non-North Munda. There are at least five sub-groups of such languages, three occupied by single languages Juang, Kharia and Gtaʔ, and two by sets of more closely related languages, Sora-Juray-Gorum and Gutob-Remo. Here a range of group and language-specific quirks are identified, but some features are found both across various subgroups of these languages and in some instances shared with Proto-North Munda as well. Therefore, we also cautiously make some preliminary suggestions about the possible nature of negation- and TAM-interdependencies in Proto-Munda.[5]

Of course all these languages have at least some previous documentation. While the vast majority of forms we cite below come from the field notes of the Munda Languages Initiative, we have consulted almost all published resources on these as languages (and many unpublished ones as well), but it is not our intention here to do a side-by-side comparison of our sources and published data on these same languages, a topic which merits its own full-length investigation. Because the database is collected by the same core group of researchers and using the same data collection techniques regardless of the medium of communication involved, the data represent a largely comparable and semi-controlled corpus that allows for the detection of meaningful micro-variation across, for example, closely related Kherwarian varieties, or to detail trends across the family as a whole. We feel our approach in this pilot study is therefore valid and defensible.[6]

---

costly to have materials ingested by the archive and so unfortunately we must await adequate dedicated funding before all the materials collected to date under the Munda Languages Initiative will be available.

[4] We are not going to give a point-by-point comparison with Pinnow since i) we have not attempted a systematic reconstruction yet but rather here offer only preliminary and broad stroke-type reconstructions. Indeed, all such reconstructions offered here should be approached therefore with extreme caution, as not all varieties of all languages have been surveyed yet and we have barely begun the process of systematic cross-language analysis for most categories.

[5] These should be taken for what they are, very preliminary observations suggestive of future research objectives. Pinnow (1966) is the only comprehensive attempt to reconstruct the verbal system of Proto-Munda. Pinnow's interpretation is heavily skewed towards assigning all Kherwarian structures into the proto-language, but does not include all relevant variation in the Kherwarian data.

[6] Moreover, in addition to being heavily Kherwarian in its feel, Pinnow's Proto-Munda reconstructions did not have recourse to any data from Gtaʔ–a language unknown to science in 1960 when Pinnow's (1966) manuscript was written. So this is the first attempt at a pan-Munda data synthesis that acknowledges rather considerable intra-Kherwarian variation as well as takes into consideration all known Munda languages insofar as possible. Also, since we have not fully researched the functional domains of all of the TAM markers in the various Munda languages as they appear in their naturally occurring text contexts in all the varied uses and permutations found, including

2 Kherwarian verb structure and negative formations

Like many other Munda and Austroasiatic languages (see chapters in Jenny and Sidwell (eds. 2015), Kherwarian languages have two formally distinct systems of negation, contrasting i) a general negative marked by *ba/ban/baŋ*, etc., in the Santali and Santali-esque varieties (Birhor, Mahali, etc.) and Korku (and thus probably Proto-North Munda) or by *ka* in Mundari-esque Kherwarian lects (Bhumij, Ho) with ii) a pan-Kherwarian prohibitive marker *alo*, itself possibly in part cognate with negative elements found in most non-North Munda branches of the family (discussed in section 6 below).

Proto-Kherwarian had a complex verbal system like most of its daughter languages still do. Two inflectional series can be reconstructed for Proto-Kherwarian, known since Pinnow (1966), roughly a perfective series (1) and an imperfective series (2), each with its own inflectional template. The former can be projected back to Proto-North Munda, the latter appears to have been innovated at the Proto-Kherwarian level. The two templates differ mainly in where the object marker and voice marker appear, either immediately after the verb stem and occupy the same templatic slot (imperfective) or after the TAM marker and occupying two separate templatic slots in the order voice-object (perfective), and that in the imperfective series transitive/active is unmarked and only intransitive/middle/passive overtly marked, while in the perfective series both receive overt formal indexation. This historical difference between the perfective and imperfective series is due to the fact that the imperfective series arose from an auxiliary verb construction in the development of Proto-Kherwarian from Proto-North Munda,[7] while the perfective series has cognates in Korku as well and is the diachronically older structure.

(1) a. Proto-North Munda maximal verb template [PERFECTIVE SERIES]

<[(NEG)/X=SUBJ]$_{p}$w> [Verb.Stem]$_{p}$w=[APPL-TAM-VOICE/VALENCE=/-OBJ=IND]$_{p}$w<=SUBJ>

b. Proto-Kherwarian maximal verb template [IMPERFECTIVE SERIES]

[(NEG)/X=SUBJ]$_{p}$w [Verb.Stem(=OBJ/VOICE)]$_{p}$w=[TAM=IND]$_{p}$w

The Proto-North Munda verb stem was quite simple,[8] a stem plus an optional reciprocal infix. An etymological causative prefix was preserved only in a lexically restricted manner, the functional category being renewed by various new auxiliary forms.

(2) Proto-North Munda Verb Stem

[<CAUS>-]Root$_{[</RCP/>]}$

The Proto-North Munda/Proto-Kherwarian patterns endure in many Kherwarian languages today, including Kera᳡ Mundari, Santali or Tamaɽia Mundari. Korku lost subject marking all together except in a small number of locative phrases, but otherwise shows the same templatic structure in the perfective series at least. Morphotactically, there is considerable variation with respect to the prosodic status of the different elements involved in the large Kherwarian grammatical words and whether these are treated (or are not) as one or

---

percentage counts of the use of different TAM + voice/valence markers under negation, so for this reason we are not including a table of the TAM markers referenced in this paper. For more on the TAM markers and person/number markers used in verb forms in the various Munda languages, see Anderson (2007) chapters 3and 4. The larger study that this will feed into naturally must cover all data from all resources. We feel that with native speakers of both North Munda and non-North Munda languages on our research team with advanced training and degrees in linguistics (including a co-author of the present study), that our judgements on the data is valid.

[7] More accurately there were competing AVCs in Proto-Kherwarian, one using **tan* one using **kan* to encode the imperfective/progressive. However, all the languages speak to a different AVC grammaticalized as an imperfect marker of the shape **tVhVn-ke-n* < 'remain-AOR-ITR/MDL'.

[8] There are a small number of lexicalized forms in Ho that also speak to a possible post-root slot available for an incorporated noun as is found in Sora (Anderson 2017). This, however, has no bearing on the present study.

more than one prosodic word across the Kherwarian languages (Anderson 2018, Ring and Anderson 2018).[9] Invariably however, subject markers in Kherwarian languages, if pre-posed, form a phonological word with the negative scope element opposing the word consisting of the verb stem and other inflectional markers:

(3) Kera‍ʔ Mundari

sukri	*[ka=i]* $_{p}\omega$	*[gɔj]-[ka-n-a]*$_{p}\omega$
pig	NEG=3SG.ANIM.SUBJ	kill-PFV-ITR/MDL-IND

'the pig was not killed' or 'the pig did not die'

(4) Santali

iŋ	*hola*	*ha:ʈ*	*[ba=iŋ]*$_{p}\omega$	*[ʧalá-ó]-[le-n-a]*$_{p}\omega$
I	yesterday	market	NEG=1SG.SUBJ	go-ITR/MDL/PSV.IPFV-ANT-ITR/MDL-IND

'I did not go to (the) market yesterday'

(5) Tamaɽia Mundari

kula	*sukri=ke*	*[ka=i]* $_{p}\omega$	*[goi]*^{ʔ}*-[k-i-a]* $_{p}\omega$
tiger	pig=OBJ	NEG=3SG.ANIM.SUBJ	kill-PFV-3SG.ANIM.OBJ-IND

'the tiger did not kill the pig'

Based on comparative data with Korku (6), we must reconstruct **ba(N)* ~ as the default preverbal negative particle in Proto-North Munda, with *ka* in the varieties that use this (e.g. Ho, Mundari lects) seemingly the innovator from a Munda-internal perspective.[10]

[9] The situation with regard to prosodic or phonological domains and the various processes that define such domains is both complicated and in need of resolution in the Munda languages (Ring and Anderson 2018, Hildebrandt and Anderson 2018ms). Thus there are domains that one can define by specific parameters or by the application of various phonological processes or prosodic phenomena. For example, in a language like Santali, domains might include i) the stem, ii) the stem + some inflectional suffixes but not all, iii) the stem plus all inflectional suffixes, and so on such that more than one phonological and grammatical word level can be argued for, in addition to levels clearly above and below these like iv) stems and v) phrases. Among other things, different processes of vowel harmony are varied in their domains of application (e.g., within a stem [*e/o* vs. *ɛ/ɔ*], vs. across stem+affix domains of different sorts which includes the instantiation of the harmonic contrasts of *o : u, e : i* and of *ə : a*, with still yet different domains of application, all of which interact with other word-boundary defining processes or phonotactic restrictions (no word initial *ŋ-* or word final *-s* in non-loans) as well as word vs. phrasal prominence tendencies, etc., all of which complicate the analysis. Thus there are a number of confounding factors that we have chosen to normalize in our transcriptions here for the sake of presentational/reading ease. Some grammatical elements in specific languages, such as the non-future marker in Hill Gta‍ʔ, or the subject agreement markers in Gutob can be variably extrametrical or part of the verb for the application of, for example stress assignment, which would suggest they might be transcribed as clitics in some uses and suffixes in others, while morphosyntactic/morphotactic distribution suggest the Hill Gta‍ʔ non-future is a suffix and the Gutob agreement markers rather clitics, since the host of the former is restricted to verbs and must occur in the final position of the verb template while the latter can appear multiple times in a single clause and take virtually any word as a host. Thus, morphotactic/morphosyntactic and prosodo-morphological mismatches are commonly encountered in the analysis of specific functional elements in individual Munda languages. Here we transcribe subject and case clitics with promiscuous or phrasal distribution (e.g., a case clitic occurring only once in a conjoined object NP) with [=] to encode a clitic boundary, and all other elements with a suffix boundary by [-], while remaining agnostic about the actual morphotactic status of such elements within the specific systems, and indeed acknowledging that this status may vary in a (quasi-)principled manner between two seemingly discrete categories, and moreover that syntactic or morphotactic considerations may conflict with prosodic ones in a given language in how to define the morphemes so designated.

[10] With possible cognates in other branches of Austroasiatic, it may also be old, but demonstrating this must await a separate study; see section in Anderson (2018) for a start.

(6) Korku

japai-ko	ɖusra-ku=ʈen	ban	manɖi-lakken
woman-PL	other-PL=OBLQ	NEG	speak-PROG

'the women are not speaking to each other'

While typologically plausible, there is no Kherwarian- nor Munda-internal evidence that the negative particles of Kherwarian originated as auxiliary or serial verbs in any period that is limited to the Munda languages themselves, as required by the application of the comparative method to Munda. However, there are possible Austroasiatic parallels that might reflect such an origin, as suggested by Jenny et al. (2015).[11] There, this suggestion appears to be based primarily on the placement of the subject clitics on the negative polarity elements. But the pre-verbal subject clitics in Kherwarian can attach to any word that occupies the correct structural position, that is, the immediately preverbal position, including case-marked or unmarked NPs (7), and even indeed overt subject pronouns themselves (8), so hosting subject clitics says nothing of the function of the pre-verbal elements:

(7) Ho

aiŋ	*hoˀ=ke=ŋ*	*goiˀ-k-i-a*
I	man=OBJ=1SG.SUBJ	kill-PRF-3SG.ANIM.OBJ-IND

'I killed the man'

(8) Santali

hɛ~	*iɲ=iɲ*	*ʧalak'-a*
yes	I=1SG.SUBJ	go:ITR/MDL/PSV.IPFV-IND

'yes I will go' (Bodding 1929:58)

Although this pre-verbal negator is the preferred locus, subject clitics do not obligatorily occur dislocated from the verb on the word immediately preceding it, as in (3)-(5), (7)-(8). They can occur at the end of the verbal complex as well, as in the following form from Keraʔ Mundari (9). Also, as we show below, in specific Kherwarian languages in specific constructions, subject markers can actually appear in both positions simultaneously.

(9) Keraʔ Mundari

era-ku	*inini=se*	*ɟagar-ɔ(ʔ)-r-a=ku*
woman-PL	each.other=PURP/DAT	speak-ITR/MDL/PSV.IPFV-PRG-IND=3PL

'the women are speaking to each other'

3 Subject/Negative Interdependencies in Kherwarian

3.1 Inanimate subjects

As mentioned above, subject-negation interdependencies show significant complexities in Kherwarian Munda languages. So while subject may be doubly marked in negative formations in Tamaɽia Mundari prohibitive forms (see 3.4 below), subject agreement is in fact typically suppressed and thus absent in a range of instances as well across the Kherwarian languages. For example, inanimate singular subjects are generally not marked in the positive conjugations as a rule in the Kherwarian languages, e.g., in Ho (10)-(11), in Bhumij (12) or in Santali (13).

[11] There are possible Austroasiatic analogs to both the *ba(n/ŋ)* and the *ka* negators as serial verbs, but no Munda-internal evidence supports either as originating in such *per se*. Same holds true for the likely origin of the default negator found in most Munda languages of Odisha as well (see 6 below).

(10) Ho
koto *rəpuɖ-o(ʔ)-tən-a*
branch break-ITR/MDL/PSV.IPFV-IPFV-IND
'the branch breaks'

(11) Ho
koto *rəpuɖ-jə-n-a*
branch break-PRF-ITR/MDL-IND
'the branch broke'

(12) Bhumij
koto rəpud-ʤa-n-a
branch break-PFV.ITR-ITR/MDL-IND
'the branch broke'

(13) Santali
ɖɛr rapud-e-n-a
branch BREAK-PFV-ITR/MDL-IND
'the branch broke'

However, Kherwarian languages show a division in how they treat inanimate singular subjects in negative formations. Languages like Ho (14)-(15) and Bhumij (16) show exact parallels with the positive forms, with subject marking suppressed, while inanimate subject encoding is typically not suppressed in the negative in Santali (17).

(14) Ho
koto *ka* *rəpuɖ-o(ʔ)-a*
branch NEG break-ITR/MDL/PSV.IPFV-IND
'the branch does not break'

(15) Ho
koto *ka* *rəpuɖ-jə-n-a*
branch NEG break-PRF-ITR/MDL-IND
'the branch did not break'

(16) Bhumij
koto ka rəpud-ʤa-n-a
branch NEG break-PFV.ITR-ITR/MDL-IND
'the branch did not break'

(17) Santali
ɖɛr ba=i rapud-kan-a
branch NEG=3SG.INAN.SUBJ break-IPFV-IND
'the branch isn't breaking'

This yields a typologically quirky system in Santali where inanimate subjects are overtly marked when the predicate is under negation, but otherwise are unmarked.

3.2 animate, non-human, singular

Animate non-human singular subjects can also show distinct behavior in various individual Kherwarian languages as well. Thus, and Tamaɽia Mundari (18)-(19), Ho (20)-(21) and Kera? Mundari (22)-(23) have overt subject clitics in negative formations but suppressed in positive conjugations, a situation similar to the pattern above that Santali shows with inanimate singular subjects. In other words, subject agreement is typically lacking in positive sentences with non-human animate subjects but is typically present when these same sentences are negated in these two Mundari lects and in Ho.

(18) Tamaɽia Mundari
kula sukri=ke ka=i goiˀ-k-i-a
tiger pig-OBJ NEG=3SG.ANIM.SUBJ kill=PRF.TR-3SG.ANIM.OBJ-IND
'the tiger did not kill the pig.'

(19) Tamaɽia Mundari
kula sukri=ke goiˀ-k-i-a
tiger pig=OBJ kill-PRF.TR-3SG.ANIM.OBJ-IND
'the tiger killed the pig'

(20) Ho
kula *sukri=ke* *ka=i* *goi^{ʔ}-ki-j-a*
tiger pig=OBJ NEG=3SG.ANIM.SUBJ kill-PRF.TR-3SG.ANIM.OBJ-IND
'The tiger did not kill the pig.'

(21) Ho
kula *sukri=ke* *goi^{ʔ}-ki-j-a*
tiger pig=OBJ kill-PRF.TR-3SG.ANIM.OBJ-IND
'The tiger killed the pig.'

(22) Keraʔ Mundari
sukri ka=i gɔʤ-ka-n-a
pig NEG=3SG.ANIM.SUBJ kill-PFV.NEG-ITR-IND
'the pig was not killed.'

(23) Keraʔ Mundari
sukri gɔʤ-e-n-a=e
pig kill=PFV-ITR-IND=3SG.ANIM.SUBJ
'the pig was killed'

3.3. Animate human subjects

Human animate subjects and first and second person pronominals are basically almost always encoded in Kherwarian languages in both positive sentences (24)-(25) and in negative ones alike (26)-(27) across all the languages.

(24) Ho
aiŋ *ho^{ʔ}=ke=**ŋ*** *goi^{ʔ}-k-i-a*
1SG man=OBJ=**1.SUBJ** kill-PRF-3.OBJ-IND
'I killed the man'

(25) Santali
am *iɲ=**em*** *ɖaɽ-oʈo-ki-d-iɲ-a*
2SG 1SG=**2SG.SUBJ** run-CAUS-TR.PFV-TR/ACT-1OBJ-IND
'you made me run'

(26) Santali
abo *baŋ=bo* *sen-le-n-a*
1PL NEG=1PL.SUBJ go-ANT-ITR/MDL-IND
'we have not gone'

(27) Tamaɽia Mundari
aiŋ hon=ke ka=iŋ abuŋ-k-i-a
1SG baby=OBJ NEG=1SG.SUBJ wash-PRF.TR-3SG.ANIM.OBJ-IND
'I did not wash the baby'

3.4 Imperative vs. Prohibitive

The templates in (1a-1b) primarily pertain to the declarative mood of Kherwarian. Imperative formations show different formal properties across the Kherwarian languages. With respect to imperative forms, a different set of templates was thus found in Proto-North Munda (28). In the imperative, no

declarative/indicative/finite marker is found, and the subject clitics attach directly to the object suffixes, as in Ho (29).

(28) Proto-North Munda Imperative
Verb.Stem-OBJ=SUBJ

(29) a. Ho
eto-ɲ-me
teach-1SG.OBJ-2SG.SUBJ
'teach me!' (Deeney 1979:18)

b. Ho
dʒom-e=ben
eat-INAN.OBJ-2DL.SUBJ
'eat it you-2!' (Deeney 1979:14)

The imperative template of Proto-North Munda can be projected to the Proto-Munda level, with identical structure in most languages (30). Prohibitives differ in non-North Munda languages, however (see section 6).

(30) Sora
ti'-iɲ=ba
give-1SG.OBJ=2PL.SUBJ
'you-PL give (it) to me' (Ramamurti 1931:141)

In the prohibitive (31), the finite marker is used and the subject markers are enclitic to the preverbal prohibitive marker (32). Prohibitives thus appear to be otherwise like declarative forms, differentiating from these primarily by the use of a different negator *alo* in the prohibitive.

(31) Proto-North Munda Prohibitive
PHB=SUBJ Verb.Stem-OBJ-IND

(32) Santali

dʒʰuɽi	*iŋ*	*alo=m*	*em-á-iɲ-a*
basket	1SG	PHB=2SG.SUBJ	give-APPL-1SG.OBJ-IND

'don't give me the basket!'

Kherwarian subject clitics often appear extrametrical except when prosodic minimal word constraints necessitate including them within a phonological word.[12] This may have been also true of all the elements in the clitic chain in Proto-North Munda. On the Santali end of the continuum,[13] one finds full forms of the subject clitic with monosyllabic stems obligatorily in the imperative (35), while in Birhor, which shares certain grammatical elements with Santali and certain with Mundari (Anderson and Jora forthcoming), consonant-final imperative verbs, regardless of whether they are monosyllabic or disyllabic, take the fully vocalized form =*me* of the subject clitic.

(33) Birhor
nir=me
run=2SG.SUBJ
'run'

(34) Birhor
gitiʧ=me
sleep=2SG.SUBJ
'go to sleep'

(35) Santali
ɖaɽ=me
run=2SG.SUBJ
'run!'

[12] A bimoraic prosodic minimal word constraint has been proposed for proto-Austroasiatic (Anderson and Zide 2002) which remains operative in all Munda languages today. Only bound functional elements can be CV in phonetic realization, satisfying minimal word constraints since they appear with other elements. Underlying noun stems used in isolation must be pronounced with two morae, for example /*ti*/ 'hand' is [*tii*] in Ho and Santali.

[13] While several authors have claimed a Mundari-esque orientation of Birhor in the Kherwarian language-dialect continuum, grammatically it aligns at least as much with Santali, and thus may constitute an intermediate node between the two (Anderson and Jora forthcoming).

In the prohibitive the short form of the subject clitic *=m* attaches to the vowel-final prohibitive particle but the verb is marked by final *-a* (36)-(38), unlike the corresponding imperative forms, which lack the final *-a*.

(36) Birhor
alo=m nir-a
PHB=2SG.SUBJ run-IND
'don't run!'

(37) Birhor
alo=m gitiʧ-a
PHB=2SG.SUBJ sleep-IND
'don't sleep!'

(38) Santali
alo=m ɖaɽ-a
PHB=2SG.SUBJ run-IND
'don't run!'

Thus, the fully vocalized form of the subject clitic *might* be triggered by a need to fill a prosodic minimal word constraint in the singular imperative in Birhor and Santali, but not in the singular prohibitive, where the particle itself is disyllabic.[14] Some forms in Tamaɽia Mundari show a similar pattern:

(39) Tamaɽia Mundari
nir=me
run=2SG.SUBJ
'run!'

(40) Tamaɽia Mundari
alo=pe nir=a
PHB=2PL.SUBJ run-FIN
'do not run!'

Note that in Tamaɽia Mundari there is, on the other hand, a preference towards a double marking of subjects in prohibitive forms that have overt object indexing and no finite marker (41)-(42).

(41) Tamaɽia Mundari
aiŋ=ke kanʧi alo=m om-a-iŋ=me
I=OBJ basket PHB=2SG.SUBJ give-APPL-1SG.OBJ=2SG.SUBJ
'do not give me the basket!'

(42) Tamaɽia Mundari
alo=m kaʤij-eŋ=me
PHB=2SG.SUBJ tell-1SG.OBJ=2SG.SUBJ
'do not tell me!'

Keraʔ Mundari also typically lacks the finite marker in prohibitive formations, both with overtly indexed objects (43) and without them (44). Unlike Tamaɽia Mundari, however, subject marking tends to occur on the lexical verb and *not* the prohibitive particle. Thus, while the prohibitive in most Kherwarian languages serves as the host for the subject clitics as it stands immediately before the verb in the template and this is the preferred locus of the subject clitics as a rule, in Keraʔ Mundari the subject clitic may also appear at the end of the complex in prohibitive forms.

(43) Keraʔ Mundari
aiŋ=ke alo kaʤi-ŋ=me
1SG=OBJ PHB tell-1SG.OBJ=2SG.SUBJ
'don't tell me!'

(44) Keraʔ Mundari[15]
alo nir=em
PHB RUN=IND:2SG.SUBJ
'don't run!'

[14] For Hasadaʔ Mundari, Osada (1992, 2008) suggests that final -Cs are moraic, thus a CVC stem meets the minimal bimoraic constraint. Whether this holds for Birhor, Santali and other Mundari varieties remains to be determined.

[15] What exactly triggers the use of the allomorph *=me* vs. *=em* for the second singular subject agreement marker both within and across the Kherwarian languages is not entirely clear. In Keraʔ Mundari, *=me* appears typically after vowel final stems and *=em* after consonant final verb stems, as in (43) even if an intervening consonantal suffix is present. However, other languages use other selectional criteria and other allomorphs, so in Bhumij and Ho, the variants are *=m* after vowel-final hosts and *=em* after consonant final ones (45), (47)-(49) for Bhumij while in Ho (50)-(52) one finds *=m* after disyllabic vowel-final hosts, but *=me* after consonant final disyllabic hosts and *=em* after monosyllabic ones. Only singular imperatives would in theory yield a potentially bare stem form of the verb, but since subject markers are obligatory, no Kherwarian verb form will fail to satisfy the minimal word constraint.

Bhumij shows yet another pattern. This Kherwarian language prefers just a single post verbal element in imperatives, whether it encodes subject (45) or object (46).

(45) Bhumij
nir=em
run=2SG.SUBJ
'run!'

(46) Bhumij
kadʒi-ŋe
tell-1SG.OBJ
'tell me!'

In prohibitives, these tendencies converge, and one finds formations similar to those of Birhor or Santali as in (47), but also to that of Tamaɽia Mundari with double subject marking (48). However, if there is an overt object, it appears instead of the second, pleonastic or redundant subject clitic on the verb (49).

(47) Bhumij
alo=m *sen-a*
PHB=2SG.SUBJ go-IND
'Don't go!'

(48) Bhumij
alo=m *nir=em*
PHB=2SG.SUBJ run=2SG.SUBJ
'Don't run!'

(49) Bhumij
alo=m *adʒi-ŋe*
PHB=2SG.SUBJ tell-1SG.OBJ
'Don't tell me!'

Ho shows yet a different tendency, but one that has various parallels to those patterns previously discussed. Intransitive imperatives behave in the expected fashion with an overt subject clitic in its full/vocalized form with monosyllabic stems (50) and a reduced form with disyllabic ones (51):

(50) Ho
nir=me
run=2SG.SUBJ
'run!'

(51) Ho
kadʒi=m
tell=2SG.SUBJ
'tell (me)!'

Transitive imperatives encode both object and subject, in that order, and the inflectional clitics tend to stack on the verb (52), and not appear on the word immediately preceding the verb as is typical in declarative and prohibitive formations.

(52) Ho
ʈola *ema-iŋ=me*
basket give:APPL-1.OBJ=2.SUBJ
'give me the basket!'

As would be by now expected, prohibitive formations in Ho follow the typical pattern with intransitive prohibitives, whereby subject clitics attach to the immediately preverbal prohibitive particle and are marked with the finite/declarative clitic =(*j)a* (53). However, object agreement may also be suppressed in Ho prohibitives, but subject marking left overt as in the example in (54).

(53) Ho
alo=m *nir-ja*
PHB=2SG.SUBJ run-IND
'do not run!'

(54) Ho
alo=m *kadʒi-ja*
PHB=2SG.SUBJ TELL-IND
'do not tell (me)'

Of course, object encoding may also be overt in Ho prohibitives. With transitive prohibitives, the object clitic appears together with the finite/declarative clitic, and the subject clitic attaches in the expected immediately preverbal position (55).

(55) Ho
ʈɔla *alɔ=m* *ema-iŋ-ja*
basket PHB=2SG.SUBJ give:APPL-1SG.OBJ-IND
'don't give me the basket!'

These findings on subject marking patterns in Kherwarian languages and the putative reconstructed formations in Proto-Kherwarian and Proto-North Munda are summarized in Table 1. It is straightforward to reconstruct Ø-subject marking in positive conjugations with inanimate singular subjects, and probably also in negative forms as well for Proto-Kherwarian, with its presence in Santali with negative *ba* (but not, importantly, with *baŋ*), a likely innovation based on a parallel with animate subjects. This may also have been true of animate non-human subjects as well. With animate human subjects and first and second person pronominals on the other hand, we can safely reconstruct a pattern to Proto-Kherwarian where subject marking is found in both positive and negative conjugations. Parallels in Korku data suggest we can project the Proto-Kherwarian system back to the Proto-North Munda level too.

Table 1: *Subject marking patterns in Kherwarian negative formations*

Language	INAN+	INAN	ANIM. NONHUM+	ANIM-	I/ANIM. HUM+	I/ANIM/HUM-	NP	PHB	VERB	IMP
Bhumij	Ø	Ø	Ø	√	√	√	√	√	Ø/√	√
Birhor	Ø	Ø	Ø	√	√	√	Ø	√	Ø	√
Santali	Ø	√ba, Øbaŋ	Ø	√	√	√	√	√	Ø	√
Kera? Mundari	Ø	Ø	Ø	√	√	√	√	Ø	√	√
Tamaɽia Mundari	Ø	Ø	Ø	√	√	√	√	√	√/Ø	√
Ho	Ø	Ø	Ø	√	√	√	√	√	Ø	√
PKherw	*Ø	*Ø	*Ø	*√	*√	*√	*√	*√	*Ø	*√
Korku	Ø	Ø	Ø	Ø	Ø	Ø	√	Ø	Ø	Ø
PNM	*Ø	*Ø	*Ø	*√	*√	*√	*√	*√	*Ø	*√

Key

√	subject marking present
Ø	subject marking absent
Ø/√	subject marking variable

4 NEG.COP TAM/SUBJ-OBJ interdependencies in Kherwarian isofunctional possessive forms

We now turn to some curious patterns seen between past and present in negative copula formations in possessive functions. First let's examine for comparison how positive and negative copula formations with animate possessa operate in the present. In Bhumij (56)-(57) or Ho (58)-(59), both positive and negative formations encode such referents as morphological objects in the present.

(56) Bhumij

iɲa(ʔ)	*bəria*	*kuɽihon-kin*	*mena(ʔ)-**kin**-a*
1SG:GEN	two.ANIM	daughter-DL	COP-**3DL.OBJ**-IND

'I have two daughters'

(57) Bhumij

iɲa(ʔ)	*bəria*	*kuɽihon-kin*	*baŋ-**kin**-a*
1SG:GEN	two.ANIM	daughter-DL	NEG.COP-**3DL.OBJ**-IND

'I don't have two daughters'

(58) Ho
*aiɲa(ʔ) bəria ku:ihon-kin mena(ʔ)-**kin**-a*
1SG:GEN two.ANIM girl.child-DL COP-**3DL.OBJ**-IND
'I have two daughters'

(59) Ho
*aiɲa(ʔ) bəria ku:ihon-kin baŋ-**kin**-a*
I:GEN two girl.child-DL NEG.COP-**3DL.OBJ**-IND
'I don't have two daughters'

To be sure, similar formations can be found in positive (60) and negative copula forms (61)-(63) in the present tense across the Kherwarian Munda languages, regardless of what the formal shape of the negative particle/copula is, e.g., *banu(ʔ), baŋ, ka…li-*, etc.

(60) Keraʔ Mundari
*aiɲa(ʔ) du ʈʰɔ kuɽihɔn hen-**kin**-a*
1SG:GEN two CLSSFR daughter COP-**3DL.OBJ**-IND
'I have two daughters'

(61) Keraʔ Mundari
*aiɲa(ʔ) du ʈʰɔ kuɽihɔn ka li-**kin**-a*
1SG:GEN two CLSSFR daughter NEG NEG.COP-**3DL.OBJ**-IND
'I don't have two daughters'

(62) Tamaɽia Mundari
*aĩja(ʔ) barija honkuɽi-kin baŋ-**kin**-a*
1SG:GEN two.ANIM daughter-DL NEG-**3DL.OBJ**-IND
'I do not have two daughters'

(63) Santali
*iŋ-rin barija kuɽigidra banu(ʔ)-**kin**-a*
1SG-GEN.ANIM.PSM two.ANIM daughter NEG.COP-**3DL.OBJ**-IND
'I don't have two daughters'

In *past* negative copular formations, animate possessa are rather encoded as *subjects*, in an *anti-ergative* type of patterning. Such a pattern is attested across the Kherwarian languages, see (64)-(69), and thus can be safely projected back to Proto-Kherwarian. The agreement clitics attach as expected to the preverbal negative particle, whether this is the *ka*/*kə* series or the *ba*(*ŋ*) series.

(64) Ho
*aiɲa(ʔ) bəria ku:ihon-kin ka=**kin** taiken-a*
1SG:GEN two.ANIM daughter-DL NEG=**3.DL.SUBJ** PST.COP-IND
'I did not have two daughters'

(65) Santali
*iŋ-rin barija kuɽigidra ba=**kin** taheken-a*
1SG-GEN.ANIM.PSM two.ANIM daughter NEG=**3DL.SUBJ** PST.COP-IND
'I did not have two daughters'

(66) Bhumij

iɲa(ʔ)	*bəria*	*kuɽihon*	*ka=**kin***	*taiken-a*
1SG:GEN	two.ANIM	daughter	NEG=**3DL.SUBJ**	PST.COP-IND

'I did not have two daughters'

(67) Tamaɽia Mundari

aĩja(ʔ)	*barija*	*honkuɽi-kin*	*ka=**kin***	*taiken-a*
1SG:GEN	two.ANIM	daughter-DL	NEG=**3DL.SUBJ**	PST.COP-IND

'I did not have two daughters'

(68) Birhor

iŋ-ren	*bəria*	*majõ*	*kə=**kin***	*təhiken-a*
1SG-GEN.ANIM.PSM	two.ANIM	daughter	NEG=**3DL.SUBJ**	PST.COP-IND

'I didn't have two daughters'

(69) Kera**ʔ** Mundari

aiɲa(ʔ)	*du*	*ʧʰɔ*	*kuɽihɔn*	*ka=**kin***	*dɔhɔnken-a*
1SG:GEN	two	CLSSFR	daughter	NEG=**3PL.SUBJ**	PST.COP-IND

'I did not have two daughters'

The particular interdependencies between subject vs. object agreement and negation in copular forms in the Kherwarian languages, and the putative reconstructed Proto-Kherwarian system are presented in Table 2.

Table 2: *OBJ vs. SUBJ encoding in Kherwarian PRS vs. PST negative copula forms*

Language	PRS.COP	OBJ	SUBJ	PRS.COP. NEG	OBJ	SUBJ	PST.COP.	OBJ	SUBJ	PST.COP.NEG	OBJ	SUBJ
Bhumij	mena(ʔ)	√	Ø	bano(ʔ)/bəno(ʔ)	√	Ø	taiken	Ø	√	ka..taiken	Ø	√
Birhor	mena(ʔ)	√	Ø	bənu/o(ʔ)	√	Ø	təhiken	Ø	√		Ø	√
Santali	mena(ʔ), Ø<anim>	-	-	banu(ʔ)/bano(ʔ); baŋ; ba	√	Ø	taheken	-	-	ba taheken	Ø	√
Kera**ʔ** Mundari	hen	√	Ø	ka likna	√	Ø	dɔhɔnken	Ø	√	ka le	Ø	√
Tamaɽia Mundari	mena(ʔ)	√	Ø	baŋ[16]	√	Ø	taiken	Ø	√	ka taiken	Ø	√
Ho	mena(ʔ)	√	Ø	baŋ	√	Ø	taiken	Ø	√	ka taiken	Ø	√
Pkherw	COP	√	Ø	*ba(N) ~ *ka	√	Ø	PST.COP	Ø	√	NEG+PST.COP	Ø	√

Key:

√ referent encoded as OBJ/SUBJ

Ø referent not encoded as OBJ/SUBJ

This putative Proto-Kherwarian system of copular formations has reflexes in Korku as well. However in Korku, unlike Kherwarian, the present forms show the same split as the past ones do, and thus all positive copular forms (70) treat the animate possessa as *objects* (and thus can be encoded in the morphological verb-

[16] Hasada? Mundari has *banoʔ* (Osada 2008:132). Thanks to an anonymous reviewer for reminding of us of this important fact. It is likely that such a form was also present in Proto-Kherwarian.

word), but in negative copular forms (71), they are encoded rather as *subjects*, and thus remain unmarked in the verbal complex, as Korku lacks subject marking (Zide 2008). However the nouns referring to the possessa themselves may take indexes of nominal number of course, as animate non-singular nouns typically do in Korku.

(70) Korku

iɲ-en	*bari*	*koŋje-kin*	*ʈa-kin*
1SG-GEN/DAT	two	daughter-DL	COP-DL

'I have two daughters'

(71) Korku

iɲ-en	*bari*	*koŋje-kin*	*baɲ*
1SG-GEN/DAT	two	daughter-DL	NEG.COP

'I don't have two daughters'

It seems likely therefore that Proto-Kherwarian reflects the original Proto-North Munda system, and that this was analogically extended to include present copular forms as well in Korku. A detailed picture of how and why this system arose must await further research.

5 TAM/NEG Interdependencies in Kherwarian: future forms

In addition to interdependencies between argument encoding and negation ±TAM there are also interdependencies seen in Kherwarian languages between negation and the formal markers of TAM themselves. Many Kherwarian languages prefer different TAM markers in perfective series negatives than they use in the corresponding positive conjugations. This variation is complex and extensive and remains a subject of ongoing research as to how to tease apart the various historical layers and the particular semantic and discourse/pragmatic factors that are interacting in the determination of this. We offer simply some brief comments here. Thus, in Santali, perfective transitive/active forms prefer the perfect TAM marker *ke-/ki-* (72), but the corresponding negative forms prefer the anterior *le-/li-* (73) However, as (74) shows, the same TAM marker as the positive conjugation is permitted in negative forms, and thus the opposition is simply a tendency or statistical preference. More research is required to fully determine what factors contribute to this.

(72) Santali

am	*iɲ=em*	*ɖaɽ-oʈo-ki-d-iɲ-a*
2SG	1SG=2SG.SUBJ	run-CAUS-TR.PFV-TR/ACT-1SG.OBJ-IND

'you made me run'

(73) Santali

am	*iŋ*	*ba=m*	*ɖaɽ-oʈo-li-d-iɲ-a*
2SG	1SG	NEG=2SG.SUBJ	run-CAUS-TR.ANT.NEG-TR/ACT-1SG.OBJ-IND

'you didn't make me run'

(74) Santali

am	*iŋ*	*ba=m*	*ɖaɽ-oʈo-ki-d-iɲ-a*
2SG	1SG	NEG=2SUBJ	run-CAUS-TR.PFV-TR/ACT-1SG.OBJ-IND

'you didn't make me run'

With intransitives, the preference is even stronger, but it is still not an absolute requirement for the use of the anterior marker *–l(e)* in Santali negative past formations.

(75) Santali

iŋ	*hola*	*haːʈ*	*ba=iŋ*	*ʧala-o(ʔ)=le-n=a*
1SG	yesterday	market	NEG=1SG.SUBJ	go-ITR/MDL/PSV.IPFV-TR.ANT.NEG-ITR/MDL-IND

'I did not go to market yesterday'

6 Negation and negative- TAM interdependencies in non-North Munda languages

Turning now to the other languages of the family, the non-North Munda languages range from relatively simple to quite complex in the systems of negation and how these interact with person encoding and TAM marking.[17] Most non-North Munda languages of southern Odisha show one or two cognate negative scope elements, and other often non-cognate negators as well. Across most of the languages is a negative element that is morphotactically a prefix and variously realized as *a-*, *ar-*, *ad-*, *aC-*, etc., depending on the form of the stem it attaches to and the language involved, in the case of the default negator, all of which derive from Proto-Austroasiatic **ʔəʔt*: Sidwell and Rau (2015), Rau (2017 ms) and Anderson (2017 ms) have independently suggested this may derive from a proto-Austroasiatic serial verb etymologically meaning 'lack' **ʔəʔt* in Proto-Austroasiatic (but lacking the glottal initial in Proto-Munda most likely and with a preglottalized final as *aˀ*d-* in Proto-Munda). It is at least possible that this same element is reflected in the (first half of the) Kherwarian prohibitive particle *alo*, but they could also be independent. The other common negator takes the shape of *ama-*, *am-*, *ma-* in Juang and Gtaʔ, which may or may not be related to the default preverbal negator in *um* in Kharia. Languages discussed here are Sora, Juray, Remo, Gutob, Juang and Gtaʔ, after first briefly mentioning some data from Kharia.

6.1 On subject inflection in Kharia negative forms

Kharia has a relatively simple system of negative formation using the particle *um* as the default negator. What is noteworthy about Kharia is that this negative element stands in immediately preverbal position (76) and may serve as the host for subject clitics–a system quite reminiscent of Kherwarian languages as discussed above, which contrasts with subject clitic placement in positive inflections (77) in Kharia, which rather typically follows the TAM marker. Whether this negative+subject pattern can be attributed to Kherwarian, specifically Mundari (78), influence in Kharia–which is plausible given present-day and likely past contact scenarios–remains an open question, as does a possible alternative explanation, that this formation is an inherited structure shared between Proto-Kherwarian and Kharia but lost in other branches. Note that Kharia is the northernmost of the non-North Munda languages, the only one spoken in Jharkhand and the only one in direct contact with Kherwarian. It is thus perhaps not surprising that it largely stands apart from the other languages in this regard, but this fact does not *a priori* favour inheritance nor metatypic shift/convergence as an explanation for the presence of this construction in Kharia, but, on the other hand, the data clearly lend themselves to this interpretation.[18]

[17] We use the somewhat infelicitous term non-North Munda here to underscore the fact that it has yet to be demonstrated that these non-Kherwarian and non-Korku languages form a coherent taxon, avoiding the term South Munda, which, while convenient or more euphonic, is not particularly useful. Even southern Munda is non-ideal as Kharia is spoken in Jharkhand and northern Odisha in the same areas as Kherwarian languages, not to mention that Sora is also spoken on various tea gardens in Assam (Horo 2017ms, 2017; Horo and Sarmah 2015).

[18] The Gutob pronominal clitics as discussed in Zide (1997) are very promiscuous in distribution when appearing outside of their normal distributional position, enclitic to the verb+tam forms. Subject clitics in Gutob may appear multiple times in a sentence or in the case of Kherwarian, any word that appears in immediately preverbal position *or* (but importantly not *and*), in a handful of examples, on the first word of the clause, but never three or more times in a single clause, as is attested in the Gutob text corpus and in Zide's (1997) publication. Thus, the Kharia forms are very likely due to Kherwarian influence, and neither have anything to do with the Gutob subject markers, whose behavior shows no analogs even in the closely related Remo language, and indeed some, like the 3rd plural marker *=nen*, appear to be of very recent origin, and are not even cognate with isofunctional markers in Remo.

(76) Kharia (Peterson 2011:337)

um=iɲ	*lam=te*
NEG=1SG	want=ACT.PRS

'I don't want'

(77) Kharia (Peterson 2008:463)

ho=ki	*tenton=ga*	*maj=te=ki*
THAT=PL	TAMARIND=FOC	MIX=ACT.PRS=3PL

'they mix in the tamarind'

(78) Tamaɽia Mundari (repeating 79 above)

aĩja(ʔ)	*ti*	*ka=iŋ*	*abuŋ-a*
I:GEN	hand	NEG=1SG.SUBJ	wash-IND

'I will not wash my hand'

6.2 Negation-TAM interdependencies in non-North Munda languages of Odisha

We begin our brief survey of negative-TAM interdependencies in the non-North Munda languages of Odisha with the largest of them, Sora. Sora appears to permit a single pre-stem inflectional slot that can be filled by either a plural subject prefix or a negative scope element, but not both.[19] In the past the negative scope operator attaches to the tense-marked verb in a combinatorial manner. In both instances the past marker *–l(i)-* is used whether under the scope of negation or not (79)-(81).

(79) Sora

a-ɲam-dʒaʔt=lɪ-n-aj

1PL.SUBJ-catch-snake-PST-ITR/MDL-1.ACT

'we caught (a/the) snake(s)'

(80) Sora

amən	*doʔŋ-ɲen*	*a-giɟ-l-iɲ*
2SG	OBJ-1SG	NEG-see-PST-1SG.UND

'you have not seen me'

(81) Sora

anindʒi	*rban*	*daʔa-n*	*a-tij=l-əm-dʒi*
3PRON:PL	yesterday	water-N.SFX	NEG-give-PST-2SG.UND-3PL.ACT

'yesterday they didn't give you water'

Sora shows formal differences in the positive and negative variants of sentences in the non-past that are slightly more complex than the addition of a negative polarity item to the positive form, as in the following examples where the non-past marker *–t(i/e)-* in (82) is suppressed when the negative prefix *ə-* is added (83).

(82) a. Sora

ɲɛm-dʒaʔt-tɪ-n-dʒi

catch-snake-NPST-ITR/MDL-3PL.ACT

'they (will) catch (a/the) snake(s)'

b. Sora

ɲen	*giʔj-t-aj*
1SG	see-NPST-1.ACT

'I (will) see'

c. Sora

ɲen	*kəmbun-an=adoʔŋ*	*tij-ɟum-t-ai*
1SG	pig-N.SFX=OBJ	give-food-NPST-1.ACT

'I will feed the pig' (field notes)

[19] This appears to be in flux and subject to individual speaker variation. Some speakers distinguish 1PL positive and negative by using a lengthened vowel in the negative, suggesting a phonetic coalescence of what remain two distinct templatic prefix slots for the first plural marker and the negative marker.

(83) Sora (Anderson & Harrison 2008b:346, 331)

ɲen	*bazar-ɪn*	*ə-jeːr-ej*
1SG	market-N.SFX	NEG-go-1.ACT

'I don't, won't go to the market'

The closely related Juray attests a somewhat similar pattern: the non-past marker (encoding future and present) is suppressed in the negative but obligatory in the positive. Note also that the non-finite converb form of the lexical verb is identical with the past marker even in present forms (84) and the auxiliary takes the tense and person encoding in Juray in the positive in the syntactic order V AUX, while in the negative the order is reversed, and we find AUX V and polarity and person marking rather on the lexical verb (85).

(84) Juray

ɲen	*əman=adoʔŋ*	*giɟ-le*	*rabti-t-am*
1SG	2SG=OBJ	see-CV	CAP-NPST-2SG.UND

'I can see you'

(85) Juray

ɲen	*əman=adoʔŋ*	*rabti*	*a-giɟ-am*
1SG	2SG=OBJ	CAP	NEG-see-2SG.UND

'I am not able to see you'

Gutob has a default negative prefix and a negative copula.[20] One majorly complex feature of Gutob conjugation however is that there are TAM elements in the positive conjugations that have different functions in the negative conjugations despite being formally identical. For example, the tense marker *-gu* marks past with class-I verbs (mainly intransitive and middle verbs) but when combined with the negative prefix *ar-*, it encodes prohibitive (Anderson 2007, Voß 2017).

(86) Gutob

ser-gu
sing-PST.ITR/MDL
'sang'

(87) Gutob

ar-ser-gu
NEG-sing-PHB
'don't sing'

Similarly the TAM suffix *-to* encodes a habitual present in the positive but when combined with the negative prefix *ar-*, a negative past tense is rather the result.[21]

(88) Gutob

ser-to
sing-HAB
'sings'

(89) Gutob

ar-ser-to
NEG-sing-NEG.PST
'didn't sing'

As alluded to above, not all negative constructions in Gutob use the prefix *ar-*. The negative copula functions as the negative polarity marker in a range of conjugations. That the element is a negative copula is clear from examples like (90), where it functions in opposition to structures like *ɖu-* in the positive (91).

[20] As well as a vanishingly rare compound anticipatory negative *mor-* that etymologically includes the default negator, and also likely included the non-finite negator mentioned in Gtaʔ, see below.

[21] Originally from Zide's field notes, and published in Anderson (2007), Griffiths (2008), confirmed in field by authors in 2013.

(90) Gutob
niŋ-nu *dʒoɽek* *oʔon* *uraʔ*
1SG-GEN two child NEG.COP
'I don't have two daughters'

(91) Gutob
niŋ-nu *dʒoɽek* *oʔon* *ɖu-tu=nen*
1SG-GEN TWO CHILD COP-NPST=3PL
'I have two daughters'

Note there is no agreement if the possessum is inanimate; note also that this agreement system is *not* the same as the one attested in Kherwarian. With past formations, one finds [*ar-*]*ɖu-gu* in Gutob (92)-(93). Thus in copula forms *ar-X-gu* is concatenative NEG + PST, but with verbs it forms a prohibitive circumfix, i.e., it is constructional semantically.

(92) Gutob
niŋ-nu *dʒoɽek* *ɖieŋ* *ɖu-gu*
1SG-GEN two house COP-PST
'I had two houses'

(93) Gutob
niŋ-nu *dʒoɽek* *ɖieŋ* *aɖ-ɖu-gu*
1SG-GEN TWO HOUSE NEG-COP-PST
'I did not have two houses'

Perhaps due to the 'unnaturalness' of this system, and perhaps due to the obsolescence of the Gutob language as a whole and/or the long-term Dravidian influence from Dravidian-speaking Gadaba, as well as the increasing dominance of Indo-Aryan Desia–both of which use negative copula forms in finite formations–there appears to be an ongoing generalization of the negative copula form *uraʔ* into finite structures in our Gutob corpus (94).

(94) Gutob
niŋ *minɖig* *(h)aʈ-boʔ* *ui=niŋ* *uraʔ*
1SG yesterday market-DIR/LOC go=1SG NEG.COP
'I did not go to market yesterday'

Thus the system of negation in Gutob appears to be breaking down somewhat in this seriously endangered language, at least for the speakers we have recorded. It is likely of course that this type of obsolescence effect is subject to considerable local and even individual variation. In our data set on Gutob (dozens of texts, thousands of sentences), we find examples of the use of the old system as predicted and described above, but one now also hears a more typologically 'normal' and altered structure for the prohibitive, with just the verb stem in a bare form (95) and the default negator, i.e., a formation paralleling the positive imperative structure (96):

(95) Gutob
o-niŋ *ar-su:n*
OBJ-1SG NEG-tell
'don't tell me!'

(96) Gutob
o-niŋ *su:n*
OBJ-1SG tell
'tell me!'

Turning now to Gtaʔ, there is a general default prefix *a(r)-* in Plains Gtaʔ and Hill Gtaʔ used in both declarative and prohibitive forms. The prohibitive in Gtaʔ, as in pre-decline Gutob, is constructional and thus non-combinatorial semantically, using an evidential/perfect marker *=gɛ/=gə* together with the negator to yield the prohibitive meaning (97).

(97) Hill Gtaʔ
a-næjŋ *na-á-basoŋ-gɛ*
OBJ-1SG 2SG-NEG-tell-PHB
'don't tell me!'

Also similar to pre-obsolescent Gutob is the constructional use of the *non-past* marker with the negator to create *past* negative formations. In Hill Gtaʔ this element is *=tɛ/=tə* (98) and in Plains Gtaʔ *=ke* (99).

(98) Hill Gtaʔ

gubug	*a-goiʔ-tə*
pig	NEG-die-NEG.PST

'the pig didn't die'

(99) Plains Gtaʔ

n-ár-aʔtʃoŋ-ke
1SG-NEG-FEED-NEG.PST

'I didn't feed (s.o.)'

In simplex predicates in Hill Gtaʔ, the negative plus the non-past TAM marker, i.e. *a-...-tə* constructionally encodes negative past tense (98), but conversely encodes negative present tense in complex predicates (100), that is concatenatively or combinatorially, not constructionally.

(100) Hill Gtaʔ

ɖiaŋkoj	*ɖiaŋkoj*	*ho(ʔ)-barsoŋ*	*a-riŋ-tə*
woman	woman	RCP-speak	NEG-IPFV-PRS

'the women are not speaking to each other'

As in Gutob, the negative plus the default past tense marker encodes a prohibitive construction. Unlike Gutob, subject marking is usually overt in Hill Gtaʔ and not suppressed in the prohibitive as in (95) above, as it also typically is in all first and second person subject forms in Gtaʔ in all the TAM forms (third person subject is unmarked), as in the perfect (101):

(101) Hill Gtaʔ

a-me	*kej*	*n-læʔ-tɛ*
OBJ-3SG.PRON	see	1SG-PRF-PRS

'I have seen her'

While future and present are not conflated in Gtaʔ, as they are for example in Sora, nevertheless the TAM marker is suppressed in negative future forms in Gtaʔ as well. It is important to mention here that the system in Gutob, Gtaʔ and Sora is similar to what can be reconstructed to Proto-Kherwarian and Proto-North Munda systems,22 which suggests negator alone with no tam marking may have marked negative future in Proto-Munda. Note that this yields a typologically quirky situation where while both forms consist of three morphemes, the negative first singular future has only two syllables (102) and is thus shorter than the positive first singular future form (103), which is rather trisyllabic.

(102) Hill Gtaʔ

kine	*hãwe*	*a-na*	*n-a-biʔ*
this	bow	OBJ–2SG	1SG-NEG-give

'I will not give you this bow'

(103) Hill Gtaʔ

gubug=kə	*m-biʔ-wə*
pig=OBJ	1SG-give-FUT

'I will give (it) to the pig'

Remo, although closely related to Gutob, has innovated away from the Gutob system. Thus, the negative present is simplex and combinatorial in Remo with the structure NEG-Verb-PRS-SUBJ (104). This is true of both class I or intransitive/middle verbs and class II or transitive/active verbs in Remo (105)-(106).

(104) a. Remo

niŋ	*a-lop-t-iŋ*
1SG	NEG-fall-NPST-1SG

'I do not fall, am not falling'

b. Remo

pe	*gulajro*	*a-goiʔ-te-pe*
2PL	all.ANIM	NEG-die-NPST-2PL

'you all do not die'

(105) Remo

niŋ	*a-no*	*dʒu-t-iŋ*
1SG	OBJ-2SG	see-NPST-1

'I see you'

(106) Remo

niŋ	*a-no*	*a-dʒu-t-iŋ*
1SG	OBJ-2SG	NEG-see-NPST-1SG

'I don't see you'

This is also true in complex predicates using the progressive or imperfective auxiliary, similar to Gtaʔ.

(107) Remo

pe	*gulajro*	*goiʔ*	*a-rĩː-te-pe*
2PL	all.ANIM	die	NEG-AUX-NPST-2PL

'you all are not dying'

However, in the negative past in Remo, although this system has been restructured with new auxiliary elements, it has preserved some older quirky features as well such as the structure of NEG-AUX-NPST-SUBJ in negative past forms with the positive present tense marker. Thus it maintains the constructional nature of the negative past. With Class I or intransitive/middle verbs (108), the auxiliary used in the negative past is *ar-ej ~ ar-oj,* presumably historically < 'go', with the *-r-* historically part of the negative prefix. For a subset of monosyllabic verbs of this type, this construction requires a reduplicated non-finite form of the lexical stem (109).

(108) Remo

niŋ	*log*	*ar-ej-t-iŋ*
1SG	fall	NEG-AUX.NEG.ITR/MDL-NEG.PST-1SG

'I did not fall'

(109) Remo

pe	*gulajro*	*gugoiʔ*	*ar-oj-te-pe*
2PL	all.ANIM	RDPL~die	NEG-AUX.NEG.ITR/MDL-NEG.PST-2PL

'you all did not die'

With Class II or transitive/active verbs, the auxiliary used is rather *boŋ* (or rather *a-boŋ*; < 'put') and the lexical verb is required to be in the converb form that is identical to the positive past tense marker, but the auxiliary nevertheless takes the negative past marker that is identical with the positive non-past. Compare (110) and (111).

(110) Remo

niŋ	*a-no*	*dʒul-oʔ-niŋ*
1SG	OBJ-2SG	see-PST.TR/ACT-1SG

'I saw you'

(111) Remo

niŋ	*a-no*	*dʒu(l)-oʔ*	*a-boŋ-t-iŋ*
1SG	OBJ-2SG	see-PST.TR/ACT	NEG-AUX.NEG.TR/ACT-NEG.PST-1SG

'I didn't see you'

In future negative forms, a negative future element is found, probably a semi-univerbated auxiliary in origin.[22] This can appear either in a structure following a negative marked lexical verb (112) reflecting a split negative inflectional paradigm, or it can appear in a formation with negative marked on both the lexical verb and the auxiliary verb (113) in a split/doubled configuration (Anderson 2006).

(112) a. Remo

niŋ	*a-log-a-niŋ*
I	NEG-fall-NEG.FUT-1

'I will not fall'

b. Remo

pe	*gulajro*	*a-goiʔ-a-pe*
you.PL	all.ANIM	NEG-die-NEG.FUT-2PL

'you all will not die'

[22] That forms like (113) exist suggest that the negative future element in Remo *-/=a-* is synchronically still an auxiliary or at least optionally remains so for some speakers. For others it appears to be moving into the function of a (restricted negative future) TAM suffix. Thanks to an anonymous reviewer for pointing this out.

(113) Remo
niŋ *a-log=ar-a-niŋ*
I NEG-fall=NEG-NEG.FUT-1
'I will not fall'

In Juang, forms typically are combinatorial semantically with the TAM forms also maintaining their meanings in positive conjugations under negation such as present (114)-(115) or past active/transitive (116).

(114) a. Juang
ne *ʤandare* *a-lara-ke*
DEM woman.TOP NEG-laugh-PRS
'the woman is not laughing'

b. Juang
arokia *baronoŋ* *a-goiʔ-ki-kia*
3PRON.DL TWO.HUM NEG-DIE-PRS-3DL
'those two do not die, are not dying'

(115) Juang
ne *ʤandare* *lara-ke*
dem woman.TOP laugh-PRS
'the woman is laughing'

(116) Juang
arokia *baronoŋ* *a-goiʔ-jo(ʔ)-kia*
3PRON.DL two.HUM NEG-die-PST.TR/ACT-3DL
'two of them did not die'

Like Gutob, the negative copula *ʤena* has also been drawn into the finite verb system in Juang, with reduplicated forms of monosyllabic verb stems. That *ʤena* is etymologically a negative copula can be seen comparing (117) and (118).

(117) Juang
ini *butare* *guba=ro*
DEM pig.TOP big=COP
'the pig is big'

(118) Juang
ne *butare* *guba=ʤena*
DEM PIG.TOP BIG=NEG.COP
'the pig is not big'

When used as a negator in finite forms in Juang, monosyllabic verb stems are obligatorily reduplicated (119) with the negative copula *ʤena*. Indo-Aryan influence is likely here as the use of the negative copula in finite verb forms is relatively common in the local variety of Oḍia.

(119) Juang
arokia *baronoŋ* *gogoiʔ=ʤena*
3PRON.DL two.HUM RDPL~die=NEG.COP
'those two have not died'

6.3 Non-finite negator in ma-

Another negative scope operator found in Munda languages with analogs in other Austroasiatic languages is the non-finite negator in *ma-* (Anderson 2018). It can be used in subordinate clauses and in attributive formations as in the following Juang (120) and Gtaʔ (121) forms.

(120) Juang
apa *a=ma=ʤim-ke* *ete* *aiŋ* *kikib=ʤena*
2DL 2DL=NEG=eat-PRS because 1SG RDPL~do=NEG.COP
'because you don't eat (it), I didn't do it' (Patnaik 2008:546)

(121) Hill Gtaʔ

ma=bihæ=nə	*ngire*
NEG.ATTR=marry=ATTR	young.man

'unmarried young man, bachelor'

6.4 On negation-TAM interdependencies in Munda languages of Odisha

Munda languages of Odisha show a range of complex TAM-negation interdependencies. Several languages with respect to TAM+negation categories express this functional nexus constructionally and not concatenatively or combinatorially. That is, the functional value of certain TAM markers in positive inflections changes when the same TAM markers appear under negation, seen in Gutob and Gtaʔ. This typologically odd system is in flux in the languages like Gutob today, which shows various degrees of obsolescence and contact-driven shift or reorganization. Some languages like Remo have innovated new structures but partly kept old patterns intact. In others, like Juang and Gutob, negative copula forms have been shifted into finite negative TAM functions. However, one feature shared between Proto-North Munda and at least two different subgroups of Munda languages of southern Odisha (Sora-Juray-Gorum and Gtaʔ) is the suppression of TAM marking to encode negative future. As such, it seems that one might posit the structure of negator plus Ø-TAM marking to encode negative future back to the Proto-Munda stage.

7 Summary

The present study represents the first attempt to study systems of negative marking across the full spectrum of Munda languages, updating the work of Pinnow (1966), who had no access to quality data on several languages including Remo and several minor Kherwarian varieties, and no data whatsoever on Gtaʔ. As a result much of Pinnow's specific reconstructions are heavily skewed to major Kherwarian languages like Santali or Hasadaʔ Mundari, and do not give sufficient weight to other data. Moreover, Pinnow worked under the premises of a now rejected branching of Munda that had Kherwarian, Korku, Kharia-Juang and 'Koraput Munda' on equal footing, with the consequence that many features that should be considered only as old as the Proto-North Munda stage have been projected back to Proto-Munda. It is now recognized (independently by Anderson 2015, 2016 and Sidwell and Rau 2015) that Korku and Kherwarian form a higher node, North Munda, while the other groups so-joined in fact do not form valid sub-taxa, but rather Kharia and Juang each form isolate branches (as does Gtaʔ, unknown to Pinnow), while Sora-Gorum and Gutob-Remo are each small defensible sub-taxa.

Many Munda languages make at least a formal distinction between two types of formations with regards to their system of negative marking, often contrasting prohibitive with other negative conjugations. Similar phenomena have been attested in a wide range of Austroasiatic languages. Such a formal opposition in negative markers can be reconstructed for Proto-North Munda at least, and probably Proto-Munda as well. In Kherwarian languages, there is a contrast between a default negative marker and a prohibitive one, both appear before the verbal complex and typically serve as the host for subject clitics. In Santali, both frequently serve in this function. In Keraʔ Mundari, the default negative particle *ka* is more likely to serve as host for the subject clitics than the prohibitive one, *alo*.

In all the Kherwarian languages but Santali, inanimate subjects are unmarked in both positive and negative structures. Animate non-human singular subjects are typically unmarked in Bhumij, Birhor, and Tamaɽia Mundari in positive forms but marked in negative ones, while in Santali this is variable and in Keraʔ Mundari subject marking is found in both positive and negative formations of this type. Speech-act participants, human singular, dual and plural subjects are typically encoded in both positive and negative formations in all languages, and these patterns can be safely reconstructed to Proto-North Munda. Subsequently subject clitics were lost in Korku, and subject marking appears preserved as third person number enclitics only in some locative expressions (Zide 2008).

In imperative forms across Kherwarian, subject marking is obligatory, while in prohibitive forms, subject can be optionally doubly marked on both the prohibitive marker and the verb in Bhumij and Tamaɽia Mundari. Unlike the other Kherwarian languages, prohibitive forms prefer subject marking on the verb and not the prohibitive element itself in Keraʔ Mundari.

Negative future formations lack any tense marker in many North Munda languages, as do positive future in most but not all conjugations across the various Kherwarian languages, with formations in Birhor,

Ho and Korku variably interpreted as having an overt index of future; this pattern with a lack of formal marking in negative futures can be safely reconstructed for Proto-North Munda. In addition, an identical Ø-marking of future or non-past in negative formations is also attested in a range of diverse languages of southern Odisha, such as Sora and Gtaʔ. As such, this pattern is likely to be reconstructable to Proto-Munda.

There is a typologically odd use of constructional not combinatorial semantics of TAM marking under negation found in both Gutob and Gtaʔ, a system restructured in Remo but with traces of the older forms remaining. This system appears to be old. Both Juang and Gutob have independently drawn the negative copula forms into new finite structures of the shape V NEG, possibly under external/contact influence.

It is likely that negators in the Munda languages may have originated in serial verb constructions in a pre-Proto-Munda Austroasiatic dialect. At least one Proto-Munda negator may have its origin in the negative scope operator found in, and putatively derived from, an original verb meaning 'lack', and this is a typologically normal grammaticalization path to assume for how verbal negators arise (Anderson 2011), but there is as yet no Munda-specific or Munda-internal evidence that suggests this *per se*, and all must be simply considered syntactically as negative particles at all historical stages within Munda proper, attested or reconstructed. Such elements appear to have been drawn into the dependent operator functional clitic chain usually in one of the two leftmost prefix slots. Originally, the negators probably hosted the subject *pro*clitics, an order reflected in Juang, Sora and Gtaʔ, both the *m*-negator series and the one putatively derived from 'lack' (cf. Khmer, Nicobarese [*ʔət*]) that has become the default or general negator in several southern Munda branches. In Proto-Kherwarian and in Kharia, the negator serves as host for subject *en*clitics, so both orders have reflexes in multiple branches of Munda or daughters of Proto-Munda, but only the latter situation (Kherwarian-Kharia) reflects a known contact history. Furthermore, it appears that all four commonly used general negators in Munda have possible Austroasiatic etyma, but working out the details of such developments awaits a separate future study. The present study is merely the first step in a full-scale reconstruction and typology of negation and negative structures and constructions found across the Munda branch of Austroasiatic.

References

Anderson, Gregory D. S. 2001. A New Classification of Munda: Evidence from Comparative Verb Morphology. In *Indian Linguistics* 62:27–42.

Anderson, Gregory D. S. 2004. Advances in Proto-Munda Reconstruction. In *Mon-Khmer Studies* 34:159–184.

Anderson, Gregory D. S. 2006. *Auxiliary Verb Constructions*. Oxford: Oxford University Press

Anderson, Gregory D. S. 2007. *The Munda Verb. Typological Perspectives*. Berlin: Mouton de Gruyter.

Anderson, Gregory D. S. 2008. Gtaʔ. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 682–763. London: Routledge.

Anderson, Gregory D. S. 2011. Auxiliary Verb Constructions (and Other Complex Predicate Types): A Functional-Constructional Typology. *Language and Linguistics Compass* 5(11): 795–828.

Anderson, Gregory D. S. 2015. Prosody, phonological domains and the structure of roots, stems and words in the Munda languages in a comparative/historical light. *Journal of South Asian Languages and Linguistics* 2(2):163–183.

Anderson, Gregory D. S. 2016. Do Koraput Munda, Lower Munda, and even South Munda really exist? Once more on the still unresolved classification of the Munda languages. In *Multilingualism and Multiculturalism: Perceptions Practices and Policy*. ed. by Supriya Pattanayak, Chandrabhanu Pattanayak, and Jennifer Bayer. Delhi: Orient Blackswan.

Anderson, Gregory D. S. 2017. Polysynthesis in Sora, with special reference to noun incorporation. In *Handbook of Polysynthesis*. ed. by Michael Fortescue and Nicholas Evans. Oxford: Oxford University Press.

Anderson, Gregory D. S. 2018. Proto-Munda in Austroasiatic comparative and South Asian areal perspectives. To appear in *Proto-Austroasiatic Syntax*, ed. by Mathias Jenny and Paul Sidwell. Leiden: Brill.

Anderson, Gregory. D. S. and K. David Harrison. 2008. Sora. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 299–380. London: Routledge.

Anderson, G. D. S., T. Osada and K. D. Harrison. 2008. Ho and the other Kherwarian languages. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 195–255. London: Routledge.

Anderson, Gregory D. S. and Bikram Jora. Forthcoming. Introduction to the templatic verb morphology of Birhor. To appear in *Languages and Linguistics*.

Bodding, P. O. 1929. *Materials for a Santal Grammar, II (mostly morphological)*. Dumka: Santal Mission of Northern Churches.

Horo, Luke. 2017ms. Phonetic comparison of Orissa Sora and Assam Sora. Presented at 1st International Conference on Munda Languages and Linguistics, Deccan College Pune, March 2017.

Horo, Luke 2017. *A Phonetic Description of Assam Sora.* PhD Dissertation. Indian Institute of Technology, Guwahati.

Horo, Luke and Priyankoo Sarmah. 2015. Acoustic analysis of vowels in Assam Sora. *Northeast Indian Linguistics* 7, ed. by Linda Konnerth et al., 69–86. Canberra: Asia-Pacific Linguistics.

Jenny, Mathias, Tobias Weber, and Rachel Weymuth. 2015. The Austroasiatic Languages: A Typological Overview. In *The Handbook of Austroasiatic Languages*, ed. by Mathias Jenny and Paul Sidwell 13–143. Leiden: Brill.

Jenny, Mathias, Paul Sidwell and Mark Alves. 2018. Position paper: Austroasiatic syntax in diachronic and areal perspective. To appear in *Proto-Austroasiatic Syntax*, ed. by Mathias Jenny and Paul Sidwell. Leiden: Brill.

Jenny, Mathias and Paul Sidwell (eds.). 2015. *The Handbook of Austroasiatic Languages.* 2 volumes. Leiden: Brill.

Jora, Bikram and Gregory D. S. Anderson. 2017 ms-b. Introduction to Birhor (BirhoR) verb morphology. Presented at South-East Asian Linguistics Society (SEALS, 27) held in Padang, Indonesia, 11–13 May 2017.

Nagaraja, K. S. 1999. *Korku Language. Grammar, Texts, Vocabulary*. Tokyo: Tokyo University of Foreign Studies, Institute for the Study of Languages and Cultures of Asia and Africa.

Patnaik, Manideepa. 2008. Juang. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 508–556. London: Routledge.

Peterson, John M. 2008. Kharia. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 434–507. London: Routledge.

Peterson, John M. 2011. *Grammar of Kharia*. Leiden: Brill.

Pinnow, Heinz-Jürgen. 1966. A comparative study of the verb in Munda languages. In *Studies in Comparative Austroasiatic Linguistics*, ed. by Norman H. Zide, 96–193. The Hague: Mouton,.

Rau, Felix. 2018. The Proto-Munda Predicate and the Austroasiatic Language Family. To appear in Mathias Jenny and Paul Sidwell (eds.) *Proto-Austroasiatic Syntax*. Leiden: Brill.

Ring, Hiram and Gregory D. S. Anderson 2018. On prosodic structures in Austroasiatic diachrony: 'Rhythmic holism' revisited in light of preliminary acoustic studies on Khasian and Munda. To appear in J*SEALS special publication. Selected papers from ICAAL 7*.

Sidwell, Paul. 2015. Austroasiatic classification. In *The Handbook of Austroasiatic Languages*, ed. by Mathias Jenny and Paul Sidwell 144–220. Leiden: Brill.

Sidwell, Paul and Felix Rau. 2015. Austroasiatic comparative-historical reconstruction: an overview. In *The Handbook of Austroasiatic Languages*, ed. by Mathias Jenny and Paul Sidwell 221–363. Leiden: Brill.

Voß, Judith. 2017. Work in progress. Verbal morphology of Gutob. Presented at ICAAL 7, Kiel, Germany.

Zide, Norman H. 2008. Korku. In *The Munda Languages*, ed. by Gregory D.S. Anderson, 256–98. London: Routledge.

CORRELATIVE-RELATIVE CLAUSES IN MUNDA LANGUAGES: AN OVERVIEW

Jurica Polančec
University of Zagreb
jpolance@ffzg.hr

Abstract
The paper deals with the correlative-relative clauses (CRCs) in the Munda branch of Austroasiatic. Two types of CRCs are distinguished: headed CRCs and headless CRCs, the former being the main focus of this paper. Headed CRCs are attested in most Munda languages, with consensus that this construction was borrowed from Indo-Aryan (IA). Despite scarce evidence for most Munda languages, this article identifies a number of relevant characteristics of both headed and headless CRCs in Munda languages. For headed CRCs, there is variation in the form of the correlative pronoun as well as variation with respect to their degree of integration into the grammar of individual Munda languages. The latter issue awaits further research, but a number of preliminary remarks are made. As for the headless CRCs, which are also attested across the Munda branch, evidence can be adduced suggesting they are original in Munda (unlike headed CRCs). Evidence for this claim is found in the history of Dravidian languages, as well as in cross-linguistic tendencies.

Keywords: Munda languages; Indo-Aryan languages; correlative-relative clauses; syntactic borrowing; language contact

ISO 639-3 codes: sat, unr, hoc, kfq, srb, pcj, khr, jun, bfw, gbj, gaq

1 Introduction[1]

Munda languages are a branch of the Austroasiatic phylum spoken in eastern central India.[2] They are the westernmost Austroasiatic branch, and, together with the Meghalayan (Khasian, Khasic) and Nicobarese languages, the only Austroasiatic languages spoken outside the Mainland Southeast Asian linguistic area. At some point in their prehistory, the grammatical profile of the Munda languages underwent a dramatic restructuring, resulting in a profile typical of most South Asian languages (Donegan and Stampe 2004). The major features typical of this profile include the rise of the verb-final (OV) constituent order and the change from postnominal, fully finite (N-Rel) to prenominal, non-finite (Rel-N) relative clauses. The prenominal, non-finite relative clauses (RCs) are one of the two relative (attributive) constructions available in Munda languages, and are the more widely used of the two.[3] The other relativization strategy is headed correlative-

[1] The initial research on this topic was made possible by a generous grant from the Government of the French Republic in 2013. I would like to thank Denis Creissels and John Peterson for encouraging this research, as well as for their comments on an earlier version of this paper. The quality of the paper was considerably improved by the comments and suggestions made by two anonymous reviewers. Parts of this paper were presented at the Syntax of the World's Languages VI in Pavia in September 2014 and at the 7th International Conference on Austro-Asiatic Linguistics (ICAAL 7) in Kiel in September 2017. The comments and feedback from the participants of both conferences are gratefully acknowledged. I am particularly grateful to Judith Voß for sharing her data on Gutob, and John Peterson for providing information on Sadri and Nepali. Special thanks go to David Edel and Filip Medar for correcting my English. Any remaining errors are my own.

[2] For more information on the geographical distribution of Munda languages see Anderson (2015:364–365).

[3] The topic of prenominal relative clauses in Munda languages is relatively more complex, as it also provides important insights into the prehistory of Munda languages as a branch. Prenominal relative clauses will be the topic of a separate paper, and preliminary findings have been presented in Polančec (2017).

relative clauses (henceforth headed CRCs).[4] This relativization strategy, typical of Indo-Aryan (henceforth IA) languages, is borrowed into Munda languages from IA. In all likelihood, the spread of CRCs occurred fairly recently and support for such a claim is laid out in this paper.[5]

Due to the lack of extensive sources on CRCs of any kind in Munda (discussed below), claims presented here are taken to be provisional. For that reason, some of the claims made in this paper are further backed by citing the research on language contact and syntactic borrowing in other language groups. This in particular concerns the more amply documented instances of CRCs in Dravidian and Tibeto-Burman, the two other major groups of languages spoken in South Asia.

The ultimate goal of this overview is twofold. First, to present what is known about CRCs (both headed and headless) in Munda languages. Second, to touch upon various relevant issues that need to be addressed in future research on this topic. In that sense, this overview will hopefully benefit researchers working on the grammatical description of Munda languages, as well as those working on language contact in South Asia.

The paper is organized as follows: the remainder of this section provides a brief introduction to Munda languages, followed by a discussion of the sources this survey is based on; Section 1 introduces headed CRCs, focusing on the features characteristic of Indo-Aryan languages; Section 2 presents data from Munda languages; Section 3 discusses headless CRCs; finally, Section 4 concludes the paper with a summary of findings and suggestions for future research.

1.1 Munda languages

According to Ethnologue (21st edition, Simons and Fennig 2018) there are 23 Munda languages. Grammatical descriptions of any length are available for only a dozen of them. This study includes 11 such languages, which are listed in the following table:[6]

Table 1: *Munda languages included in the study.*

Language (alternative name)	*Affiliation (branch – subbranch)*[7]	*ISO 639-3*
Santali	(N Munda – Kherwarian)	sat
Mundari	(N Munda – Kherwarian)	unr
Ho	(N Munda – Kherwarian)	hoc
Korku	(N Munda – Korku)	kfq
Sora (Savara)	(S Munda – Sora-Gorum)	srb
Gorum (Parengi)	(S Munda – Sora-Gorum)	pcj
Kharia	(S Munda – Kharia)	khr
Juang	(S Munda – Juang)	jun
Remo (Bonda, Bondo)	(S Munda – Gutob-Remo)	bfw
Gutob (Bodo Gadaba)	(S Munda – Gutob-Remo)	gbj
Gtaʔ (Gata', Didayi)	(S Munda – Gtaʔ)	gaq

[4] In this paper the distinction between headed and headless CRCs is made. Headed CRCs are the main topic of the paper, whereas headless CRCs are covered in less detail (see also the next footnote). In most of the literature the two subtypes are not clearly distinguished (at least not terminologically). This paper observes the distinction consistently, both notionally and terminologically.

[5] As it will become clear in the course of the paper, an important claim made here is that only headed CRCs were borrowed from IA, whereas headless CRCs might be considered an original Munda construction.

[6] The languages not included here are numerous small Kherwarian languages (about a dozen languages/dialects), which are still largely undescribed, but are known to be close to the other three major Kherwarian languages (Santali, Mundari, Ho). Further, little is known about Juray, a South Munda language said to be close to Sora.

[7] Genetic affiliation within the Munda family here combines the traditional view that there is a division between the North and South Munda branches, but abandons the division of South Munda into two branches (Koraput and Kharia-Juang), following a more recent proposal by Anderson (2001). See also Anderson (2015:365-369).

1.2 Sources the study is based on

Kharia is singled out in ***Table 1*** for being the only Munda language described in recent times in a comprehensive grammar (Peterson 2011). Peterson's grammar contains an exhaustive description of Kharia syntax, including a thorough treatment of relative clause formation. A number of languages have been described in less exhaustive grammars (Mundari, Santali, Ho, Korku – the list of sources is found in §2), some of which are predominantly based on earlier sources (e.g. Neukom's 2001 grammar of Santali). South Munda languages (with the exception of Kharia) are mainly known from the grammar sketches in the 2008 volume *Munda Languages*, edited by G. Anderson (2008a; henceforth ML). ML is also a valuable source for all the other Munda languages represented there, considering that all the grammar sketches in that volume include a section on relative clauses.[8] The full list of sources providing the data used in this study is given in ***Table 3*** in §2.

One should be well aware that the level of detail presented in Peterson's 2011 grammar of Kharia is not matched in the sources for other Munda languages. The mentions of CRCs of any kind for other Munda languages are in most cases brief, with only a couple of examples provided. Such cursory presentations leave numerous open questions. The lack of exhaustive sources notwithstanding, we find the study of the kind presented here worthwhile. Crucially, despite the lack of comprehensive sources, we were able to present a number of preliminary findings which can be updated, amended or rejected in subsequent research as more data become available.

1 Headed correlative-relative clauses

Headed correlative-relative clauses (CRCs) are defined as a relative clause (RC) formation strategy in which "the head noun appears as a full-fledged noun phrase in the relative clause and is taken up again at least by a pronoun or other pronominal element in the main clause"[9] (Comrie 1998:62).[10] This is illustrated in the following example from Hindi, an Indo-Aryan language:[11]

3) Hindi (Lipták 2009a:1):[12]

[jo	*laṛkī*	*khaṛī*	*hai]*	*vo*	*lambī*	*hai*
CREL	girl	standing	is	that	tall	is

'The girl who is standing is tall.'

Lit. '[Which girl is standing], that is tall.'

---

8 The chapter on Kharia in ML by Peterson (2008) contains the same information presented in Peterson (2011).

9 As formulated, Comrie's definition covers only headed CRCs, but this is not made explicit in his text. This seems to reflect a convention found in most typological literature on RCs, whereby headed CRCs are simply referred to as CRCs. This practice can create confusion as it blurs the distinction between headed and headless CRCs.

10 The **head** is the referent of the NP whose reference is restricted by the relative clause. It is the participant shared by both the main clause and the relative clause. For instance, in the sentence *The book [I bought yesterday] was a trade paperback*, the head is *(the) book*, which is a participant in both the main clause – as the subject (*The book was a trade paperback*) and the RC – as the direct object (*I bought the book yesterday*).

11 In the remaining examples in this article we will use the following conventions: RCs will always be enclosed in square brackets; the head together with the RC marker (correlative pronoun or any other) will be underlined, as well as its co-referent in the main clause. Abbreviations used are: A/ACT active, ACC accusative, ANIM animate, AUX auxiliary, CNTR contrastive focus, COP copula, CREL correlative marker/pronoun, DU dual, ERG ergative, FIN finite, FOC focus, GEN genitive, HAB habitual, HUM human, INAN inanimate, INF infinitive, ITR intransitive, LOC locative, M/MID middle, N noun, NEG negation, NMLZ nominalizer, NOM nominative, NPST non-past, OBJ object, OBL oblique, OPT optative, PL plural, PLUP pluperfect, PRF/PERF perfect, PROG progressive, PST past, Q interrogative clitic, QUAL qualitative predication, REDPL reduplication, SEQ sequential converb, SFX suffix, SG singular, SUBJ subject, TOP topic

12 Transliteration has been slightly modified to reflect current practices. The Hindi phoneme transliterated here by the grapheme *j* is a voiced postalveolar affricate (IPA ʤ). The same grapheme is used in transcriptions of related phonemes in other IA languages as well as in Munda languages, even though the exact details of pronunciation may differ between them (e.g. see Peterson 2011:29 for Kharia).

In this example the RC *jo laṛkī khaṛī hai* is found to the left of the main clause *vo lambī hai*. The marker of the RC in this case is the correlative pronoun *jo,* which inflects for case and number (but not gender) in Hindi (Koul 2008:77). The correlative pronoun can be more generally referred to as a **correlative marker** (glossed as CREL), since in some languages this element can be uninflected. The correlative pronoun (or marker) *jo* is followed by and forms a constituent with the head *laṛkī*. The demonstrative *vo* found in the main clause is co-referent with the head.[13] The predicate found in the RC is the fully finite form *khaṛī hai*.

Another important property of headed CRCs is that they are considered adjoined to the main clause, that is, they "do not occupy a sentence-internal position corresponding to an argument/adjunct slot" (Lipták 2009a:7; cf. also Hendery 2012:17–19). Headed CRCs in Hindi are reported not to have any restrictions as to the syntactic slot that can be occupied by the head in the relative clause. This is also true for Indo-Aryan (IA) languages in general (Subbārāo 2012a:271–274). Headed CRCs in Hindi have the following set of features:

Table 2: *Major properties of Hindi headed CRCs.*

position of the RCs:	left-adjoined
type of relative clause marker:	correlative pronoun (*jo*)
type of verb in the relative clause:	finite
treatment of the head in the RCs:	full NP

Virtually all IA languages making use of headed CRCs share properties presented in ***Table 2***.[14] All of these properties can be contrasted with the European-type postnominal RCs, as in English (Andrews 2007:207):

4) *The dog bit <u>the man</u> [<u>who</u> was shouting]*.

In this type of RC, the head *the man* occurs in the main clause, and is represented in the RC by the pronoun *who*. The RC is postnominal and, unlike in Hindi, the head forms a constituent with the RCs (that is, it is embedded, and not adjoined). Specifically, the RC *who was shouting* forms a constituent with the head noun *(the) man*. As is the case with CRCs, the RC contains a fully finite predicate.

Headless CRCs are similar to headed CRCs, and will be introduced and discussed in §3. Both headed and headless CRCs should not be confused with the comparative correlative construction of the type *The more you read, the less you understand.* (Lipták 2009a:11, 18–21). These constructions are similar to CRCs, but are outside the scope of this paper.

Even though the left-adjoined structure illustrated in ex. 3) above is "the normal IA construction" (Masica 1991:412; cf. Hendery 2012:179), this prototypical structure can be modified in various ways. These modified, non-prototypical constructions will be discussed in more detail in §3 below, but only to the extent they are attested in Munda languages.

IA languages provide by far the most representative sample of languages with headed CRCs. Headed CRCs are common heritage of IA and were attested early on in (Vedic) Sanskrit (Speijer 2009 [1886]:347–379). As demonstrated for Hindi/Urdu by Davison (2009), the modern headed correlative structure is the result of a long-term development from a more paratactic-like old IA correlative structure of (Vedic) Sanskrit. Today, headed CRCs are a dominant strategy in almost all modern IA languages.[15] From IA, the

[13] The head can also be repeated after the demonstrative: *[<u>jo larkī</u> khaṛī hai] <u>vo larkī</u> lambī hai* (Lipták 2009a:3).

[14] This includes the repetition of the head in the main clause.

[15] There are IA languages where CRCs are absent, e.g. Sinhalese, long isolated from the rest of the Indo-Aryan languages (Chandralal 2010:63, 131–134, 195–197), and closely related Dhivehi (or Maldivian; Cain and Gair 2000:35–36). This can be attributed to contact with Dravidian (Masica 1991:408; Cain and Gair 2000:35). The same is true for IA varieties transplanted into predominantly Dravidian-speaking South India, such as Southern Konkani and Saurashtri (Lipták 2009a:10; Masica 1991:408). In Nepali correlative RCs exist, but appear to be less common than prenominal RCs (John Peterson, p.c.; Masica 1991:415; cf. Lipták 2009a:12–13).

headed CRCs have presumably spread to neighboring non-IA language groups: most importantly Munda, but also to Dravidian and Tibeto-Burman.[16]

On a final note, headed CRCs (illustrated in ex. 3) above) are uncommon in the world's languages outside South Asia and the IA group of languages, particularly as a major or main relativization strategy. According to the data in Dryer (2013), which takes into account only main/major relativization strategies, such constructions are attested outside South Asia only in a small area in West Africa. There, headed CRCs are by and large confined to the Mande family (Nikitina 2012; Kuteva and Comrie 2006).

2 The headed correlative-relative clauses (CRCs) in Munda

Headed CRCs in their prototypical (left-adjoined) form[17] are attested in most Munda languages. There is little doubt that they have been borrowed into Munda languages from IA languages (Subbārāo 2012a:312; Peterson 2011:425, fn. 26).[18] The borrowing of CRCs from IA into Munda languages constitutes an instance of what Appel and Muysken call grammatical borrowing or "incorporation of foreign rules into a language" (1987:153–154). It is also a fairly typical instance of contact situation (cf. Sakel 2007:21), whereby borrowing typically occurs from hierarchically higher or dominant languages (IA) into lower, dominated languages (Munda).[19] Headed CRCs are attested in the following Munda languages:

1 North Munda languages: Santali, Ho, Mundari and Korku;
2 South Munda languages: Kharia, Sora, and Juang.

It was not possible to confirm the existence of headed CRCs in four small tribal South Munda languages: Remo, Gorum, Gutob and Gtaʔ. In the latter two, only headless CRCs are attested (see §3). In this section, we review only the evidence of headed CRCs in the prototypical (left-adjoining) form. The few cases of non-prototypical headed CRCs attested in Munda will be mentioned briefly in §3. As to the possible reasons why CRCs are not attested in Remo, Gorum, Gutob and Gtaʔ, incomplete documentation is possible. However, this explanation may not suffice, at least in the case of Gutob, where no instances of headed CRCs have been attested after a recent thorough investigation (Judith Voß, p.c.).

An alternative explanation could be advanced, based on the assumption that the contact between IA and Munda languages has not been equally intense across different regions where Munda languages are spoken. In that respect, one may observe that headed CRCs are found in the Munda languages spoken in Jharkhand,[20] e.g. Santali, Ho and Kharia, all of which have large number of speakers. In contrast, all four languages without headed CRCs mentioned above are small tribal languages mainly spoken much further to the south, in the isolated hilly areas of the Koraput and Malkangiri districts of the Indian state of Odisha (Orissa) and in the neighboring areas of Andhra Pradesh.[21]

Seemingly, the absence of headed CRCs from small tribal languages could be explained by the fact that they have been in less intense contact with IA due to their relative isolation. This argument seems to be strengthened further by the fact that the larger languages of Jharkhand have many speakers in urban areas, for which extremely high rates of bilingualism have been reported.[22] This line of argument could also be

---

16 The extent to which headed CRCs are common in Dravidian and Tibeto-Burman is unclear. Most examples quoted in Subbārāo (2012a, 2012b) are headless CRCs (§3), not headed CRCs, though headed CRCs are also found.

17 This includes the possibility of repeating the head after the demonstrative in the main clause (see fn.13).

18 Some authors, such as Patnaik (2008:546) are noncommittal about this claim, but the evidence gathered in this paper, in my opinion, leaves little doubt about the IA origin of Munda CRCs.

19 According to Sakel, dominance can have different aspects: "a language is dominant when used for administration, as a lingua franca, and when it has to be learnt by the speakers of the dominated language" (2007:21). The dominant status of IA languages with respect to Munda and Dravidian tribal languages is briefly addressed in Abbi (1997:133–135) and Ishtiaq (1997:335–336).

20 There is extensive evidence of the longstanding and intense contact between IA and Munda languages in Jharkhand (Abbi 1995, 1997; Peterson 2010, 2017a, 2017b).

21 References are in fn. 24. Maps of this area are at http://ethnologue.com/map/IN_06 and in Anderson (2014:364).

22 Abbi (1997:134–135) suggests that near 85 percent of tribal language speakers in urban Jharkhand are bilingual.

extended so to imply that the evidence of headed CRCs in the languages of Jharkhand comes from speakers that live in urban areas and have therefore been exposed to a much larger extent to the IA influence.[23]

This explanation however has multiple weak points. First, we should mention that headed CRCs are barely registered in Mundari (see also §0 below), another large Kherwarian language with a considerable number of urban speakers, which is spoken in the same area as Santali, Ho and Kharia. Second, all Munda languages can be convincingly shown to have been under considerable influence of IA, regardless of their geographical position and sociolinguistic status. This includes the four tribal languages lacking CRCs.[24] Finally, headed CRCs are also attested in two other tribal languages, Sora and Juang, spoken much farther to the east and northeast in Odisha, respectively.

Therefore, the assumption that the four tribal languages (Remo, Gorum, Gutob and Gtaʔ), where headed CRCs are not attested, have been subject to less intense contact with IA, must be rejected. Instead, a more plausible explanation for the absence of CRCs in Remo, Gorum, Gutob and Gtaʔ may be linked to the apparent absence of some correlative pronouns (and consequently headed CRCs) in Desiya (Mathews 2003:58), a variety of Oriya spoken in southern Odisha.[25] Desiya is widely used as a second language among Munda speakers in that region and has accordingly had considerable influence on the Munda languages of the area (see fn. **Error! Bookmark not defined.**). This issue cannot be satisfactorily resolved here given our current level of knowledge about Munda languages and some of the languages they have been in contact with (in particular Desiya).

Instead, in what follows we will turn to further issues relevant for headed CRCs in Munda languages. In particular, we will address the differences with respect to the usage of CRCs revealed through a comparison of Munda languages. These differences concern the three following issues. The first issue has to do with the kind of pronoun used in the CRC, which can be either a native interrogative pronoun or the borrowed IA correlative pronoun or both (§0). The second issue concerns the extent to which the CRC structure has been integrated into the grammar of a language (§0). The third issue concerns when CRCs were borrowed into the respective Munda languages (§0). These three questions will now be discussed in turn.

The source of the correlative marker

In this section we discuss the origin of the correlative marker in Munda headed correlative-relative clauses (headed CRCs). As illustrated in ex. 3) from Hindi (§1), the correlative marker is the element preceding the head in the relative clause. In IA languages this element is called the correlative pronoun and is one of the *j*-series pronouns (Masica 1991:253, 410), often called 'relative pronouns' in IA literature. Such pronouns are called correlative here because of their special use in CRCs, distinct from interrogative (*k*-series) pronouns.

Munda languages lack such native pronouns, as do Dravidian and Tibeto-Burman languages. Thus Munda languages resort to two means of filling this gap. The first consists of replacing the IA correlative pronoun by a native element recruited from native interrogative pronouns. In the other, the IA correlative pronoun is borrowed. Munda languages are attested to make use of either only one of these two possibilities, or both. Below we illustrate the two patterns in examples from Santali (N Munda) and Kharia (S Munda). In Santali, the borrowed form *je* is illustrated in 5); in 6) the native interrogative *oka* 'which' is found:[26]

---

[23] The extent to which everyday language of urban Munda speakers, often educated in dominant languages such as English and/or Hindi, has been subject to change remains to be more thoroughly investigated, but initial reports such as Abbi (1997) on urban speakers in Jharkhand suggest significant influence of IA on their language.

[24] See Griffiths (2008:634, 670–671) for Gutob, Anderson and Harrison (2008a:557) for Remo, Anderson and Rau (2008:382) for Gorum, and Anderson (2008b:756) for Gtaʔ. Masica (1991:426–427), citing K. Mahapatra, reports that Desiya, a variety of the IA language Oriya, has been a "natural second language" of the tribal populations in Koraput District of Odisha from the 15th century.

[25] A note of caution is in place here as Mathews' (2003) sketch of Desiya offers little information about syntax. However, the coverage of morphology is quite detailed and the section on pronouns mentions some correlative items such as *ʤene* 'where' and *ʤar* 'whose', but no correlative items that would correspond to the Hindi correlative pronoun *jo.* In addition, the subsection on relative clauses (p. 16) mentions only prenominal RCs.

[26] Glosses have been slightly modified.

5) Santali (N Munda – Kherwarian; Ghosh 2008:84):
[je *hilok'* *uni-ɲ* *ɲɛl-led-e-a]*
CREL day 3SG-1SG.SUBJ see-PLUP.A-3SG.OBJ-FIN
un *hilok'* *dɔ* *sombar* *tahɛ̃kan-a*
that day TOP Monday COP.PST-FIN
'The day I saw him was Monday.'

6) Santali (N Munda – Kherwarian; Ghosh 2008:84):
[oka *disɔm-rɛ* *onko* *gaɖel* *hɔɽ-ko* *jarwa-akan-tahɛ̃kan-a]*
CREL country-LOC 3PL.SUBJ crowd man-3PL:SUBJ gather-PRF.M-COP.PST-FIN
ona *disɔm-rɛn* *raj* *dɔ* *gɔj-akan-a*
that country-GEN king TOP die-PRF.M-FIN
'The king of the country where the crowd of people had gathered has died.'

A parallel situation is found in Kharia, where headed CRCs are formed either with the borrowed correlative pronoun *je*, as in 7), or with the native interrogative pronoun serving as the correlative, as in 8):

7) Kharia (S Munda – Kharia; Peterson 2011:408–409):
[je *khajar* *tar=sikh=oʔ=may]* *ho=kaɽ=aʔ* *komaŋ=ko* *nalage, …*
CREL deer kill=PERF=ACT.PST=3PL that=SG.HUM=GEN meat=CNTR NEG.QUAL.PRS
'It isn't the meat of the deer that they had killed …'
Lit. '[Which deer they had killed], his meat it is not …'

8) Kharia (S Munda – Kharia; Peterson 2011:409):
[a=boʔ=te *pujapaɽh* *karay=na* *aw=ki],* *ho* *boʔ=te*
Q=place=OBL sacrifice do=INF QUAL=MID.PST that place=OBL
ɖam=ke *ho=ki* *ho* *ɖoli=te* *maɽay=oʔ=may.*
arrive=SEQ that=PL that palanquin= OBL put.down=ACT.PST=3PL
'Having arrived at the place where the sacrifice was to be done, they put the palanquin down.'

Table 3 summarizes our findings for all Munda languages with headed CRCs.[27] The exact form of the borrowed correlative pronoun depends on the IA variety that serves as the source of the borrowing. The variant *jo* is found in Standard Hindi.[28] The form *je* is found in Sadri, a Hindi variety that serves as the lingua franca of the part of eastern central India where Kharia and the languages of the Kherwarian group are spoken.[29] In the case of Korku, the source of *je* is in all likelihood the Marathi *dze*/*dʒe*, which is the neuter singular direct ("nominative") form of the correlative pronoun *dzo* (Pandharipande 1997:77).[30] In Juang, we

[27] Remo, Gorum, Gutob and Gtaʔ are thus not represented in the table. The latter two feature only headless RCs and are therefore discussed in §3. Kobayashi and Murmu report existence of the IA correlative pn in Keraʔ Mundari (2008:186), but give no examples. Keraʔ Mundari is thus not included in the table and is not discussed further.

[28] In Hindi the correlative pronoun inflects for number and case (Koul 2008:77).

[29] This is why Sadri is assumed to be the probable source of borrowing for Santali and Mundari, but we were unable to find any reference to back this claim. Unlike in Hindi (and Marathi), the correlative pronoun is uninflected in Sadri (Peterson and Kiran 2011; John Peterson, p.c.) and has no separate oblique stem. Generally speaking, there are no separate oblique stems for nouns and most pronouns in the IA languages of the eastern zone (Peterson 2017a). The same is the case in Oriya, another IA language of the eastern zone (see fn. **Error! Bookmark not defined.**38).

[30] In Marathi, the correlative pronoun inflects for gender (masculine/feminine/neuter), number (singular/plural) and case (direct ["nominative"]/oblique) (Pandharipande 1997:77; Dhongde and Wali 2009:52). Note also that the Marathi phonemes written as *dz* and *dʒ* here are very similar to the Hindi phoneme written as *j* (see fn. **Error! Bookmark not defined.**12). To be more precise, in Pandharipande's transliteration, *dz* represents the voiced (and unaspirated) alveolar affricate (IPA d͡z), whereas *dʒ* represents the voiced (and unaspirated) alveolo-palatal affricate (IPA d͡ʒ) (Pandharipande 1997:540). The latter (*dʒ*) is a variant of the former (*dz*) that arises through palatalization

find the form *ju*, which is probably a rendering of the Oriya correlative pronoun *jẽũ* (Neukom and Patnaik 2003:393).[31]

Table 3: *Munda languages with headed CRCs.*

Language	**IA CREL**	**Source of IA CREL**	**Native CREL**	**Reference**[32]
Santali	je	Sadri (?)	*oka*[33] *jãhae/jãhã*[34]	Ghosh (2008:84); Neukom (2001:199–200)
Mundari	je	Sadri (?)	oko	Osada et al. (2015:81)
Ho	n/a	n/a	*okon-*[35]	Burrows (1915:64); Deeney (1976:76); Koh and Subbārāo (ms), cit. in Subbārāo (2012b:116)
Korku	jo je	Hindi (for *jo*) Marathi (for *je*)	-	Zide (2008:290–291); Zide (2010:186–189)
Sora	n/a	n/a	*aieŋən*[36]	Anderson and Harrison (2008b:365); Starosta (1967:243)
Kharia	je	Sadri	*a=* 'Q'[37]	Peterson (2011:408–409)
Juang	ju	Oriya	-	Patnaik (2008:546)

The two patterns of borrowing in Munda can be characterized in terms of the distinction introduced by Y. Matras and J. Sakel between matter (MAT) and pattern (PAT) borrowing. These are "the two basic ways in which elements can be borrowed from one language into another" (Sakel 2007:15) and are defined by Sakel as follows (2007:15):

> We speak of MAT-borrowing when morphological material and its phonological shape from one language is replicated in another language. PAT describes the case where only the patterns of the other language are replicated, i.e. the organization, distribution and mapping of grammatical or semantic meaning, while the form itself is not borrowed. In many cases of MAT-borrowing, also the function of the borrowed element is taken over, that is MAT and PAT are combined.

The instances of CRCs in Munda making use of native interrogative pronouns (ex. 6) and 8)) constitute a clear instance of pattern (PAT) borrowing. In cases where IA correlative pronouns are attested instead of, or parallel to, native interrogative pronouns (ex. 5) and 7)), we are dealing with the combination of pattern (PAT) and matter (MAT) borrowing, the matter borrowed being the IA correlative pronoun of various forms.

(Pandharipande 1997:543-544). Wali (2005:16) has only the form *dʒe* (transliterated as *ǰe*), whereas Dhongde and Wali (2009:52) have only the form *dze* (transliterated as *je*).

31 Like in Sadri, the correlative pronoun in Oriya does not inflect for gender, and examples show that *jẽũ* is used with both animate and inanimate heads (Neukom and Patnaik 2003:394), and there is no separate oblique stem for it (Neukom and Patnaik 2003:46, 393). The plural marking appears to be optional.

32 The range of works surveyed for this paper is larger, but not all include information on correlative-relative clauses.

33 According to Ghosh (2008:43), *oka* is the inanimate form (the animate form is *ɔkɔe*). We found no examples of CRCs with the animate form.

34 From Ghosh (2008:83). These are originally indefinite pronouns (*jãhãe* [animate] and *jãhã* [inanimate]). Interestingly enough, Santali is the language in our sample where indefinite pronouns are used in headed CRCs (along with interrogative ones). The indefinite pronouns are also employed in headless CRCs (see §3).

35 According to Pucilowski (2013:199), the forms are *okon-iʔ/ko*, which is used with singular/plural animate NPs, and *okona*, which is used with inanimate NPs.

36 Segmented as *a-ieŋ-ən* by Anderson and Harrison (2008:365), but individual morpheme functions are unclear.

37 Bound morpheme *a=* combines "with free morphemes to derive other interrogatives" (Peterson 2011:178). In CRCs we find combinations such as *a=boʔ=te* [Q=place=OBL] 'where', *a=te* [Q=OBL] 'where', or *a=kaɽ* [Q=SG.HUM] 'who'.

The degree of integration

In the previous section we listed the Munda languages for which the headed CRC construction has been attested. Such a simple listing paints only a superficial picture of this phenomenon, as it does not provide information on the degree of integration of headed CRCs into the grammar of individual languages. As will become clear from what follows, there are significant differences among individual languages, even though many details remain unknown.

On the one end, there are Munda languages for which it can be claimed that headed CRCs are found only among educated bilingual speakers. For instance, headed CRCs are reported for Mundari only in a recent coursebook (Osada et al. 2015:81). They are said to be used by "educated bilingual people" and are rare in traditional narratives. Standard reference works on Mundari mention no CRCs (Osada 1992, Osada 2008). For Ho, Deeney notes that headed CRCs are used by "people who become accustomed to thinking in Hindi" (Deeney 1976:76).

The use of headed CRCs only during elicitations is reported for Korku and Gutob (Zide 2010:187). Zide remarks that such "instant-calques" are made by the informants in order to demonstrate to the linguist their knowledge of the more prestigious language (Hindi/Marathi in the case of Korku, and Oriya in the case of Gutob).[38]

The imitation of a prestigious IA variety is probably one of the ways headed CRCs entered at least some Munda languages and were accepted by at least some speakers. Crucial evidence is provided in the 1915 Ho grammar by Burrows, where he notes that headed CRCs are often utilized by those who want to imitate "more advanced language" (1915:64). We could assume that, later on, as education became more widespread and awareness of the prestige of IA languages became more established, the construction became more frequent and more integrated in the grammars of individual languages.[39] This assumption requires further corroboration.

Appel and Muysken (1987:162) find "imitation of a prestige language pattern" an important factor in the borrowing of RC formation strategies, documented all over the world, for instance in Turkish, Nahuatl and Quechua. According to them, this kind of syntactic borrowing is rather superficial because "[o]nly aspects of the grammar that are easily perceived can be imitated" (1987:158).

Furthermore, there are languages where the available data appears to demonstrate that headed CRCs are fully integrated into the language.[40] Making such a claim requires a thorough investigation of relativization strategies, and the only language for which this is investigated is Kharia. One way to assess the extent to which a relative construction is integrated into the language in question is by looking at its productivity with respect to the syntactic positions available for relativization. According to the table presented in Peterson (2011:410), headed CRCs are quite productive with respect to the position relativized.[41] This is not unexpected, as headed CRCs can relativize any position in IA (see §1 above), as well in Dravidian and Tibeto-Burman (Subbārāo 2012a:278).[42]

As for Santali and Sora, two languages that have not yet been touched upon, sources do not comment on the productivity of CRCs, but the construction is well attested, though it is unclear how widespread it is. Moreover, the case of Santali is interesting because it is one of the two languages in our sample (the other being Ho) for which CRCs were attested early on. Older material on Santali published by Boding in the 1920s and 1930s and condensed in Neukom's 2001 short grammar attest that only native interrogative pronouns were employed in headed CRCs (Neukom 2001:199), whereas newer material presented in Ghosh (2008) shows that nowadays borrowed IA correlative pronouns can be found as well (see ex. 5) above).

---

38 The prestige of Hindi among speakers of tribal languages in India is well attested (see e.g. Abbi 1997:135)

39 What complicates these matters in the case of Ho is the fact that in a recent survey of the aspects of Ho grammar by Pucilowski (2013) no instances of left-adjoined CRCs are attested. Since it is not a comprehensive account of Ho grammar, it could well be that these constructions simply did not come up in the material collected by the author.

40 When this happens, headed CRCs are expected to come into competition with prenominal RCs, the native Munda relative construction (cf. §59 above). This question is not addressed in any of the works we have consulted, and therefore will not be pursued here for lack of information.

41 Peterson's data also shows that there are some minor differences in the productivity of the construction with native interrogative pronouns and the construction with the borrowed correlative pronoun.

42 Subbārāo claims the same for Munda in general but cites no evidence for such a claim.

Hopefully, this section has demonstrated that much more research on individual languages is needed before more generalizations can be drawn regarding the factors governing the use of CRCs in Munda languages. In the following section, we discuss the age of borrowing of CRCs into Munda languages. This question is closely related to the one discussed in this section since a recent date of borrowing may entail that the CRC construction has not had much time to become integrated into the linguistic system of a language. However, this is another hypothesis awaiting further research.[43]

How recent are headed CRCs?

The third question we asked at the beginning of §2 concerns the date of borrowing of headed CRCs into Munda languages. As Munda languages have been attested only since the late 19th century, we cannot ascertain if CRCs had been borrowed prior to that date. However, some evidence presented above, i.e. the practice of imitating a more prestigious language by Ho speakers recorded by Burrows (1915), suggests that the borrowing and spread of CRCs may have been occurring from the beginning of the 20th century onwards.

This would coincide with the gain in the prominence of IA languages that has occurred since that time, and that made them the dominant languages with a prestigious, official status (cf. Abbi 1997:133–135). Admittedly, IA languages had been the culturally and politically dominant languages for a long time before that (Peterson 2017a:217–222), but the prestigious status of the major IA languages and English has been on the rise due to multiple reasons. These include the spread of education and growth of literacy, which entails second-language learning, urbanization, and rapid industrialization (Abbi 1997:135; Ishtiaq 1997:335–336). In addition, the intensity of language contact has dramatically increased across the world due to globalization (Sakel 2007:26). This implies that some languages nowadays have a much larger influence over speech communities than before.[44] Former colonial languages as well as languages that dominate the media are or particular concern here. As Sakel puts it, we observe "the rise of these already highly dominant, major languages through increased bilingualism" (2007:26).

In the context of our investigation, the result of these tendencies was a further increase in Munda-IA bilingualism (cf. Abbi 1997:133). Presumably, this has played a crucial role in the borrowing of headed CRCs, as there is an assumption in the literature that PAT-borrowing, of which headed CRCs are instances, can only occur if there is a degree of oral bilingualism (Sakel 2007:25). Another assumption is that the adoption of the IA correlative pronoun in place of native interrogative pronoun correlates with more intense influence of IA languages, but this claim needs further corroboration (see fn. 44**Error! Bookmark not defined.**).

In conclusion, it seems reasonable to infer that borrowing is quite recent (the beginning of the 20th century onwards) even in the languages which have been using headed CRCs the longest, such as Santali (and probably Ho). A scenario of spread stressing the role of bilingual educated speakers was outlined above.

Variations on the prototypical left-adjoining structure

In this subsection we only briefly remark on non-prototypical headed CRCs. Two types of variation regarding the prototypical left-adjoining structure were found.

Alternative positions of the headed CRCs

Headed CRCs in IA languages are not found only in the left-adjoined position, as it was the case in ex. 3) from Hindi above and in all of the above examples from Munda languages. They can also occur in the right-adjoined position, as well as in the postnominal position.[45] The latter structure uses the IA correlative pronoun to form a structure akin to the European-type embedded RCs, as in 4) above. Note that, strictly

[43] Another correlation that would be worth investigating concerns the possibility that the adoption of the IA correlative pronoun can be correlated with a greater degree of integration of CRCs in individual Munda languages.

[44] Cf. the pressure on children to learn Hindi among Ho speakers mentioned by Pucilowski (2013:4).

[45] For further variations on the basic left-adjoining structure in Hindi see Lipták (2009a:1–6), Srivastav (1991:641–652), Bhatt (2003:3–4). For Oriya see Neukom and Patnaik (2003:393–399) and for Marathi see Pandharipande (1997:78–80, 85–86). Further examples can be found in Masica (1991:411–412).

speaking, postnominal embedded RCs are not CRCs, but they are grouped together with CRCs here because they are considered a variant of the basic left-adjoining structure in IA.

The left-adjoined headed CRCs are the basic construction in IA (§1), but alternative orders are becoming increasingly common, perhaps under the influence of English which uses a postnominal embedded construction. Therefore, the question is asked if any of these variations are available in Munda languages, or in any of the two other major non-IA families of South Asia.

In Munda languages, postnominal RCs are attested in Kharia (Peterson 2011:421–422),[46] and in Ho (Pucilowski 2013:199–200). Peterson provides two examples (his examples 207 and 209), the first of which is cited here (note the IA correlative pronoun *je*):

9) Kharia (S Munda – Kharia; Peterson 2011:421–422):

ro	*brahman=ki*,	*[je*	*tama*	*pujapaʈh karay=te=may],*
and	Brahman=PL	CREL	now	sacrifice do=ACT.PRS=3PL
ho=ki	*ɖoli=te*	*goʔ=na*		*laʔ=ki=may.*
that=PL	palanquin=OBL	carry=INF		IPFV=MID.PST=3PL

'And the Brahmans, who now do sacrifices, they used to carry the palanquin.'

Pucilowski provides four examples, the first of which is cited here.

10) Ho (N Munda – Kherwarian; Pucilowski 2013:199):

coke	*en*	*gles-re*	*kanju-aka-n-a*	*[okon-a*	*tebul-re*	*em-aka-n-a]*
frog	that	glass-LOC	throw.into-PRF-ITR-FIN	what-INAN	table-LOC	put-PRF-ITR-FIN

'the frog has thrown himself into the glass which is put on the table'

Neither of the two descriptions addresses these constructions in much detail, and the constructions seem to be marginal at best. Once again, further research is needed. As for Dravidian and Tibeto-Burman families, Subbārāo (2012b:155) reports that such constructions are not allowed in neither of the two.

This issue would make an interesting topic of research because we would expect variations on prototypical construction only in cases of extreme contact and strong exposure to an IA variety. Variations on the basic left-adjoining structure are used for various expressive purposes in IA, and it would be unexpected to find that expressive potential of IA headed CRCs transferred into Munda.

Omission of the demonstrative phrase in the main clause

In recent fieldwork on Ho, Koh and Subbārāo (ms), cited in Subbārāo (2012b:116), observed that the occurrence of the demonstrative phrase in Ho headed CRCs is optional. In the following example, the demonstrative phrase *en cakūi* is enclosed in brackets to indicate its optionality:

11) Ho (N Munda – Kherwarian, Koh and Subbārāo (ms), cited in Subbārāo 2012b:116):

[okon	*cakūi-te*	*proj*	*ūtu-ko*	*hāḍe-tan-a]*	*(en cakūi)*	*leser-a*
CREL	knife-with	they	vegetable-3PL	cut-PROG-FIN	that knife	sharp-FIN

'The knife with which they are cutting the vegetable is sharp.'

In prototypical left-adjoined CRCs, the demonstrative is obligatory and the noun is not (see ex. 3) above). However, there are exceptions and omission of both elements is possible under certain conditions, e.g. in Hindi (Bhatt 2003:35–38) and Oriya (Neukom and Patnaik 2003:394), but not, apparently, for other Munda languages. Subbārāo reports that such omissions are not found in Tibeto-Burman or Dravidian (2012a:276).

[46] Postnominals are also attested in an earlier description of Kharia by Malhotra, as reported by Peterson (2011:408, fn. 21), as well as in Abbi (1997:143–144). Note that Kharia also provides the only possible example of a right-adjoined CRC (Peterson (2011:422, ex. 208), but that interpretation is not certain.

3 Headless correlative-relative clauses

The difference between prototypical headed CRCs and headless CRCs concerns the absence of the head after the correlative marker (CREL). In headless CRCs the head is not explicitly stated (cf. Andrews 2007:213; Epps 2012:192–193).[47] Headless CRCs are illustrated in the following examples from Hindi. In 12) we see a headless CRC with indefinite reference (reinforced by the use of the particle *bhī*), whereas 13) has definite reference. In both examples there is no noun after the correlative pronoun *jo*, as there was in ex. 3) above.

12) Hindi (Montaut 2005:237):[48]

[jo	*bhī*	*usne*	*tumse*	*kahā]*	*vah*	*sahī*	*hai*
CREL	ever	3SG.ERG	you.to	said	that	true	is

'Whatever he told you is true.'

13) Hindi (McGregor 1995:51):[49]

[jo	*kahtā*	*hũ]*	*vah*	*sac*	*hai*
CREL	saying	am	that	true	is

'What I say is true.'

Headless CRCs are attested in the following Munda languages (Table 4):

Table 4: *Headless CRCs in Munda languages.*

Language	Source
Sora	Bai (1985:188–189), Starosta (1967:242–243)
Santali	Neukom (2001:200), Ghosh (2008:83)
Kharia	Peterson (2011:409)
Gutob	Griffiths (2008:644–645, 667–668)
Gtaʔ	Anderson (2008b:709)

The first examples to be cited are from Sora, Santali and Kharia, with Gtaʔ and Gutob being discussed afterwards. In Sora, headless CRCs are well attested and are formed with the native marker *etente* (*itente* in Bai), which contains the interrogative pronoun *ete(n)-* 'what'.[50]

14) Sora (S Munda – Sora; Anderson and Harrison 2008b:365, ex. 213c):[51]

dɔ	*[etente*	*j-ən-om-jom-ən*	*ɲaŋ-tɛ-ji]*	*kun*	*batte*	*aninji*	*mɛɛŋ-tɛ-ji*
so	what	eat-NMLZ-eat-N.SFX	get-NPST-3PL	that	with	they	live-NPST-3PL

'so whatever food they get that's what they live on'

In Santali, headless CRCs are attested with the native indefinite pronouns *jãhãe* 'whoever' (animate) and *jãhã* 'whatever' (inanimate).[52]

---

[47] Another widely used term for headless CRCs is free (headless) relatives. However, this is imprecise, as headless CRCs are only a subtype of free relatives. For instance, free relatives do not necessarily contain the co-referent demonstrative in the main clause, as in *[Whoever comes] will be welcome.* They can also be postnominal, as in *You can invite [whoever you want]*.

[48] Glosses have been slightly modified.

[49] Glosses are ours.

[50] This fact is nowhere stated as such, but can be inferred from various places in Anderson and Harrison (2008b). There the form *etente* is either glossed as 'what' (e.g. on p. 325, ex. 79b or on p. 345, ex. 140a) or analyzed as *ete-n-te*, where *ete-* is 'what', *-n-* an unidentified nominal suffix and *-te* a focus marker (e.g. on p. 365, ex. 213c).

[51] Glosses have been slightly modified. The example is originally from Starosta (1967:243, ex. 4).

[52] It appears that the indefinite pronoun spread from such contexts to the prototypical headed CRCs (see fn. **Error! Bookmark not defined.**31 above).

15) Santali (N Munda – Kherwarian; Neukom 2001:200):
[nukin *jãhãe-ge-kin* *hɔrɔk'-a-e]*
these(ANIM).DU whoever-FOC-3DU.SUBJ put-APPL-3SG.OBJ
uni-ge *raj-e* *hoe-y-ok'-a*
that(ANIM)-FOC king-3SG.SUBJ become-y-MID-FIN
'Whomever they put it (i.e. a chain) on, he shall be king.'

In Kharia, the headless CRCs exemplified in Peterson (2011:409) contain native interrogative pronouns.

16) Kharia (S Munda – Kharia; Peterson 2011:421–422):
[a=kaɽ *seŋ* *ɖaʔ* *kuy=e],* *ho=kaɽ* *u*
Q=SG.HUM first water find=ACT.IRR that=SG.HUM this
daru=te=ga *yo=ga* *ɖe=na".*
tree=OBL=FOC see=FOC come=MID.IRR
'He who first finds water, he should come to this tree, looking [for the others].'

It may be recalled that in Gtaʔ and Gutob, headless CRCs are the only instances of CRCs attested. Headless CRCs with indefinite reference are illustrated in the following example from Gtaʔ:

17) Gtaʔ (S Munda – Gtaʔ, Anderson 2008b:709):
[ja *par-le]* *mœ* *paŋ* *ccoŋ* *diŋ-le*
CREL can-OPT he come REDPL:eat AUX-OPT
'Whoever wins, let him come and eat.'
The form *ja*, glossed here as CREL is in fact the interrogative animate pronoun (Anderson 2008b:707). One can assume it is unrelated to the Indo-Aryan (IA) correlative pronoun, given the fact that an identical form is found in Remo (Anderson and Harrison 2008a:579).[53] Headless CRCs with definite reference are illustrated with the following example from Gutob, where the native interrogative pronoun *laj* is employed:[54]

18) Gutob (S Munda – Gutob-Remo, Griffiths 2008:668):
[laj *mara + mari* *ɖeŋ-gu-men]* *o-maj* *razi* *ɖem-to*
CREL beat + ECHO become-MID.PST-PL OBJ-3.PL agreed make-HAB
'He makes those who have fought with each other settle their dispute.'

A separate discussion of headless CRCs was warranted mostly because there appears to be evidence suggesting that this construction is original in Munda, and not borrowed from IA.[55] This claim is supported by the fact that headless CRCs are found in numerous languages where there are no headed CRCs, including Dravidian and Tibeto-Burman (see fn. **Error! Bookmark not defined.** above), as well as in languages from other parts of the world.[56] This shows that headless CRCs are more common than headed CRCs, which are restricted to a very small number of languages.[57]

---

[53] Or it is borrowed in both languages from IA, as suggested by one of the reviewers.

[54] Griffiths (2008:667) claims that either a native or borrowed item can be utilized as the correlative pronoun in Gutob. No example is given of the construction with the borrowed item, and Judith Voß (p.c.) informs me that she has found no instances of the IA correlative pronoun in her corpus.

[55] Of course, the IA origin is still possible. Headless CRCs are well attested in IA languages, e.g. in Marathi (Pandharipande 1997:86-87) and Hindi – in addition to the examples cited as ex. 12) and 13) above, there are numerous instances of headless CRCs found, e.g. in McGregor (1995:47ff., 91ff.), Jain (1995:312–313), and Montaut (2004:235–238).

[56] This can be inferred from numerous examples cited by Lipták (2004) for Hungarian, Rebuschi (2009a, 2009b) for Basque and Cable (2009) for Tibetan.

[57] The headed CRCs are crosslinguistically rare, in particular as a major or main relativization strategy (this was already pointed out in §1).

A claim essentially identical to ours here, namely that headless CRCs should not be considered a borrowing from IA, is promoted for Dravidian by Bai (1985). The author claims that headless CRCs are a native Dravidian construction, which is based on multiple arguments, including the attestation of headless CRCs in the earliest attested Dravidian texts, as well as in various tribal Dravidian languages. Bai briefly discusses Munda languages as well, specifically headless CRCs in the tribal language Sora, extending her claim so as to suggest the native origin of headless CRCs in Sora. The native origin of Munda headless CRCs seems to be further supported by the observation that all the examples of that construction in Munda we have encountered so far contain only native interrogative pronouns (see above). Interestingly, according to the examples from Subbārāo (2012a, 2012b), the same is true for Dravidian headless CRCs.[58]

The native origin of Munda headless CRCs could help us reconstruct the manner in which borrowing of CRCs into Munda languages played out, as the prior existence of headless CRCs in Munda may have facilitated the borrowing of the IA headed CRCs into Munda by serving as a so-called "common pivot" (Matras and Sakel 2007).

4 Conclusions and prospects

This paper presented an overview of the correlative-relative clauses (CRCs) in the Munda branch of the Austroasiatic family. A distinction between headed and headless CRCs is drawn. Out of 11 Munda languages included in the survey, headed CRCs have been attested in seven. These include the North Munda languages Santali, Ho, Mundari and Korku, as well as the South Munda languages Kharia, Sora, and Juang. Headless CRCs, on the other hand, are attested in five Munda languages (Sora, Santali, Kharia, Gutob, and Gtaʔ).

The major areas of variation between individual Munda languages have been established. This includes variation in the form of the correlative marker (which can be a native interrogative pronoun or a borrowed IA correlative pronoun), as well as variation with respect to the closely related issues of the age of borrowing and the degree of integration of headed CRCs into the grammar of individual Munda languages. It is claimed that only headed CRCs are borrowed from IA, whereas headless CRCs are assumed to be original in Munda. It is hypothesized that the headless CRCs may have facilitated the borrowing of the headed CRCs.

The scarcity of the data on which this overview is based is emphasized throughout the paper. Thus, the conclusions presented should be considered only as preliminary, since almost every aspect of Munda CRCs requires further research. In that respect, this paper will hopefully serve as a starting point for future research in language-specific studies, as well as in subsequent comparative research benefitting from those studies.

References

Abbi, Anvita. 1995. Language contact and language restructuring: A case study of tribal languages in Central India. *International Journal of the Sociology of Language* 116(1):175–185.

Abbi, Anvita. 1997. Languages in contact in Jharkhand. In *Languages of tribal and indigenous peoples of india: The ethnic space*, ed. by Anvita Abbi, 131–148. Delhi: Motilal Banarsidass Publishers.

Anderson, Gregory D. S. 2001. A new classfication of Munda: Evidence from comparative verb morphology. *Indian Linguistics* 62:27–42.

Anderson, Gregory D. S. (ed.). 2008a. *The Munda languages*. Routledge Language Family Series. London: Routledge.

Anderson, Gregory D. S. 2008b. Gtaʔ. In Anderson 2008a, pp. 682–763.

Anderson, Gregory D. S. 2014. Overview of the Munda Languages. In The Handbook of Austroasiatic Languages (2 vols), ed. by Mathias Jenny and Paul Sidwell, 364–414. Leiden: Brill. doi:10.1163/9789004283572_006.

Anderson, Gregory D. S. and Felix Rau. 2008. Gorum. In *The Munda languages*, ed. by Gregory D. S. Anderson, 381–433. Routledge Language Family Series. London: Routledge.

Anderson, Gregory D. S. and K. David Harrison. 2008a. Remo (Bonda). In *The Munda languages*, ed. by Gregory D. S. Anderson, 557–632. Routledge Language Family Series. London: Routledge.

[58] The evidence from Tibeto-Burman is too scarce to allow for any conclusions; see Subbārāo (2012b:112–114).

Anderson, Gregory D. S. and K. David Harrison. 2008b. Sora. In *The Munda languages*, ed. by Gregory D. S. Anderson, 299–380. Routledge Language Family Series. London: Routledge.

Andrews, Avery. 2007. Relative clauses. In *Language typology and syntactic description. Volume 2: Complex constructions.* 2nd edition, ed. by Timothy Shopen, 206–236. Cambridge: Cambridge University Press.

Appel, René and Pieter Muysken. 1987. *Language contact and bilingualism*. London: Arnold. (Republished in 2005 by Amsterdam Academic Archive).

Bai, B. Lakshmi. 1985. Some notes on correlative constructions in Dravidian. *Oceanic Linguistics Special Publications* 20:181–190.

Bhatt, Rajesh. 2003. Locality in correlatives. *Natural Language and Linguistic Theory* 21(3):485–541.

Burrows, Lionel. 1915. *Grammar of the Ho language: An eastern Himalayan dialect*. London: Trubner and Company.

Cable, Seth. 2009. The syntax of the Tibetan correlative. In *Correlatives cross-linguistically*, ed. by Anikó Lipták, 195–222. Amsterdam: Benjamins.

Cain, Bruce D. and James W. Gair. 2000. Dhivehi (Maldivian). Munich: Lincom Europa.

Chandralal, Dileep. 2010. Sinhala. Amsterdam: Benjamins.

Comrie, Bernard. 1998. Rethinking the typology of relative clauses. *Language Design* 1:59–86.

Davidson, Alice. 2009. Correlative clause features in Sanskrit and Hindi/Urdu. In *Historical syntax and linguistic theory*, ed. by Paola Crisma and Giuseppe Longobardi, 271–291. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199560547.001.0001.

Deeney, J. S. J. 1975. *Ho grammar and vocabulary*. Chaibasa, Bihar: Xavier Ho Publications.

Donegan, Patricia and David Stampe. 2004. Rhythm and the synthetic drift of Munda. In *The Yearbook of South Asian languages and linguistics*, ed. by Rajendra Singh, 3–36. Berlin: Mouton de Gruyter.

Dryer, Matthew S. 2013. Order of Relative Clause and Noun. In *The world atlas of language structures online*, ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/90 (accessed: September 22, 2018).

Epps, Patience. 2012. Between headed and headless relative clauses. In *Relative clauses in languages of the Americas: A typological overview*, ed. by Bernard Comrie and Zarina Estrada-Fernández, 191–212. Amsterdam: Benjamins. doi:10.1075/tsl.102.09epp.

Ghosh, Arun. 2008. Santali. In *The Munda languages*, ed. by Gregory D. S. Anderson, 11–98. Routledge Language Family Series. London: Routledge.

Griffiths, Arlo. 2008. Gutob. In *The Munda languages*, ed. by Gregory D. S. Anderson, 633–81. Routledge Language Family Series. London: Routledge.

Hendery, Rachel. 2012. *Relative clauses in time and space: A case study in the methods of diachronic typology*. Amsterdam: Benjamins.

Ishtiaq, M. 1997. Typology of language change and maintenance among the Santals and Mundas. In *Languages of tribal and indigenous peoples of india: The ethnic space*, ed. by Anvita Abbi, 335–46. Delhi: Motilal Banarsidass Publishers.

Jain, Usha R. 1995. *Introduction to Hindi grammar*. Berkeley, CA: Centers for South and Southeast Asia Studies, University of California.

Kobayashi, Masato and Ganesh Murmu. 2008. Kera? Mundari. In Anderson 2008a, pp. 165–94.

Koh, T. J. and Kārumūri V. Subbārāo. ms. *A grammar of Ho*.

Koul, Omkar N. 2008. *Modern Hindi grammar*. Springfield, VA: Dunwoody Press.

Kuteva, Tania and Bernard Comrie. 2006. The typology of relative clause formation in African languages. In *Studies in African linguistic typology*, ed. by F. K. Erhard Voeltz, 209–28. Amsterdam: Benjamins.

Lipták, Anikó. 2004. On the correlative nature of Hungarian left-peripheral relatives. In *Proceedings of the Dislocated Elements Workshop (ZAS Berlin; November 2003)*, ed. by B. Shaer, W. Frey and C. Maienborn (eds.), 287–313. Berlin: ZAS.

Lipták, Anikó. 2009a. The landscape of correlatives: An empirical and analytical survey. In *Correlatives cross-linguistically*, ed. by Anikó Lipták, 1–46. Amsterdam: Benjamins. doi:10.1075/lfab.1.02lip.

Lipták, Anikó (ed.). 2009b. Correlatives cross-linguistically. Amsterdam: Benjamins. doi:10.1075/lfab.1.

Masica, Colin P. 1991. *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Mathews, Susan. 2003. *Aspects of Desiya grammar*. Asha Kiran Society.

Matras, Yaron and Jeanette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language* 31(4):829–865. doi:10.1075/sl.31.4.05mat.

McGregor, R. S. 1995. *Outline of Hindi grammar: with exercises*. Third edition, revised and enlarged. Oxford: Oxford University Press.

Montaut, Annie. 2004. *A Grammar of Hindi*. Munich: Lincom Europa.

Neukom, Lukas. 2001. *Santali*. Munich: Lincom Europa.

Neukom, Lukas and Manideepa Patnaik. 2003. *A Grammar of Oriya*. Zürich: Universität Zürich.

Nikitina, Tatiana. 2012. Clause-internal correlatives in Southeastern Mande: A case for the propagation of typological rara. *Lingua* 122(4):319–334. doi:10.1016/j.lingua.2011.12.001.

Osada, Toshiki, Madhu Purti, Nishaant Choksi and Nathan Badenoch. 2015. *A Course in Mundari*.

Osada, Toshiki. 1992. *A reference grammar of Mundari*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Osada, Toshiki. 2008. Mundari. In *The Munda languages*, ed. by Gregory D. S. Anderson, 99–164. Routledge Language Family Series. London: Routledge.

Pandharipande, Rajeshwari V. 1997. *Marathi*. London: Routledge.

Patnaik, Manideepa. 2008. Juang. In *The Munda languages*, ed. by Gregory D. S. Anderson, 508–556. Routledge Language Family Series. London: Routledge.

Peterson, John and Savita Kiran. 2011. Sadani /Sadri jazyk. In *Jazyki mira: novye indoarijskie jazyki*, ed. by Tatiana I. Oranskaia, , Yulia V. Mazurova, Andrej A. Kibrik, Leonid I. Kulikov and Aleksandr Y. Rusakov, 367–379. Moskva: Academia. English version: http://www.southasiabibliography.de/uploads/Sadri.pdf (accessed: September 22, 2018).

Peterson, John. 2008. Kharia. In *The Munda languages*, ed. by Gregory D. S. Anderson, 434–507. Routledge Language Family Series. London: Routledge.

Peterson, John. 2009. Language contact in Jharkhand: Linguistic convergence between Munda and Indo-Aryan in eastern-central India. *Himalayan Linguistics* 9(2):56–86.

Peterson, John. 2010. *A grammar of Kharia: A South Munda language*. Leiden: Brill.

Peterson, John. 2017a. Fitting the pieces together – Towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics* 4(2):211–257. doi:10.1515/jsall-2017-0008

Peterson, John. 2017b. Jharkhand as a 'linguistic area': language contact between Indo-Aryan and Munda in eastern-central South Asia. In *The Cambridge Handbook of Areal Linguistics*, ed. by Raymond Hickey, 551–574. Cambridge: Cambridge University Press. doi:10.1017/9781107279872.021.

Polančec, Jurica. 2017. On the history of Munda prenominal relative clauses. Paper presented at the 7th International Conference on Austro-Asiatic Linguistics (ICAAL 7), Kiel, Germany (September 29 - October 1, 2017).

Pucilowski, Anna. 2013. *Topics in Ho morphophonology and morphosyntax*. Eugene: University of Oregon PhD Thesis. https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/13241/Pucilowski_oregon_0171A_10666.pdf (accessed: September 22, 2018).

Rebuschi, Georges. 2009a. Basque correlatives and their kin in the history of Northern Basque. In *Correlatives cross-linguistically*, ed. by Anikó Lipták, 81–130. Amsterdam: Benjamins. doi:10.1075/lfab.1.05reb.

Rebuschi, Georges. 2009b. Position du basque dans la typologie des relatives corrélatives. *Langages* 174(2):25–38. doi:10.3917/lang.174.0025.

Sakel, Jeanette. 2007. Types of loan: matter and pattern. In *Grammatical borrowing in cross-linguistic perspective*, ed. by Yaron Matras and Jeanette Sakel, 15–29. Berlin: Mouton de Gruyter.

Simons, Gary F. and Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the world.* Twenty-first edition. Dallas: SIL International. Online version: http://www.ethnologue.com (accessed: September 22, 2018).

Speyer, J. S. 1886. Sanskrit Syntax. Repr. 2009. Delhi: Motilal Banarsidass.

Srivastav, Veneeta. 1991. The syntax and semantics of correlatives. *Natural Language and Linguistic Theory* 9(4):637–86. doi:10.1007/BF00134752.

Starosta, Stanley. 1967. *Sora syntax: A generative approach to a Munda language*. University of Wisconsin PhD Thesis.

Subbārāo, Kārumūri V. 2012a. *South Asian languages: A syntactic typology*. Cambridge: Cambridge University Press.

Subbārāo, Kārumūri V. 2012b. *South Asian languages: A syntactic typology*. Web Material. Cambridge: Cambridge University Press. http://www.cambridge.org/fr/download_file/138991/ (accessed: September 22, 2018).

Zide, Norman H. 2008. Korku. In *The Munda languages*, ed. by Gregory D. S. Anderson, 256–98. Routledge Language Family Series. London: Routledge.

Zide, Norman H. 2010. Review of *Korku Language: Grammar, Texts and Vocabulary*, by K.S. Nagaraja, 1999. *Mon-Khmer Studies* 39:177–92.

A PHONOLOGICAL ANALYSIS OF RIANG LANG

Elizabeth Hall
Payap University Linguistics Institute
ellie_hall@sil.org

Abstract
Riang is a tonal language belonging to the Palaungic branch of the Austroasiatic languages. Based on new data, this paper presents a phonological description of a variety of Riang Lang, spoken in Namsang township of Shan State in Myanmar. Results differ somewhat from earlier analyses of Riang varieties by Shintani (2014) and Sidwell (2015), showing 12 vowels instead of 11 and 21 consonants, including one not previously documented. Lang maintains a relatively large inventory of reduced syllables, including open and sonorant or stop-final reduced syllables. Two contrastive tones reflect the proto-Austroasiatic initial consonant voice contrast.

Keywords: phonology, phonetics, Palaungic, tone

ISO 639-3 codes: ril, yin

1 Introduction
Riang is an Austroasiatic language spoken in Myanmar. Sidwell (2015) classifies it as Palaungic, Palaung-Riang, Riang. Riang Lang (ril) is one of two major varieties of Riang, the other one being Riang Lai, also called Sak (yin); the two are very close. In Burmese, Riang Lang is called Yinnet, and Riang Lai, Yinchia; *-net* meaning black and *-chia* meaning striped. The Riang themselves do not use the terms Lang, Lai or Sak, usually calling themselves only Riang (*rə jaːŋ*). However, if they want to differentiate, their own terms for themselves are *bàːn rói* (Riang Lai) and *trám* (Riang Lang).

Riang was first mentioned in wordlists by Scott (1900), used by Schmidt in his work on Khasi (1904). Data on two Riang varieties, Lang (Black Riang) and Lai (White striped Riang) was provided by Luce (1965, with additional data published by Shorto 2013), and used by Mitani (1979) in his comparative work. Riang Lang and Sak are very similar (Sidwell 2015) and Riang Lang speakers in this study reported being able to understand speakers of Riang Sak. Sidwell (2015) presents a phonology for Riang Sak based on Luce's data in his reconstruction of proto-Palaungic, which finds 21 consonants, with eight possible onset clusters, and 14 codas, as seen in the following tables.

Table 1: *Riang Sak onset consonants (Sidwell 2015)*

p	t~ʈ	ts	k	ʔ
pʰ	tʰ		kʰ	
b~ɓ	d~ɗ		g	
m	n	ɲ	ŋ	
vw[ʊ]	l, r	j		
	sʰ			h

Table 2: *Riang Sak onset clusters (Sidwell 2015)*

pr	tr	kr
pl		kl
p^hr		k^hr
		k^hl

Table 3: *Riang Sak codas (Sidwell 2015)*

p	t~t̪	ic	k	ʔ
m	n	iɲ	ŋ	
u	r, l	e/I		h

The Luce data (Sidwell 2015) also shows 11 vowels, nine monophthongss and two diphthong (Table 4) and two tones. The present study of Riang Lang finds one more vowel and consonant but not the consonant *g*.

Table 4. *Riang Sak vowels (Sidwell 2015)*

i		u
e (é)	ə	o
ɛ (è)		ɔ
	a	ɑ
iɛ		uɔ~ua

Shintani (2014) presents a phonology of a Riang Lang variety spoken around Löy Lëm, Paanglong. He finds 20 consonants, of which 11 may occur as finals. Unlike the Luce data (Sidwell 2015), he does not find the voiced velar *g*. Shintani (2014) finds 10 monophthong vowels and two tones. He does not discuss vowel length or diphthongs, but his transcription does include VV syllables. Shintani notes that *v* is realized as [v~w], *c* may sometimes be realized as [ts], and *s* may sometimes be realized as [ɛ] before *i*.

Table 5. *Riang consonant initials (Shintani 2014)*

p	t	c	k	ʔ
ɓ	ɗ			
p^h	t^h		k^h	
m	n	ɲ	ŋ	
	s			h
	l			
	r			
v		j		

Table 6. *Riang consonant codas (Shintani 2014)*

p	t	c	k	ʔ
m	n	ɲ	ŋ	
	l			
	r			

Table 7. *Riang vowels (Shintani 2014)*

i			ɯ	u
e		ɤ	o	
ɛ	ə	ɔ		
	a			

Shintani (2014) finds 10 monophthong vowels, noting that *ɯ* occurs only in Shan loans. He does not discuss vowel length or diphthongs, but his transcription shows VV syllables, including *aa*, *ia*, and *ua*. He does not transcribe final *-j*, *-w*, handling them instead as diphthongs or triphthongs ending in *-i* or *-o*. Like Sidwell (2015), Shintani finds two tones.

Tone in Austroasiatic languages is well known to relate to loss of initial consonant voicing distinctions, as well as to loss of finals (cf Haudricort 1954, Matisoff 1973). Various Mon-Khmer languages have developed two-tone systems, as seen in Northern Khmu (Kammu) and Plang (Blang), or register systems, as seen in Lamet or some varieties of Khmu (Svantesson 1989, Suwilai 2001) from a historical initial consonant voicing contrast. In this model, proto-voiceless initials result in high tone, proto-voiced initials in low tone. Thurgood (2007) reframes this model in terms of laryngeal features rather than segments, positing a laryngeal intermediate stage to provide a more convincing phonetic explanation for the phenomenon. Loss of initial consonant voicing contrast cannot account for all tonogenesis in the Palaungic languages, however, as Angkuic languages are known to develop two-tone systems from vowel length contrasts instead, with short vowels resulting in low tone and long vowels in high tone as seen in Hu (Svantesson 1991) and Mok (Hall and Devereux 2018). Some Angkuic languages also show effects of final consonant loss on tonogenesis resulting in more complex tonal systems, as in U (Svantesson 1988) and Muak Sa-aak (Hall 2014).

This phonology is based on a wordlist of 1537 items collected by Johanna Sayk with the help of Myint Myint Phyu, from speakers from the village of Sam Kha in Namsang township, southern Shan State, in February 2015. The analysis is based on the author's transcription of the recordings. The language of elicitation was Shan. The list is based on Luce's wordlist (Shorto 2013). Some further data was elicited by this author for confirmation in 2016 and 2017, with a different speaker from Sam Kha village. A shorter 436 item wordlist collected by Myint Myint Phyu from six other Riang villages, four Lang and two Lai, was available for comparison. This paper presents the consonant inventory of the variety of Riang Lang spoken in Sam Kha village, then vowels, and then word structure. It then examines the suprasegmental system of Riang Lang, including the tonal system in relation to the historical initial voicing contrast and the glottal stop.

2. Consonants

There are 21 consonants in Riang Lang, shown in Table 8; of these, 20 occur as single-consonant onsets. The glottal stop occurs predictably with vowel-initial syllables, and is not transcribed as an onset in this paper.

Table 8. *Riang Lang consonants*

Labial	Alveolar	Alveolo-palatal	Velar	Glottal
p	t	c	k	ʔ
b	d			
p^h	t^h	c^h	k^h	
m	n	ɲ	ŋ	
	s			h
	l			
	r			
w		j		

The voiced stops *b, d* have implosive allophones [ɓ, ɗ] occurring in free variation with [b, d]. As noted by Sidwell (2015), the implosive allophones in this data occur with the high tone. However, in this data, all voiced *b, d* occur almost entirely with the high tone, as in examples 1-8.

(1)	[ɓíl]	*bíl*	‘to loose, be lost’ / ‘to disappear’
(2)	[ɓɔ́ʔ]	*bɔ́ʔ*	‘to carry on back’
(3)	[ɓə́q]	*bə́k*	‘to draw water’
(4)	[kə̆ ɓúj]	*kə̆ búj*	‘bamboo rat, mole’
(5)	[də̆k]	*də́k*	‘to halt, stop (of rain)’
(6)	[kə̆ dɪ̆an]	*kə̆ dían*	‘thigh’
(7)	[kə̆ dɔ̆ʔ]	*kə̆ dɔ́ʔ*	‘nose’
(8)	[kn̩ dó]	*kə̆n dó*	‘to stumble’

There are three labial plosives, *p, b,* p^h, as seen in *pír* ‘winnowing tray’, p^h*ír* ‘bee’ and *pɔ́*^{ʔ} ‘father in law, uncle {younger}’, *bɔ́ʔ* ‘to carry on back’. There are three alveolar plosives *t, d,* t^h: *tùp* ‘gable’, *dúp* ‘to cover’ and t^h*úk* ‘to rub {ointment}’. The alveopalatal plosive *c* is usually affricated; in Sidwell (2015) this appears to be the affricate *ts*. In this data it is realized as [ʦ].

The aspirated alveopalatal c^h does not appear in previous data on Riang. It is uncommon but distinct from *c*, as only two examples, ‘spit’ and ‘rose,’ were found in the wordlist of 1537 items. The minimal pairs in examples (12) and (13), (14) and (15) were elicited later in an orthography workshop with a speaker from the same village.

(9)	*tə̆k* c^h*ú ɲáːŋ*	‘spit’	(12)	*cáːr*	‘locust’
(10)	k^h*ín* c^h*ɔ́ːm*	‘midnight’	(13)	c^h*áːr*	‘recover’
(11)	*dák nán* c^h*ì*	‘rose’	(14)	*cáːn*	‘repent’
			(15)	*kàn* c^h*áːn*	‘crawl’

Although Sidwell’s summary of Riang phonology (2015) does not include the consonant c^h, the dictionary includes example (9) in the comparative lexicon for both Riang Sak and Lang as *tək*^{2} c^hu^1 *ɲaŋ*^{1}, ‘spit’. Example (12) agrees with the comparative lexicon, which reconstructs **caːr* ‘grasshopper’ (Sidwell 2015). Example (13) is unlikely to be a borrowed word as the *-r* coda is not found in Shan or Burmese.

There are two velar plosives *k,* k^h: *ké* ‘to weave cloth’ and k^h*é* ‘to wash {hands, plates}’. Uvular allophones [q, q^h] may occur in environments preceding vowels *a, aː, ɔ* as in *ká* [*qáʔ*], ‘fish’ or k^h*rɔ́* [q^h*rɔ́*] ‘rust’; compare to *kát* [*kɜ́t*] ‘cold’ and *ké* [*kéː*] ‘to weave cloth’. In final position, *-k* may become a fricative as in *plɛ́ʔ ʔàk pɨt* [plɛ́ʔ ʔàχ pɨ́t] ‘pellet’.

Nasals occur at four points of articulation, *m, n, ɲ, ŋ*: *mɨk* ‘cattle’, *nɨk* ‘full’, and *ɲɨk* ‘sticky, glutinous’; *mɔ̀m* ‘sister in law’ and *ŋɔ̀m* ‘to wait for’. There is one voiceless syllabic nasal in the data: *ñ̥*^{ʔ} ‘he/she’. The semivowel *w* may be realized as *ʋ* syllable initially.

Table 9. *Riang Lang consonant codas*

Labial	Alveolar	Alveolo-palatal	Velar	Glottal
p	t	c	k	ʔ
m	n	ɲ	ŋ	
	s			
	l			
	r			
w		j		

There are 14 coda consonants, as shown in Table 9. These include voiceless plosives *-p, -t, -c, -k* as in *káːp* 'chin', *kát* 'cold', *mác* 'sand', *kák* 'to bite'; the glottal stop, as in *sóʔ* 'dog'; nasals *-m, -n, -ɲ, -ŋ* as in *mɛ̀m* 'tea', *kúan* 'child', *prɨ̀ɲ* 'ant', *kàːŋ* 'house'; the fricative *-s* as in *bás* 'carry (hanging from head)'; the liquids *-l, -r* as in *kə́l* 'to play', *lír* 'smooth'; and the semivowels *-j, -w* as in *kə̆ dáw* 'liquor', *màj* 'elder sibling'. There are no aspirated or voiced plosive codas and *-h* does not occur as a coda in this variety of Riang. Plosive codas are unreleased.

The smaller set of data available for six other Riang villages is very similar even between the Riang Lang and Riang Lai villages. The most notable phonological difference was that for one Riang Lang village, the *-s* coda did not occur. It is sometimes dropped but is frequently replaced by *-h*, which does not occur as a coda in the other varieties of Riang.

Consonant clusters:

The liquids and semivowels *l, r, j, w* may occur as the second consonant in a cluster. Possible clusters seen in the data include those in Table 10.

Table 10. *Consonant clusters*

pr	pl		
pʰr			
tr			
sr			
kr	kl	kj	kw
kʰr	kʰl	kʰj	

Both the cluster *sr-* and a reduced syllable *sə̆* followed by the main syllable onset *r* may occur, as seen in the pair *sráːŋ* 'sweat' (n) and *sə̆ráːŋ* 'amount of land a buffalo can plow in one morning or one afternoon'.

3. Vowels

Riang Lang has 10 monophthong vowels, as shown in Table 11, and two diphthongs. This analysis differs from Luce (1965) and Sidwell (2015) primarily in finding one additional monophthong vowel, *ɨ*. This is consistent with Shintani's inventory.

Table 11. *Riang Lang vowels*

	Front	Central	Back
Close	i	ɨ	u
Close-mid	e	ə	o
Open-mid	ɛ	a	ɔ
Open		a:	
Diphthongs	ia		ua

The vowels *a, a:* have slightly different vowel qualities; phonetically they are [ɜ, a:]. They contrast in length as well as vowel quality. In this analysis it will be considered primarily a length distinction, as the contrast is neutralized in open syllables. A minimal set for the ten monophthong vowels is given in Table 12.

Table 12. *Monophthong vowel examples*

kít 'sink (v.)'	*kɨ́t* 'remainder'	*túk* 'tie'
két 'fish scales'	*kə́t* 'sunset'	*kók* 'neck'
két 'run over'	*kát* 'cold'	*kɔ́k* 'handle'
	tá:k 'tongue'	

The length distinction between *a* and *a:* is not large, as may be seen in Table 6. In these examples, *a* before glottal stop coda was longer than *a* followed by other stop codas. In addition, the open syllable *a:* was more than double the length of *a:* with a coda.

Table 13. *Length distinction between /a/ and /a:/*

item	gloss	token 1	token 2	average (seconds)
kák	'to bite'	.097	.098	.098
kà:k	'hang from shoulder'	.153	.121	.137
ká:p	'lower jaw'	.148	.119	.134
káp	'to wear [clothes]'	.087	.077	.082
ka:	'dance'	.361	.270	.316
kaʔ	'fish'	.128	.111	.120

Other monophthong vowels do not show a length contrast. They are predictably lengthened in open syllables and short in closed syllables. The two diphthongs, *ia* and *ua*, like the long vowel *a:*, do not occur with the glottal stop coda. The diphthong *ia* may sometimes be realized as [ea], as in *vìam* [vèam] 'to grind teeth'.

4. Word structure:

Riang Lang is sesquisyllabic; words include a main syllable and an optional reduced syllable preceding the main syllable. Main syllables occur with the full inventory of phonemes discussed above. The reduced syllable has a limited inventory of consonants, and vowel and tone are neutralized. Although the minor syllable vowel pronunciation may vary, it is not contrastive and will be transcribed here as *ə̆*. Overall word-structure may be summarized as follows:

$$((C)\breve{ə}(C)).(C)(C)V(C)^{T}$$

Examples (16-26) show possible Riang Lang word structures.

(16)	V	*óʔ*	'I' [1S]
(17)	VC	*úp*	'speak'
(18)	CV	*sáː*	'sell'
(19)	CVC	*rùp*	'fishing net'
(20)	CV.VC	*rə̆ áːŋ̰*	'stone'
(21)	CV.CVC	*mə̆ rə̀ŋ*	'horse'
(22)	CV.CCVC	*kə̆ trép*	'flat, level'
(23)	CVC.CV	*sə̆k ŋíʔ*	'day'
(24)	CVC.CCV	*tə̆k klɛ́*	'to cause to fall'
(25)	CVC.CVC	*kə̆n mɨ̀r*	'pregnant'
(26)	CVC.CCVC	*tə̆r plə̀ŋ*	'to be different'

Reduced syllables may include an onset, a neutralized vowel [ə] and tone, and optionally a consonantal coda, which may be a sonorant or stop. There is therefore a relatively large inventory of reduced syllables in comparison to some other Palaungic languages. Onsets are restricted to *p, t, c, k, m, r, s* and codas to *k, m, n, ŋ, l, r*. Continuant onsets or codas may become syllabic, so that there is no phonetic vowel. Reduced syllable codas need not match the following syllable onset in place of articulation; see examples (27-41). In examples 29, 35, 38 and 40, the reduced syllable coda agrees with the following onset, but in 28, 30, 33, 36 and 39, they are at different points of articulation.

(27)	*sə̆ kɔ́l*	'ten'	(32)	*cə̆n áːŋ̰*	'bone'	(37)	*pə̆ náʔ*	'water buffalo'
(28)	*sə̆m tɔ́r*	'cock's comb'	(33)	*cə̆k nɛ̀ŋ*	'lean on'	(38)	*pə̆n lùː*	'cemetery/ grave'
(29)	*sə̆k ŋíʔ*	'day'	(34)	*kə̆ bú*	'rat'	(39)	*pə̆k lìat*	'snail'
(30)	*sə̆k túʔ*	'wash clothes'	(35)	*kə̆l dɨ́c*	'to limp'	(40)	*tə̆r làːk*	'bat'
(31)	*rə̆ áːŋ̰*	'stone'	(36)	*kə̆n mɨ̀r*	'pregnant'	(41)	*mə̆ rə̀ŋ*	'horse'

6. Suprasegmentals:

6.1 Tone

There are two tones in Riang Lang, high and low. As expected, these two tones appear to correlate with the historical initial voicing distinction which has been lost in Riang.

Table 14 shows that Riang Lang tones line up with the tones in Northern Khmu, where the historical initial voicing distinction has been replaced by a tone contrast, and with the voicing distinction in Southern Khmu, which retains it. Further, the glottal fricative *h* occurs almost exclusively with high tone, as in examples 54-56. This is consistent with high tone developing from voiceless initials.

Table 14. Riang Lang tone in comparison to Northern Kammu, and Southern Kammu [Khmu] (adapted from Svantesson 1991)

Northern Kammu	Southern Kammu	Riang Lang	Gloss
**voiceless*			
píŋ	*piŋ*	*púŋ*	'to shoot'
táaŋ	*taaŋ*	*tá:ŋ*	'to weave'
káap	*kaap*	*ká:p*	'jaw''
**voiced*			
--	--	*kǎn rà:*	'white'
prì²	*bri²*	*rùk*	'forest'
ktáak	*kdaak*	*plák tí?*	'palm (of hand)'
kàaŋ	*gaaŋ*	*kà:ŋ*	'house'
--	--	*kák*	'to bite'

Of 39 *h* onsets in the data, only one (*hà:n cʰím*, bird's nest) is low tone. This could reflect borrowing from Shan *háŋ*, but it is not inconsistent with **su:m* which is reconstructed for proto Palaungic, with **s* having merged with **h* (Sidwell 2015), and borrowing from Shan would also not explain the resulting low tone.

(54) *húk* 'body hair'
(55) *hír* 'iron'
(56) *hé* 'to lean sideways'

6.2 Glottal stop

The glottal stop in Riang Lang does not behave like the other consonants. Syllable-initially, the glottal stop occurs predictably in Riang Lang vowel-initial syllables. It also occurs syllable finally in open syllables with phonetically short vowels. All vowels are phonetically long in open syllables and shorter in glottal-stop final syllables, as shown in Table 15.

Table 15. *Vowel length comparisons, glottal stop final and open syllables*

item	gloss	token 1	token 2	average (seconds)
pá:	'father'	.354	.366	.360
pá?	'blanket or sheet'	.139	.125	.132
là:	'petticoat, skirt'	.382	.330	.356
lá?	'leaf'	.113	.110	.112
sá:	'to sell'	.376	.243	.310
sʰá?	'body louse'	.138	.121	.130
kɔ̀:	'to bark'	.351	.316	.334
kɔ̀?	'brother-in-law'	.159	.149	.154
lù:	'hole'	.389	.363	.376
lú?	'thread'	.139	.104	.123

The diphthongs *ia* and *ua*, which are always long, cannot be followed by the glottal stop, which further supports this. In addition, the contrast between the short and long vowels *a* and *a:* is neutralized in open or glottal stop coda syllables. In open syllables the vowel quality is usually realized as [a:]; in glottal stop coda

syllables it is usually [ɜ]. However, it may also be realized as [a] in glottal-stop coda syllables, and there is no contrast between [ɜ] and [a] in open or glottal stop coda syllables. The glottal stop might therefore be considered to be a marker of shortness rather than a coda.

Table 16. *Distribution of possible nuclei according to syllable finals*

Syllable type	Possible nuclei
open	phonetically long vowels *a:* *ia, ua*
-ʔ	phonetically short vowels *a* no diphthongs
other finals	*a, a:* other vowels phonetically short *ia, ua*

Table 16 shows the difference in distribution for syllable types and vowel nuclei. Unlike Angkuic languages such as U or Muak Sa-aak for which final consonants have been a factor in tonogenesis, tone is not part of the distributional differences. Open syllables, glottal stop coda syllables, and syllables with other codas may all occur with either high or low tone in Riang Lang. If glottal stop coda syllables and open syllables are combined in Table 16, the combined group would have the full range of nuclei possibilities seen for closed syllables with non-glottal stop codas.

7. Conclusions:

The current study of Riang Lang largely agrees with Sidwell (2015) and Shintani (2014), finding 21 initial consonants and 12 vowels. Like earlier studies, this one also finds two tones.

This analysis finds Riang Lang to be similar to the variety studied by Luce, with a few exceptions. Unlike the Luce data (Scott, 1900 and Luce, 1965; cited in Sidwell, 2015), the analysis reveals 12 vowels instead of 11, including 10 monophthong vowels and two diphthongs. It posits four instead of two central vowels and only three instead of four back vowels. This is consistent with Shintani's finding of 10 monophthong vowels. Length contrast in this study is found to be phonemic only in closed syllables. There are 20 consonants, including one not previously documented. In contrast to Shintani's data, this variety of Riang Lang has one more initial consonant (c^h) and one more final (-s). Further study of other varieties of Riang Lang, and of Riang Lai, would be useful to clarify the extent of internal variation among Riang varieties.

References

Hall, Elizabeth. 2014. An analysis of Muak Sa-aak tone. *Journal of the Southeast Asian Linguistics Society* 7:1–10.

Hall, Elizabeth and Shane Devereux. 2018. Preliminary Mok Phonology and Implications for Angkuic Sound Change. Paper presented at the 28th meeting of the Southeast Asian Linguistics Society, Kaohsiung, Taiwan, May 24-26, 2018.

Haudricourt, André-George. 1954. De l'origine des tons en vietnamien. *Journal Asiatique* 242:69–82.

Luce, Gordon H. 1965. Danaw, a dying Austroasiatic language. *Lingua* 14:98–129.

Matisoff, James A. 1973. Tonogenesis in Southeast Asia. In *Consonant type and tone*, ed. by Larry M. Hyman. 71–95. Los Angeles: Linguistics Program, University of Southern California.

Mitani, Yasuyuki. 1979. Vowel correspondences between Riang and Palaung. In *Studies in Thai and Mon-Khmer Phonetics and Phonology in Honour of Eugénie J.A. Henderson*, ed. by Theraphan L. Thongkum, Vichin Panupong, Pranee Kullavanijaya, and M. R. Kalya Tingsabadh. 142–50. Bangkok: Chulalongkorn University Press.

Schmidt, P. Wilhelm. 1904. Grundzuge einer Lautlehre der Khasi-Sprache inihren Beziehungen zu derjenigen der Mon-Khmer-Sprachen. Mit einem Anhang: die Palaung-, Wa-, und Riang-Sprachen des mittleren Salwin. *Abhandlungen der bayerischen Akademie der Wissenschaft* 22(3):677–810.

Scott, James George and John Percy Hardiman. 1900. *Gazetteer of Upper Burma and the Shan State*. Rangoon: Superintendent, Government Printing.

Shintani, Tadahiko L. A. 2014. *The Riang Language. Linguistic Survey of the Tay Cultural Area*, No. 101. Tokyo: Research Institute for Languages and Cultures of Asia and Africa.

Shorto, Harry. L. 2013. *Palaung Word List: Based on Materials Collected from Pan Shwe Kya, Namhsan, Sept-Oct, 1957*. Asia-Pacific Linguistics Open Access Monographs, A-PL 004; SEAsian Mainland Languages E-Series (SEAMLES), 003. Canberra, ACT: Asia-Pacific Linguistics, available at http://hdl.handle.net/1885/9782.

Shorto, Harry. L. 2013a. *Riang-Lang Vocabulary: Compiled from the Materials Collected by G. H. Luce*. Asia-Pacific Linguistics Open Access Monographs, A-PL 005; SEAsian Mainland Languages E-Series (SEAMLES), 004. Canberra, ACT: Asia-Pacific Linguistics, available at http://hdl.handle.net/1885/9781.

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

Suwilai Premsrirat. 2001. Tonogenesis in Khmu dialects of SEA. *Mon-Khmer Studies* 31:47–56.

Svantesson, Jan-Olof. 1988. U. *Linguistics of the Tibeto-Burman Area*, 11(1): 64–133.

Svantesson, Jan-Olof. 1989. Tonogenetic Mechanisms in Northern Mon-Khmer. *Phonetica* 46:60-79.

Svantesson, Jan-Olof. 1991. Hu - a language with unorthodox tonogenesis, in Austroasiatic *Languages, Essays in honour of H. L. Shorto*, ed. J.H.C.S. Davidson, 67–80. London: School of Oriental and African Studies, University of London.

Thurgood, Graham 2007. Tonogenesis revisited: Revising the model and the analysis. In *Studies in Tai and Southeast Asian Linguistics*, ed. by Jimmy G. Harris, Somsonge Burusphat, and James E. Harris 263–291. Bangkok: Ek Phim Thai Co.

VERBAL AFFIXES IN RUMAI, PALAUNG

Rachel Weymuth
University of Zurich
rachel.weymuth@uzh.ch

Abstract
Rumai, a variety of Palaung, a Palaungic language of the Austroasiatic language family, has several verbal affixes. Some of them can be traced back to lexical morphemes, so that the grammaticalization path is obvious, for others there is no such trace. The aim of this paper is to describe the functions and where possible the origin of the Rumai verbal affixes. Additionally, the connections of the affixes with secondary verbs with similar meanings will be discussed.

Keywords: Austroasiatic, Palaung, Rumai, verb, affix

ISO 639-3 codes: rbb, pll, pce, khm, shn, mya

1 Introduction
Rumai (rbb) is a variety of Palaung, an Austroasiatic language of the Palaungic branch. The other two varieties, so far mentioned in the linguistic literature, are Shwe (pll) and Ruching (pce). Rumai is mainly spoken in Northern Shan State, Myanmar, but also in the adjacent province Yunnan, China. There are some 140,000 speakers, who traditionally live on the slopes and ridges of mountains, but today many of the Palaung, especially young people, live in towns and cities in Shan State valleys or in the plain of Myanmar for studies or for work. The main contact languages are Shan and Burmese.
The prevalent syntactic structure of Rumai is verb-medial, but most of the dependent clauses are verb-initial, such that the arguments follow the verb. Full or partial grammatical descriptions of Rumai are so far not available. Therefore, all the examples are from the corpus of the present author, collected on several field trips to Myanmar. It includes elicited sentences as well as written texts and recordings of conversations, interviews and picture stories. sIn this paper, the verbal affixes will be described and discussed, including secondary verbs[1] with similar functions. The paper is work in progress, and many questions are still open to discussion. A list of example sources and consultants is provided following the conclusion.[2]

1.1 Affixes in Rumai
Traditionally, linguistic studies and descriptions of Southeast Asian languages classify them as morphologically non-complex, that means, each morpheme, be it lexical or grammatical, is an independent entity. In Rumai however, some morphemes are not in free occurrence, but they are verbal or nominal affixes. These morphemes are never stressed and some of them show phonetic features, which don't occur with independent morphemes. One of them is the durative prefix *ʔuN-*, of which the final nasal adapts to the place of articulation of the following consonant like in *ʔun-sə̂w* (DUR-hurt).

---

[1] Secondary verbs are verbs occurring in a clause in addition to the main verb. They are verbs like "can", which are often classified as auxiliaries in other languages. Such verbs precede the main verb in Rumai. Additionally, there are verbs which can occur as main verbs, but take a grammatical function when following the main verb.

[2] Abbreviations used are: 1SG first person singular; 1DU.EXCL first person dual exclusive; 1PL.INCL first person plural inclusive; 1PL.EXCL first person plural exclusive; 2SG second person singular; 3SG third person singular; 3DU third person dual; 3PL third person plural; ; ANA anaphoric demonstrative; ASRT assertive; CLF classifier; COMP complementizer; COMPAR comparative; COND conditional; DEM demonstrative; DES desiderative; DISC discourse marker; DIST distal demonstrative; DUR durative; EMPH emphasis; EXP experiential; INCEP inceptive; IPFV imperfective; IRR irrealis; LOC locative; MEDL medial demonstrative; NEG negative; NSIT new situation; OBL1 oblique 1; OBL2 oblique 2; PN proper name; POL polite; PROH prohibitive; PROX proximal demonstrative; Q question marker; RECP reciprocal; RESTR restrictive; TAG tag question; TCL topic-comment linker.

Each of these affixes is exclusively either adjacent to a verb stem or to a noun stem or to another verbal or nominal affix and except for one, namely the negative suffix *-maʔ*, the affixes, as is common in V-O languages (Payne 1997:72), are prefixes. In Rumai, there are a small number of lexemes that may be used as verbs or as nouns, that means, they can occur in verbal as well as in nominal function and each of them can take the corresponding affixes. Among these lexemes are for example *gɔ̂ːj* 'stay', *kəmɛh* 'love' and *dâː* 'use; usage'. In this paper, these lexemes are simply called verbs or nouns, according to their respective function in a clause.

2 The verbal affixes

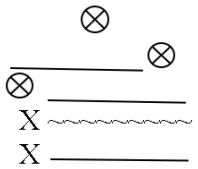
The verbal affixes are summarized in the following table:

Table 1: *Verbal affixes in Rumai*

Affix(es)	Meaning	Source	Domain
giːj-	imperfective	*gɔ̂ːj* 'stay'	aspectual
ʔuN-	durative	*ʔûːn* 'keep'	
ʔə-	inceptive	-	
hɔ̆j-	new situation	*hɔ̀ːj* 'finish'	
tʌm-	experiential	-	
nʌŋ-	irrealis	-	modality
siŋ-	desiderative	-	
buː- / ʔaːw- / -maʔ	negative	-	negation
ɲjʌm-	'not yet'	('dilatory, stiff')	
kʰuː-	prohibitive	-	
kə- / laj-	reciprocal	- / *lâːj* 'take'	reciprocal

The first five affixes: imperfective *giːj-*, durative *ʔuN-*, inceptive *-ʔə*, new situation *-hɔ̆j* and experiential *tʌm-* (sections 2.1 to 2.5) are aspectual markers that will be categorized according to the following table that was compiled by Roos (2001) according to Johanson (2000:33):

Table 2: *Aspectual categories*

Category	The actional content is conceptualized	
[+tf] finitransformative	1. as implying final transformation…	
[+tf, +mom]	…without a salient cursus	
[+tf, −mom]	…with a salient cursus	
[+ti] initiotransformative	2. as implying initial transformation	
[+t, +dyn]	3. without transformation, as dynamic	
[+t, −dyn]	4. without transformation, as static	

(Roos 2001:49)

Sections 2.6 and 2.7 cover the modality prefixes *nʌŋ-*, irrealis, and *siŋ-*, desiderative, followed by the negative affixes in 2.8. Finally, the reciprocal prefixes *kə-* and *laj-* are discussed in section 2.9.

2.1 Imperfective gi:j-

The imperfective marker *giːj-* denotes an action or state as going on or holding for an indeterminate time span in the present (1), past (2) or future (3). There is no transformation, and the situation may be dynamic as well as static, what includes the last two categories in table 2 and results in [+t, ±dyn]. Smith (1991:111) describes the imperfective as follows:

> Imperfective viewpoints present part of a situation, with no information about its endpoints. Thus imperfectives are open informationally. The unmarked imperfective spans an interval that is internal to the situation; this conforms to the principle that unmarked viewpoints have a span that coincides with all or part of the temporal schema of the situation.

(1) *caŋnaj ʔʌ̂w **giːj**-ɻɤh tʒ̂w.hlâː*
now 1SG IPFV-buy vegetable
'Now I am buying vegetables.' (ENEG_13_M_1/2_054)

In example (2), the temporal adverbial does not define any boundaries of the action, but the addressee had already started work prior to the preceding year.

(2) *sənâm ʔàːj ɻûŋ siː mâj **giːj**-rɛ̂n*
year before work what 2SG IPFV-do
'Last year, what work were you doing?' (I_17_MS_3_028)

(3) *nʌŋ-**giːj**-bâːj lôj hŋjɛ̂n hnaːŋ ʔʒː gʌ̀ː jʌ̂ːm*
IRR-IPFV-happen only DEM how.many 1PL.INCL old die
'It only will be like this, until we are old, and we die.' (C3_17_MS_3_155)

The imperfective marker, which among other meanings expresses the "continuous" function, is probably derived from the verb *gɔ̂ːy* 'stay'. Heine and Kuteva (2002) list 'stand' as one of the sources for the category "continuous" (Heine and Kuteva 2002:330). Note, that the Rumai verb *gɔ̂ːj* 'stay' also means 'stand'. This verb is not used as a secondary verb, but there is *ʔûːn* 'keep', another verb that is mentioned by Heine and Kuteva (2002:330) as a source for a continuous marker. *ʔûːn* as a post-verbal secondary verb has a similar function like *giːj-*. *ʔûːn* is probably the origin of the durative prefix *ʔɯN-* and therefore, its use as a secondary verb will be presented together with this marker in the following section.

2.2 Durative ʔɯN-

A situation of which neither the initial nor the final boundary is known or of importance can be marked by the durative prefix *ʔɯN-*. Therefore, the function of this prefix is similar to that of the imperfective marker *giːj-*, but *ʔɯN-*, contrary to *giːj-*, is never directly used with activity verbs and its category is therefore [−t, −dyn]. The prefix is derived from the verb *ʔûːn* 'keep' and the final nasal adapts to the point of articulation of the following consonant.

(4) *kjɔh mʌː **ʔɯm**-buː-bɤ̂n-maʔ lʌː kjɔh*
language RESTR DUR-NEG-get-NEG COMP speak
'He also could not speak.' (WA_15_M_3_046)

The following example is similar to (2), as also introduced by a temporal adverb that encompasses a definite timeframe. Here, the woman had pains in different parts of her body already for some weeks, and presumed that it would not disappear the same day. Therefore, the durative marker focusses on her ongoing pains.

(5) *ʔɯndih*[3] *tʌŋ jaʔ ʔɯnhnîː dih **ʔɯn**-sʒ̂w*
today LOC shoulder this DISC DUR-hurt
'Today, in this shoulder it hurts.' (I_17_MS_7_056)

The prefix *ʔɯN-* is often used to mark generic situations:

(6) *pəkjɤ̂ː nîː **ʔɯŋ**-kəpʌ̀ː kətâːj-kʌmpʰaː nîː*
moon PROX DUR-circulate ground-world PROX
'This moon goes around this earth.' (WA_15_M_3_059)

[3] There are several lexemes, which seem to have a prefix *ʔɯN-*, but where the meaning of the two parts is not transparent anymore.

As already mentioned in section 2.1, *ʔûːn* 'keep', used as a secondary verb, has a similar function like the progressive prefix *giːj-*. It describes a situation or action with unknown or unimportant boundaries [±t, ±dyn]:

(7) *sohoː* *mʌː* *ʔɯmbjâːj-**ʔûːn*** *tôː* *ʔân*
mosquito RESTR prepare-keep body 3SG
'The mosquito was getting ready.' (WA_15_M_3_023)

(8) *kʰʌː* *ʔəsaːk* *ʔʌ̂w* *dâːŋ* *mɛh-**ʔûːn*** *ʔîːːpân* *nʌŋ-tɤh* *ʔʌ̂w*
COND life 1SG big exist-keep woman IRR-look 1SG
'When I'm old, there is a woman who will look after me.' (WA_15_M_3_077)

The prefixes *ʔɯN-* and *giːj-* can occur together (9) and so can each of the two prefixes with the secondary verb *ʔûːn* (10; 11):

(9) ***ʔɯŋ-giːj-**moh* *lôj* *baj* *ɻɛn*
DUR-IPFV-be.so only how alike
'It is always like this.' (I_17_MS_7_029)

(10) *kʰʌː* *bâːj* *ɻɛn* ***ʔɯm-**bɤ̂n-**ʔûːn*** *vîː* *hôm* *pʌ-sɜŋ* *dîː*
COND happen alike DUR-get-keep return eat OBL2-shop MEDL
'In this case, we can go again to eat in that shop.' (EQB_17_M_3_063)

(11) *bɤː* *hôm* *kɜ̂ː* *tʌ-pôm* *sʌw-kətâːj* *dîː* ***giːj-**bjâː-**ʔûːn*** *lʌ*
When eat 3PL OBL1-rice dog-ground MEDL IPFV-steal-keep COMP
tɤh *kɜ̂ː*
look 3PL
When they were eating, the fox kept on watching them stealthily. (PCHT_16_L_6_010)

The common feature of the two prefixes *giːj-* and *ʔɯN-* and the secondary verb *-ʔûːn* is their lack of a transformation and therefore their affiliation to the imperfective domain. The two prefixes cover to a certain extent different parts of the domain, as *giːj-* is used with stative as well as with activity verbs and therefore can express a progressive meaning (1) and *ʔɯN-*, only occurring with stative verbs, is often used for generic situations (6). However, there may be an overlap of their functions in examples (2) and (5). To what extent the imperfective prefix *giːj-* and the secondary verb *-ʔûːn* differ in their function is not yet clear and neither is the interplay between all the three morphemes.

2.3 Inceptive ʔə-

The inceptive prefix *ʔə-* marks a situation that has been established after a change of state [+ti]. This category is described by Smith (1991) as follows:

> Sentences with an inceptive focus may in effect present an Activity indirectly. The inceptive focusses on the beginning of the event. With no information to the contrary, the receiver could reasonably infer that the Activity continues. (Smith 1991:48)

(12) *kɜ̂ː* *dʌh* *sʌw-kətâːj* *dîː* ***ʔə-**jok* *hâːw* *kɔ̂ːn* *ʔjɤ̂ː* *dɤ̂ː*
3PL say dog-ground MEDL INCEP-lift go child chicken ANA
'They said: "That fox took away that chick."' (PCHT_16_L_6_014)

(13) *ʔân* ***ʔə-**jɔh* *lɛp* *kʰɛ̂ŋ* *ʔɛ̂m-kʰôm* *hlɛ̂m*
3SG INCEP-fall enter LOC water-reservoir deep
'He just fell into a deep drinking water reservoir.' (W4_15_M_3_003)

Often, *ʔə-* is used to mark a sequence of actions and/or states:

(14) *ʔân* ***ʔə-****ʔɔ̂ŋ* ***ʔə-****hlə̂ːn* *ʔân* ***ʔə-****mɔj* *ʔân* ***ʔə-****hlə̂w* *lʌ* *dâː*
3SG INCEP-shout INCEP-long 3SG INCEP-tired 3SG INCEP-rest COMP use
ʔuː *taj* *dɜʔ*
one moment small
'He shouted for a long time, he got tired and then he took a rest for a moment.' (W4_15_M_3_009)

The inceptive prefix and the new situation marker *hɤ̆j-* have very similar functions, their interplay will be discussed in the following section.

2.4 New situation **hɤ̆j-**

A marker that is derived from a verb with the meaning 'finish' and that denotes a new situation, that is a situation having "been established after a change of state", is very common in the languages of Southeast Asia (Jenny et al. 2015:97–98). Jenny, in an earlier publication, describes the marker as a "new (but expected) situation after a limit has been transgressed" (Jenny 2001:125). In Rumai, the marker with this function is *hɤ̆j-*, derived from the verb *hɔ̀ːj* 'finish'. Like the inceptive marker *ʔə-*, it shows the feature [+ti].

(15) *pʌŋtâːj* *mʌː* ***hɤ̆j-****jə̀ːj-vɤ̌h* *mâj.ploh*
rabbit RESTR NSIT-rise-open window
'The rabbit has opened the window.' (PCHT_16_L_6_007)

(16) *ʔîː* ***hɤ̆j-****bɤ̌n* *lʌː* *hok* *tìː* *pʌ-nʌm.ʔom*
others NSIT-get COMP climb plant OBL2-PN
'Others have got to plant up in Nam Om.' (C3_17_MS_9_024)

As mentioned in section 2.3, the inceptive and the new situation markers have very similar meanings. This is shown in the following examples which are question and answer in an interview:

(17) *hnaːŋ* *sənâːm* ***ʔə-****bɤ̌n* *pəcaːŋ* *cʌ̂ːm-dɛh* *pʌ-nîː*
how.many year INCEP-get monk reach-come OBL2-PROX
'How many years ago did you come here?' (I_15_M_1_007)

(18) *ʔʌ̂w* *cʌ̂ːm-dɛh* *pʌ-nîː* ***hɤ̆j-****bɤ̌n* *pʰʌːn-sənâːm*
1SG reach-come OBL2-PROX NSIT-get five-year
'I came here five years ago.' (I_15_M_3_008)

The prefixes are, according to one consultant, often interchangeable, as is shown in examples (12) and (19).

(19) *kə̂ː* *dʌh* *sʌw-kətâːj* *dîː* ***hɤ̆j-****jok* *hâːw* *kɔ̂ːn* *ʔjɤ̂ː* *dɤ̂ː*
3PL say dog-ground MEDL NSIT-lift go child chicken ANA
'They said: "That fox has taken away that chick."' (pc_3)

The prefix *ʔə-*, in contrast to the prefix *hɤ̆j-*, cannot be traced back to any lexical source. Therefore, the inceptive marker is probably an old feature of Rumai, while the new situation marker may be a more recent development due to areal influence. The functions of the two markers overlap to a large extent; a major distinction is that the inceptive marker can also express sequential actions and states.

The verb *hɔ̀ːj* can be used as a secondary verb (20) and in addition, the NSIT-marker may be prefixed to the main verb (21):

(20) *ʔʌ̂w* *hmɔ̀ː-****hɔ̀ːj****,* *ʔân* *mʌː* *buː-mɛh-maʔ* *nəriː*
1SG ask-finish 3SG RESTR NEG-exist-NEG hour
'I have asked, he also doesn't have a watch.' (EQB_17_M_3_047)

(21) *ʔân* ***hɔ̆j-****mɛh-****hɔ̀ːj*** *dûːn*
3SG NSIT-exist-finish sweetheart
'He already has a girlfriend.' (WPKT_14_L_391)

The prefix and the secondary verb in (21) seem to have similar meanings and one may reinforce the other. This similarity of meaning is also supported by the following examples that are elicited from the same sentence but uttered by different consultants. Thus, the new situation marker and the secondary verb are at least sometimes interchangeable.

(22) *ʔân* *hôm-****hɔ̀ːj*** *doʔ*
3SG eat-finish exhausted
'He has eaten it all.' (EEN_16_M_3_022)

(23) *ʔân* ***hɔ̆j-****hôm* *doʔ*
3SG NSIT-eat exhausted
'He has eaten it all.' (EEN_16_M_1_023)

Generally, regarding the aspectual prefixes and their lexical sources which are partly also used as post-verbal secondary verbs, there is not only the question about the differences between the meanings of the prefixes and the secondary verbs, but also about the relationship between them, their order of appearance and the areal influence. Moreover, the age and the origin (dialect) of the consultants may play a role in the use of the aspect markers and the secondary verbs.

2.5 Experiential **tʌm-**

The aspectual marker *tʌm-* expresses perfectivity [+tf, +mom], marking an event that is looked at in its entirety without a salient cursus. However, it also connotes having done or experienced something at least one time. Due to its perfective meaning, it is included here in the aspectual domain.

(24) Q: *pəcaːŋ* ***tʌm-****hlɜː* *lʌh* *diː* *hâj*
monk EXP-visit Q place other
'Have you ever visited another place?' (I_15_M_1_041)

A: ***tʌm-****hlɜː* *ʔɤː*
EXP-visit ASRT
'Yes, I have.' (I_15_M_3_043)

As *tʌm-* is an affix, it is always bound to a verb, and therefore, a simple *tʌm* is not possible in the answer.

2.6 Irrealis **nʌŋ-**

The irrealis category is not as easily identifiable as other categories like perfective or progressive (Bybee 1998:264). Bybee explains it as follows:

> For any given language, there are several grams that mark off portions of the conceptual space for situations that are not asserted to exist, or if there is a highly generalized gram, it does not cover all "irrealis" situations and furthermore does not actually have one invariant meaning, but rather takes its meaning from the construction in which it occurs. (Bybee 1998:264–65)

This is the case for the irrealis *nʌŋ-*, which may have a connection with the Khmer morpheme *nɯŋ* 'future' (Haiman 2011:263). The prefix isn't used, for example, to express negative or imperative meanings, which don't assert a situation to exist, but depending on the situation, it expresses future (25), possibility (26), intention (27) or supposition (28):

(25) *ʔʌ̂w* ***nʌŋ-****tɔ̀ː* *mâj* *hâːw* *pənʌ̂w* *pəŋaː*
1SG IRR-follow 2SG go learn education
'I will follow you to learn.' (W1_030)

(26) *kɔ̂ːn* *mâj* ***nʌŋ-****pənʌ̂w* *bèː* *pʌ-cʰôŋ* *jêː*
child 2SG IRR-learn text OBL2-school 1PL.EXCL
'Your son may study at our school.' (EGGD_16_M_3_134)

(27) *maː* *ʔân* *ʔə-nɅ̀ː* ***nʌŋ-****hâːw* *ʔân* *pʌ-mantəlêː*
mother 3SG INCEP-know IRR-go 3SG OBL2-Mandalay
'His mother found out about his plans to go to Mandalay.' (EEN_16_M_1_062)

(28) ***nʌŋ-****gɔ̂ːj-ʔûːn* *pûː* *tâː* *pəjʌː*
IRR-stay-keep seven eight hundred
'There may be seven or eight hundred [houses].' (C3_17_MS_9_020)

The irrealis marker, at least in its future meaning, is not obligatory and it is, especially in negative clauses, often omitted. The following sentence is elicited, but in an appropriate context, it could also be translated as 'I did not go with my brother.' Whether or to what extent the irrealis marker can be left out in its other meanings is still to be investigated.

(29) *ʔʌ̂w* *tɔ̀ː-hâːw-maʔ* *piː* *ʔʌ̂w*
1SG follow-go-NEG elder.sibling 1SG
'I will not go with my brother.' (ENEG_13_M_1/2_181)

Although *nʌŋ-* is often omitted in negative clauses, it precedes sometimes the negative prefixes *buː-* (30) and *ʔaːw-* (31):

(30) *maː* *ʔʌ̂w* ***nʌŋ-buː-****pʰlâːn-maʔ*
mother 1SG IRR-NEG-poor-NEG
'My mother will not be poor.' (WA_15_M_3_038)

(31) *ʔʌ̂w* *ʔaːw-hŋjɛm-maʔ* ***nʌŋ-ʔaːw-****moh-maʔ* *dih*
1SG NEG-believe-NEG IRR-NEG-be.so-NEG DISC
'I don't believe it, it is impossible.' (EQB_17_M_3_034)

Other prefixes so far found which are preceded by *nʌŋ-* are *giːj-* in example (3) and *hɔ̆j-* in (32):

(32) *ʔɛ̂* ***nʌŋ-hɔ̆j-****loː-vîː*
1PL.INCL IRR-NSIT-need-return
'We should go back home.' (WPKT_14_L_399)

Some of the personal pronouns merge with the irrealis marker like in the following example, where *ʔʌŋ-* is a merger of *ʔʌ̂w* (1SG) and *nʌŋ-*:

(33) ***ʔʌŋ-****kəɟɔj* *jâː* *tɔ̂j* *ʔundʌ̂ːn* *hmɛː* *ʔân* *siŋ-hâːw* ***ʔʌŋ-****tôn-tʰîw*
1SG.IRR-help missis DIST way which 3SG DES-go 1SG.IRR-send-give
'I will help that woman, which way she wants to go, I will send her.' (W2_15_M_3_020)

Other merged pronouns in the corpus are *ʔɜŋ-* from *ʔɜ̂ː* (1PL.INCL), *mʌŋ-* from *mâj* (2SG), *pɜŋ-* from *pɜ̂ː* (2PL) and *kɜŋ-* from *kɜ̂ː* (3PL). The absence of merged forms of the other pronouns is probably due to the

rare use of them, especially of the dual pronouns. For *ʔân* (3SG) this reason does not seem valid, but a merged form would probably coincide with the first singular irrealis *ʔʌŋ-* and thus might not be used.

2.7 Desiderative siŋ-

The marker *siŋ-* is only tentatively called "desiderative", as although its most usual use is to express a wish of the agent, there are clauses where this is not the case (37).

(34) *ɲûŋ* *siː* *mâj* ***siŋ****-rɛ̂n*
work what 2SG DES-do
'What kind of job do you want?' (WPKT_14_L_017)

Although *siŋ-* is usually translated as 'want', it is not a free morpheme, as first, it cannot be negated on its own (35) and second, it is not used as a main verb, but is always bound to *bɤ̂n* 'get' in the case of "wanting something" (36):

(35) *ʔʌ̂w* ***ʔaːw-siŋ-sôm-maʔ*** *rɤ̂n*
1SG NEG-DES-waste-NEG silver
'I don't want to waste money.' (WPKT_14_L_427)

(36) *ʔʌ̂w* ***siŋ-bɤ̂n*** *pleː* *kʰjîː* *ʔuː-cuŋ* *ʔɔ̂ː*
1SG DES-get earring gold one-CLF POL
'I want a pair of golden earrings.' (WPKT_14_L_014)

As already mentioned, in some sentences, *siŋ-* does not have a desiderative function:

(37) *kɔ̂ːn* *dɜʔ* *kâːj* *dîː* ***siŋ****-dɜʔ* *lôj* *kʰuː-jâːj*
child small 3DU MEDL DES-small only COMPAR-1DU.EXCL
'That two children were younger than we were.' (W2_15_M_3_016)

The preceding example is from a book of short stories (Sa Pe 2015) and the present translation is from the author of the book. A further clarification with that author resulted in 'more younger' for *siŋ-dɜʔ* and therefore the marker may have an emphasizing function. This is compatible with other non-desiderative sentences in that *siŋ-* occurs. More data is needed to clarify whether there are two homophonous morphemes with different functions or whether there is only one morpheme with a more general meaning.

2.8 Negative affixes

There are five negative affixes in Rumai: The prefixes *buː-* and *ʔaːw-* and the suffix *-maʔ* all express on their own simply a negative meaning of the verb, but their use depends on different clause types, namely independent and dependent clauses. Furthermore, the two prefixes can in independent clauses be used together with the suffix. The other two negative affixes are *ɲjʌm-* 'not yet' and the prohibitive *kʰuː-*. While *ɲjʌm-*, like *buː-* and *ʔaːw-*, sometimes occurs with the suffix *-maʔ*, this is never the case with *kʰuː-*.

2.7.1 Negators buː-, ʔaːw- *and* -maʔ

Of these three negative markers, *ʔaːw-* and *-maʔ* occur only in independent clauses while *buː-* is used in dependent as well as in independent clauses, but in the latter ones, it is always accompanied by the suffix *-maʔ*. Two factors are relevant for this distribution: First, the change of the constituent order in independent clauses and second, the reinforcement of negation.

In Rumai, all independent clauses have the constituent order SV/AVP, but most of the dependent ones have the order VS/VAP. Example (38) shows this feature with a relative clause, following the main clause:

(38) *piː* *ʔʌ̂w* *nʌŋ-dih-ʔûːn* ***cəʔuʔ*** ***tʰî̂w*** ***ʔʌ̂w*** ***tʌ-ʔân***
elder.sibling 1SG IRR-read-keep book give 1SG OBL1-3SG
'My sister will read the book that I have given to her.' (ENEG_13_M_1/2_145)

Conditional clauses, although they have the same constituent order as independent clauses, namely SV/AVP, never contain the suffix *-maʔ*. This indicates their dependent character.

(39) *paːt* *bʌ̂ːn* *kʰʌː* *ʔʌ̂w* ***buː-tʌ̌n*** *ʔʌ̂w* ***buː-hâːw-maʔ*** *[…]*
week back COND 1SG NEG-free 1SG NEG-go-NEG
'"Next week, if I am not free, I will not go […]."' (ENEG_13_M_1/2_149)

Given the observation, as described e.g. by Bybee (2001), that dependent clauses are more conservative than independent ones, one can assume that the original general constituent order in Palaung was VS/VAP.

Dryer (1988) investigates in his study the position of negation markers. In his sample, there are 53 verb-initial languages and all but of one have preverbal negation. Therefore, post-verbal negation in verb-initial languages is highly unusual (Dryer 1988:97). This leads to the assumption, that the preverbal negator *buː-* was originally the general negator and only after the completed constituent order change in the independent clauses, the post verbal negator *-maʔ* was introduced.

Initially, *-maʔ* was probably used to reinforce the negation and therefore, the development of the interplay between preverbal *buː-* and post verbal *-maʔ* in Rumai independent clauses can be compared to the cycle explained by van der Auwera (2010:78-79). The possible stages for Rumai are the following:

Figure 1: *Negation cycle*

Stage 1 →	Stage 2 →	Stage 3 →	Stage 4 →	Stage 5
buː-	*buː-*	*buː-*		
	buː- … -maʔ	*buː- … -maʔ*	*buː- … -maʔ*	
		-maʔ	*-maʔ*	*-maʔ*

(Adapted from van der Auwera 2010:79)

Examples for the contemporary use of the two affixes:

(40) *ʔəkʰjɯŋ* ***buː-dɛh*** *piː* *ʔʌ̂w* *pʌ-kʰəlɛp,* *tɔ̂ːn* *bɔːk* *ʔʌ̂w* *nɔh.hŋjaːp*
time NEG-come elder.sibling 1SG OBL2-house every time 1SG worried
'When my brother does not come home, I am always afraid.' (ENEG_13_M_1/2_167)

(41) *ʔʌ̂w* ***buː-təŋnâːw-maʔ*** *redijo*
1SG NEG-listen-NEG radio
'I will not listen to the radio.' (ENEG_13_M_1/2_088)

(42) *ʔʌ̂w* *dʌh-**maʔ*** *kjɔh* *ʔiŋkəlik*
1SG speak-NEG language English
'I don't speak English.' (ENEG_13_M_1/2_022)

The examples show that the dependent clauses (40) are still on stage one, while the independent clauses (41; 42) are on stage four. The affixes *buː-* and *-maʔ* in combination often express the meaning 'not anymore' and example (41) can also be translated as 'I will not listen to the radio anymore'. As already mentioned, the negative prefix *ʔaːw-* only occurs in independent clauses. It can be used alone, as in example (43), or together with *-maʔ* (44).

(43) *kɜ̂ː* ***ʔaːw-mɛh*** *lʌː* *bʌk* *siː*
3PL NEG-exist COMPL ride what
'They didn't have something to ride.' (PCHT_16_L_6_060)

(44) *kâːj ciː* ***ʔaːw-mɔj-maʔ***
3DU TCL NEG-tired-NEG
'They (two) were not tired.' (PCHT_16_L_6_059)

(45) ***ʔaːw-moh-maʔ*** *nɔ̂ː dɜʔ*
NEG-be.so-NEG hill small
'It wasn't a small hill.' (PCHT_16_L_6_046)

For the use of the negative marker *ʔaːw-*, the sample shows a very vague picture. What is clear, is that it is only used in independent clauses and the consistent negation of *moh* 'be so' that is *ʔaːw-moh-maʔ* (45).

2.7.2 'not yet' ɲjʌm-
The prefix *ɲjʌm-* 'not yet' can occur with or without the suffix *-maʔ*. In the following example, although both clauses are not verb-initial, the absence of *-maʔ* in the first clause probably indicates its dependency from the second one.

(46) *ʔʌ̂w* ***ɲjʌm-****mɛh* *ʔəkʰjuŋ* *ʔʌ̂w* ***ɲjʌm-****dih-****maʔ*** *caʔuʔ* *nîː*
1SG not.yet-exist time 1SG not.yet-read-NEG book PROX
'As I did not yet have the time, I have not yet read this book.' (ENEG_13_M_1/2_154)

A preverbal marker with the meaning 'not yet' seems to be a common feature in Palaung, as it occurs also in Shwe with *ɦnɐm* (Milne 1921:176) and in Ruching with *ɦŋam* (Deepadung et al. 2015:1078). In the dictionary (Unknown 2012:131), these two morphemes are cited as verbs with the meaning 'dilatory, stiff'[4]. Therefore, a former common verb of the Palaung varieties with this meaning is probably the source of the negation marker *ɲjʌm-*.

2.7.3 Prohibitive kʰuː-
The following examples show the use of the prohibitive prefix *kʰuː-*:

(47) ***kʰuː-****bîː* *biʔ* *ŋâj*
PROH-forget close fire
'Don't forget to turn off the light.' (WPKT_14_L_091)

(48) *mâj* ***kʰuː-****bîː* *biʔ* *kɔmpjutaː* *vâːj* *dâː* *mâj*
2SG PROH-forget close computer after use 2SG
'Don't forget to turn off the computer after you have used it.' (WPKT_14_L_090)

The agent pronoun can be present in a prohibitive clause (48). This may be a kind of emphasis, but also here, more investigation is needed.

2.9 Reciprocals kə- *and* laj-

The original and main reciprocal marker in Rumai is *kə-*, and a similar morpheme is also found in Shwe with *kʌr* (Milne 1921:52) and in Ruching with *ka* (Deepadung et al. 2015:1073). There are also similar reciprocal markers in some Munda languages with for example *kol-* in Kharia and *ko-* in Juang (Pinnow 1966:115), which probably have the same origin as the Palaung markers, as Pinnow (1966:115) and Sidwell and Rau (2015:323) mention. However, the marker cannot be traced back to proto-Austroasiatic (Sidwell 2015:323).

The meaning of the verbs prefixed by *kə-* is often the classical acting of two participants equally on each other, but sometimes the result is figurative like in *kə-vîː* 'spin, rotate' from *vîː* 'return' and *kə-lɛp*

[4] The corresponding verb in Rumai is according to the dictionary *moj* (Unknown 2012:131).

'wrong' from *lɛp* 'enter'. *kə-* is very productive, as it is also used with loanwords like *kʌp* 'tighten', that is a loan from Shan, in the last of the examples in (49):

(49)	*gʌk*	'bite'	*kə-gʌk*	'bite each other'
	ŋʌh	'hit'	*kə-ŋʌh*	'hit each other'
	lɤ̂j	'pursue'	*kə-lɤ̂j*	'pursue each other, race'
	câːm	'test'	*kə-câːm*	'test each other, fight'
	kʌ̂w	'play'	*kə-kʌ̂w*	'play together'
	kjɤ̀j	'change'	*kə-kjɤ̀j*	'exchange'
	vîː	'return'	*kə-vîː*	'spin, rotate'
	lɛp	'enter'	*kə-lɛp*	'wrong'
	kʌp	'tighten'	*kə-kʌp*	'compose'

laj- which is derived from the verb *lâːj* 'take' is probably a loan from Shan *lɐy* 'get' and it is a more recent development, as its use is more restricted than *kə-*. In (50) it has rather the meaning 'each':

(50)	*ʔɜ̂ː*	***laj-****caʔ*	*lôj*	*pʰʌh*	*bʌ̂ːn*	*dîː*
	1PL.INCL	RECP-start	only	only	back	MEDL
	'We just start each [task] after another.' (C1_17_MS_8_015)					

(51)	*kâːj*	*giːj-****laj****-kəbɜː*	*pʰʌh*
	3DU	IPFV-RECP-level	only
	'Both [works] alternate.' (C1_17_MS_8_007)		

Example (51) shows an important function of *laj-*, namely the reinforcement of the reciprocal meaning. The verb *kəbɜː* 'be level' is derived from *bɜː* 'can' with the common reciprocal prefix *kə-*, literally resulting in 'can each other'. Other verbs of this kind are the following:

(52)	*mɛh*	'exist, have'	*kəmɛh*	'love'	*laj-kəmɛh*	'love each other'
	vâːj	'after'	*kəvâːj*	'pity'	*laj-kəvâːj*	'pity each other'
	**ɟɔj*[5]	'help'	*kəɟɔj*	'help'	*laj-kəɟɔj*	'help each other'

There are two verbs that are preferably or only used with *laj-*. The first one of them is *jɛ̂w* 'see'. This verb can be used with both prefixes, but *kə-jɛ̂w* has rather the literal meaning of 'see each other', whereas *laj-jɛ̂w* is much more common and means 'meet each other' (53). The other verb is *leh* 'descend' and it is only used with the prefix *laj-*, resulting in the meaning 'go away' (54).

(53)	*buː-hlʌ̂ːn-dâːŋ*	*ʔɜ̂ː*	*nʌŋ-****laj-jɛ̂w****-ʔûːn*	*bôː*
	NEG-long-big	1PL.INCL	IRR-RECP-see-keep	EMPH
	'We will meet each other very soon.' (WPKT_14_L_265)			

(54)	*kɜ̂ː*	***laj-leh****-hâːw*	*daʔ*	*lôŋ*	*bjâj*	*ʔʌ̀ː*	*nîː*
	3PL	RECP-descend-go	other.than	field	forest	dark	PROX
	'They left from this dark jungle.' (PCHT_16_L_6_028)						

The reinforced reciprocals in (52) and the use of *jɛ̂w* 'see' and *leh* 'descend' with the prefix *laj-* show a grammaticalization path in the sense of the bleaching of a grammatical morpheme, and its reinforcement by another one. The spread of the reinforcing *laj-*, however, is still minimal.

The verb *lâːj* 'take', from which *laj-* is derived, is much less used as a main verb, than the original Rumai verb *tɛh* 'take' and the only use of *lâːj* following another verb in the corpus is shown in (55):

[5] **ɟɔj* 'help' is a loan from Shan *cɔj* 'help'.

(55) *pəŋaː* *kəlok* *kʰrîː* *ciː* *tuː.ʔîː* *ʔaːw-bɜː* *lʌ* *bjâː-**lâːj*** *nʌː*
education pot gold TCL person NEG-can COMP steal-take TAG
'Education is a golden pot that nobody can steal, isn't it? (W1_034)

Therefore, *lâːj* is not a post-verbal secondary verb with a grammatical function, but it is used here in conjunction with a verb of a similar meaning, the two verbs denoting together a single event. There are other compounds of this kind like *kyɔh-dʌh* 'speak-say'. The use and function of such compounds need further investigation.

3. Conclusion

The Rumai verbal affixes can, as shown in table 1 and in the sections of this paper, be categorized into the four domains aspectual, modality, negation and reciprocal. Of the aspectual prefixes, three, namely the imperfective marker *giːj-*, the durative marker *ʔɯN-* and the new situation marker *hɔ̆j-* are likely derived from the lexical verbs *gɔ̂ːj* 'stay', *ʔûːn* 'keep' and *hɔ̀ːj* 'finish', respectively. The affixes are semantically as well as phonetically bleached and they show grammaticalization paths that are also found in other languages: Heine and Kuteva (2002) list for the category "continuous" the sources 'stand' and 'keep'. (Heine and Kuteva 2002:330). A source verb 'finish' for the new situation marker is wide spread in the languages of Southeast Asia (Jenny et al. 2015:97–98).

The two source verbs *ʔûːn* 'keep' and *hɔ̀ːj* 'finish' are also used as post-verbal secondary verbs, the former has a similar function to the imperfective prefix *giːj-* and the function of the latter is similar to that of the new situation marker *hɔ̆j-*. The source of the inceptive prefix *ʔə-* and the experiential prefix *tʌm-* is as of now unknown. The same is true for the modality categories, the irrealis *nʌŋ-* and the desiderative *siŋ-*. Moreover, the two markers have the vaguest meanings of the Rumai affixes.

Rumai has a rather large inventory of negation markers. The "neutral" negators also include the only verbal suffix *-maʔ* that was probably invented for reinforcement. However, the fact that there is only one suffix shows the general pre-verbal "modification" in Rumai. Of the negation markers, only *ɲjʌm-* 'not yet' can probably be traced back to a verb, this one having the meaning 'dilatory, stiff'.

The two reciprocal prefixes *kə-* and *laj-* have mainly complementary distributions. While *kə-* is the original reciprocal marker, still productive since it is also freely used with loan verbs, the marker *laj-* is mostly used to reinforce reciprocity with verbs that have been lexicalized, including the prefix *kə-* like *kəmɛh* 'love'.

This study has provided initial observations regarding verbal affixes in Rumai. Many questions are still left open, one of the most important being the interplay between affixes and between prefixes and post-verbal secondary verbs that have similar functions. Future research is needed to clarify such questions.

Sources of the examples

Short name	Description
C1_17_MS	Conversation about tea and rice cultivation in a village
C3_17_MS	Conversation about the life in a village
EEN_16_M/L	Questionnaire about emphasis and nominalization, Burmese/English
EGGD_16_M/L	Questionnaire about the use of "give" and "get" and about ditransitive constructions, English
ENEG_13_M	Questionnaire about negation, Burmese/English
EQB_17_M	Questionnaire about Burmese constructions, Burmese/English
I_15_M	Interview of a young woman with a monk
I_17_MS	Interview of a monk and two villagers about their life in the village
PCHT_16_L	Picture story: *The chicken thief*, Béatrice Rodriguez, 2008, *Der Hühnerdieb*, Wuppertal: Peter Hammer Verlag
WA_15_M	Sentences containing aspectual markers from: *Being clever by texts*, Sa Pe 2015, Mandalay: self-publishing
WPKT_14_L	Phrase book Rumai – Shwe – Burmese – English 2014, Lashio: Ta'ang Students and Youth Union

Short name	Description
W1	Story in *Our Ta'ang magazine*
W2_15_M	Short story in: *Being clever by texts*, Sa Pe 2015, Mandalay: self-publishing, pp. 78-79
W4_15_M	Short story in: *Being clever by texts*, Sa Pe 2015, Mandalay: self-publishing, pp. 38-39

Structure of the labels indicated by the examples:

1. Short name: shortcut_year_place of recording/edition
2. _consultant(s)/author_example in toolbox

First letter of the shortcuts

E	elicitation
P	picture story
I	interview
C	conversation
N	narrative
W	written text

Places

M	Mandalay
MS	Man Sat village (Namkham Township)
L	Lashio

Consultants

Nr.	Gender	Y.o.b	Origin	Education	Occupation	Languages
1	f	1996	Na Aw Gyi village, Man Ton township		bachelor student	Rumai, Burmese, English
2	m	1994			bachelor student, novice	Rumai, Burmese, English
3	m	1988	Man Sat village, Nam Hkam township	Bachelor diploma 2017	monk	Rumai, Burmese, English, Pali
6	m	1993	Sar Lu village, Nam Hkam township			Rumai, Burmese
7	f	1959	Man Sat village, Nam Hkam township	public school in Shan, grade ?	housewife, farmer	Rumai, Shan, (Burmese)
8	m	1981	Man Sat village, Nam Hkam township	public school in Burmese, grade ?	farmer	Rumai, Burmese
9	f	1998	Pha Daen village,	public school in Burmese, grade 10	housewife, farmer	Rumai, Burmese, (English)

References

Bybee, Joan L. 1998. "Irrealis" as a Grammatical Category. *Anthropological Linguistics* 40(2):257–271.

Bybee, Joan L. 2001. Main clauses are innovative, subordinate clauses are conservative'. In Joan L. Bybee and Michael Noonan (eds.) *Complex Sentences in Grammar and Discourse. Essays in Honor of Sandra A. Thompson*: 1–17.

Bybee, Joan L. and Michael Noonan (eds.). 2001. *Complex Sentences in Grammar and Discourse. Essays in Honor of Sandra A. Thompson*. Amsterdam, Philadelphia: John Benjamins.

Dahl, Östen (ed.). 2000. *Tense and Aspect in the Languages of Europe*. Berlin: Mouton de Gruyter.

Deepadung, Sujaritlak, Ampika Rattanapitak and Supakit Buakaw. 2015. Dara'ang Palaung. In Mathias Jenny and Paul Sidwell (eds.) *The Handbook of Austroasiatic Languages*:1065–103.

Dryer, Mathew S. 1988. Universals of negative position. In Michael Hammond, Edith Moravcsik, and Jessica Wirth (eds.) *Studies in syntactic typology*: 93–124.

Ebert, Karen H. and Fernando Zúñiga. 2001. *Aktionsart and Aspectotemporality in Non-European Languages*. ASAS 16. Zürich: ASAS.

Haiman, John. 2010. *Cambodian : Khmer*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Hammond, Michael, Edith Moravcsik and Jessica Wirth (eds.). 1988. *Studies in syntactic typology*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Heine, Bernd and Tania Kuteva (2002). *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Horn, Laurence R. 2010. *The Expression of Negation*. Berlin: Walter de Gruyter.

Jenny, Mathias. 2001. The aspect system of Thai. In Karen H. Ebert and Fernando Zúñiga (eds.) *Aktionsart and Aspectotemporality in Non-European Languages*: 97–140.

Jenny, Mathias and Paul Sidwell (eds.). 2015. *The Handbook of Austroasiatic Languages*. Leiden: Brill.

Jenny, Mathias, Tobias Weber and Rachel Weymuth. 2015. The Austroasiatic Languages: A Typological Overview. In *The Handbook of Austroasiatic Languages*, ed. by Mathias Jenny and Paul Sidwell, 13–143. Leiden: Brill.

Johansson, Lars. 2000. Viewpoint operators in European languages. In *Tense and Aspect in the Languages of Europe*, ed. by Östen Dahl, 27-187. Berlin: Mouton de Gruyter

Milne, Mary Lewis Harper. 1921. *An Elementary Palaung Grammar*. Oxford: Clarendon Press.

Payne, Thomas E. 1997. *Describing morphosyntax. A guide for field linguists*. Cambridge: Cambridge University Press.

Pinnow, Heinz-Jürgen. 1966. A Comparative Study of the Verb in the Munda Languages. In *Studies in Comparative Austroasiatic Linguistics*, ed. by Norman H. Zide 96–193. The Hague: Mouton.

Roos, Olivier. 2001. Mandarin Chinese *-zhe*. In *Aktionsart and Aspectotemporality in Non-European Languages*, ed. by Karen H. Ebert and Fernando Zúñiga, 49–71. Zürich: ASAS.

Sa Pe. 2015. *bjat gawj khaeng bae* (Being clever by texts). Mandalay: Self-publishing.

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

Smith, Carlota S. 1991. *The Parameter of Aspect*. Dordrecht: Kluwer Academic Publishers.

Unknown. 2012. *Pap om ngae bran–ta'ang (sam long, rumaj, rucing)*. [Dictionary Burmese–Palaung (Shwe, Rumai, Ruching)]. Namhsan: Ta'ang Literature and Culture Committee.

van der Auwera, Johan. 2010. On the diachrony of negation. In *The Expression of Negation*, ed. by Laurence R. Horn, 73–109. Berlin: Walter de Gruyter.

Zide, Norman H. (ed.). 1966. *Studies in Comparative Austroasiatic Linguistics*. The Hague: Mouton.

PROTO-NICOBARESE PHONOLOGY

Paul Sidwell
Australian National University
paulsidwell@gmail.com

Abstract

This paper presents a reconstruction of Proto-Nicobarese phonology comprising a segmental inventory and syllable structure. Nicobarese is a branch of Austroasiatic languages located on an island chain in the Andaman sea. Being the only branch of the phylum located on islands, and on a well-known trade route, Nicobarese provides an important point of comparison with other AA languages in India and Mainland South-East Asia. While much work still needs to be done, the current effort brings together relevant known work on these languages.[1]

Keywords: Nicobarese, Proto-Nicobarese, phonology, reconstruction

ISO 639-3 codes: caq, crv, tef, ncb, nik, sii

1 Introduction

The Nicobarese languages form a small branch of the Austroasiatic[2] phylum, with just a few thousand speakers on an island chain in the Andaman Sea. The 2011 India census lists a total population for the islands at 36,842 (down from 42,068 in the 2001 census due to the 2004 tsunami) with about 30% of that population being from the mainland such as government staff and plantation workers (including speakers of Hindi, Tamil, Telugu, Santali). Since the 1960s in particular, the islands have been off limits to outsiders, as a consequence of the attitude of the government of India. Historically the islands lie on the sea route between India and the Far East, and it is reported that the islands were conquered by the Indian Chola dynasty in the eleventh century (Murthy 2005:21). The earliest mention of the Nicobars is apparently in Ptolemy's atlas circa 150 A.D.

De Röepstroff (1875) describes the islanders a century and a half ago having extensive trade relations with the outside world and many being familiar with outside languages. He remarks of the Nicobarese:

> They are great linguists. You may, to a certain extent, tell hie history of the islands as far as it has been connected with trade through the languages spoken. The oldest men yet speak the corrupted Portuguese that still lingers in the East. Middle-aged men speak very often a little bad sailor-English; the young men, especially South and East, speak Burmese; the boys a little Hindistani: all talk Malay and their own language. At Car Nicobar they talk English pretty well.
>
> (De Röepstroff 1875:14)

We see the linguistic impact of this history of contact in many loan words noted in the available dictionaries/lexicons. Some examples from Man's (1889) dictionary of Nancowry: ***lēbare*** 'book', ***arōe*** 'rice', ***shapēo*** 'hat' < Portuguese, ***kapo*** 'cattle', ***lapu*** 'gourd', ***koching*** 'cat' < Malay. It is clear that loans have reached well into the cultural vocabulary, and in this paper the approach taken endeavours to exclude loan words as much as possible in order to reflect the native phonology and lexicon.

The most extensive sources available for Nicobarese are the colonial era dictionaries and grammars that deal with just two (Car and Nancowry) of the six languages conventionally distinguished in the literature.

[1] I would like to thank individuals who assisted with comments and feedback in the preparation of this draft: Jessica Johnson, Ryan Gehrmann, Mathias Jenny, and the remarks of two annonymous reviewers who were harsh but helpful in their contributions. Also I acknowledge with gratitude financial support I received from the Australian Research Council under Future Fellowship award FT120100241, and assistance from the Max Planck Institute for the Science of Human History (Jena), which supported work on this project

[2] Abbreviated to AA in tales and formulas.

The languages are mostly named after the islands on which they are spoken. Those languages, and the best available resources are listed here:

19) **Car:**[3] Whitehead (1925) dictionary with 6705 entries extracted, Das (1977) lexicon of which 2282 items are extracted, Critchfield-Brain (1963) 87 page typescript lexicon in close transcription.

20) **Chowra**: Man (1889) index provides approx. 380 words.

21) **Teresa and Bompoka**: Man (1889) index provides approx. 380 words.

22) **Central** (Nancowry/Müot,[4] Camorta, Trinkat, Katchall): Man (1889) dictionary of Nancowry 5961 entries extracted, Radhakrishnan (1981) study of Nancowry morphology lists 778 lexical root and their derivatives.

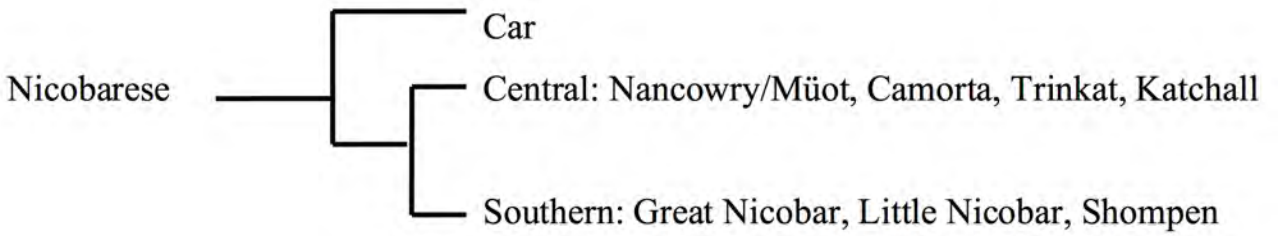
23) **Southern** (Great and Little Nicobar, Pulo Milo, Kondull): Man (1889) index provides approx. 380 words.

24) **Shompen** (interior of Great Nicobar Island): Man (1889) index provides approx. 380 words, Chattopadhyay & Mukhopadhyay (2003) list approx. 700 words, Gnanasundaram & Rangantha (1995) list some 70 words.

Of the above the Car and Nancowry sources are the most extensive and reliable, so the analyses and reconstruction presented in this paper are based primarily on just those two languages. This is a fundamental limitation that may never be overcome.

In Sidwell (2015) I presented a preliminary statistical analysis which suggests that the Nicobarese lects of the Central island group (Nancowry, Katcall, Camorta, Kondul, Pulo Milo, Teressa) form a coherent dialect grouping that coordinates with Car, forming a tree with two main branches. Since then I prepared a more extensive dataset for phylogenetic analysis, incorporating data from the comparative word lists in the appendices to Man (1889).[5] The results of that work are reported separately (a paper is in preparation) support the provisional classification followed here which places the Nicobarese lects into three primary groups, consistent with the geographical distribution of the islands as seen in the Wurm & Hattori map reproduced below. This scheme is diagrammed as follows:

Figure: *Nicobarese varieties*



The above configuration supersedes the study by Blench and Sidwell (2011), which hypothesized that Shompen may be more closely related to Aslian or otherwise represent a branch intermediate between Nicobarese and Aslian. An unpublished statistical analysis[6] suggests that the internal diversification of Nicobarese began around 2,200 years BP, based on calibrations with Austroasiatic languages with well known histories, Khmer and Mon, plus inferences regarding the internal diversification of Bahnaric and Vietic.[7] It is also possible that this estimate is actually too old; accelerated lexical change due to word

---

[3] These resources are largely extracted and available online, the Whitehead, Das and Man data at: http://sealang.net/monkhmer and the Braine and Radhakrishnan data at http://sites.google.com/view/paulsidwell/nicobarese-languages-project.

[4] *Müot* is preferred in place of *Nancowry* by V.R. Rajasingh (CIIL Mysore) who has been relatively active in Nicobarese research of late. The term *Central Nicobarese* also enjoys use in the literature.

[5] https://docs.google.com/spreadsheets/d/14pLIzYnj4Vvoscv4Zy7ACmQVQrNgDFsqeBAUZwlnF0c/edit#gid=0

[6] Phylogenetic analyses presented at the workshop "Integrating inferences about our past" June 22-23 2015, Max Planck Institute for the Science of Human History (Jena, Germany) offering a Maximum Clade Credibility Tree of the CTMC + Gamma Relaxed analysis - contact the author for further details.

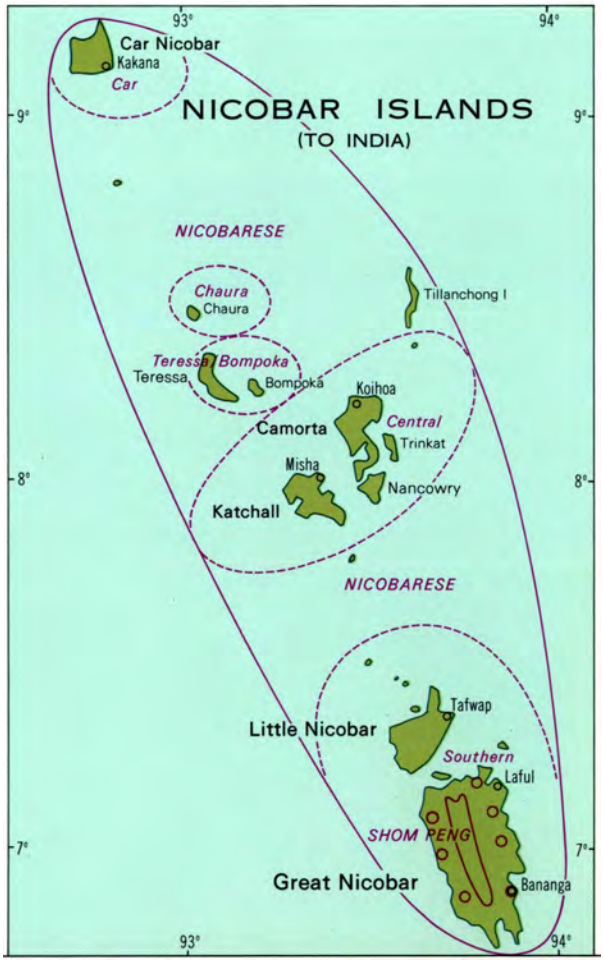
[7] Both Mon and Khmer have a history of writing that goes back to the middle of the first Millennium CE, and reasonable assumptions about the diversification of Bahnaric and Vietnamese can be made based on the known

tabooing could indicate a significantly younger age. On the issue of the effect of tabooing on the lexicon, it is worth quoting Man at length:

> The diversities of speech which have sprung into being among the four communities in question, are, moreover, no doubt in great measure ascribable to the operation of a superstitious custom, which here, as in various other remote regions, has effected constant changes in the language of the inhabitants; but in every instance of this kind such changes have been limited to the area of the particular community concerned. The practice referred to is based on a firm belief in an after-existence, and requires that the names of deceased relatives and friends shall be tabued for a certain lengthened period-generally about one generation-for fear of summoning or offending the ghost of the person so named. Therefore, as their system of personal nomenclature not only permits anyone to invent or adopt a name for him or herself, but also to take for this purpose any word in the language without consideration of its being in general use, it naturally follows that new words have to be instantly coined to take the place of those whose use is tabued in consequence of death, and thus many striking changes are introduced into the language in the course of each generation. (Man 1889:viii)

At the same time we are also pulled in the other direction by the possibility of undocumented diversity (recent or extinct) that could indicate a greater age, but we lack the resources to pursue the problem further in this paper. We can also hope that estimates of initial Austroasiatic settlement of the islands might also to be calibrated with archaeological evidence, something that is also still lacking.

Map: Nicobar languages from Wurm & Hattori (eds.) (1981/83) Language Atlas, (fragment from full map prepared by D. Bradley).



history of Indo-China and interaction with Chamic (Austronesian) settlement on the Vietnamese coast since the mid first first Millennium BCE. These facts allow for at least four calibration points in modeling the rate of change in the Austroasiatic family tree.

Taking the family tree above as our starting point, we note that the two lects for which substantial documentation is available—Car and Nancowry—fall across the two principal coordinating branches and we can hypothesize that features found to be held in common may be reconstructed to the pNicobarese level, thus our working method treats Car and Nancowry as criterion languages for the comparative reconstruction. Additionally, a root attested in only one of these lects, but having apparently cognates elsewhere in Austroasiatic, can be assumed to belong to proto-Nicobarese (abbreviated to pN in tables/formulas). Words only found in one Nicobarese sub-group and not otherwise attested in Austroasiatic are not reconstructed to proto-Nicobarese. It is acknowledged that we have no real sense of the extent to which Car and Nancowry have influenced each other after diverging from proto-Nicobarese, nor to what extent present or now extinct Nicobarese lects may have played a role in the history of the group, but we cannot base a study on unknowns. Consequently, the present reconstruction is a synthesis of bottom-up and top-down method. There is no reasonable alternative given the state of the available data and the obstacles to field work.

The Nicobarese data are generally difficult to work with. The large colonial era dictionaries are written in Roman orthographies[8] that fail to mark some distinctions while also over-representing some meaningless detail and variation. The languages are highly synthetic (in that regard they are more like Austronesian than Austroasiatic) and yet the published lexicons generally do not segment words morphologically, or segment everything simply into syllables such that the same roots may be represented differently in a variety of contexts. On top of this, the principle works that have attempted morphological analyses (Braine 1970, Radhakrishnan 1970, 1981) incorrectly assume that the lexical roots are generally monosyllabic and thus even their results have to be reassessed item by item by item for a full morphological analysis. In this short paper the focus is on lexical roots, and for identification of these we take advantage of wider Austroasiatic resources, which are now quite extensive, especially Short's (2006) *Mon-Khmer Comparative Dictionary* and the data and search tools available online at http://sealang.net/monkhmer.

This study is not the first to investigate the historical phonology and lexicon of Nicobarese. I obtained from Norman Zide (Chicago) a typescript apparently from 1963 by N. Zide and D. P. S. Dwarikesh titled *The comparative phonology of proto-Nicobarese as derived from Kar Nicobarese and Central Nicobarese: Preliminary version*. This is a 57 page draft that lays out phonological correspondences and a comparative lexicon of of 191 items. That study does not present a reconstruction as such, but is helpful in terms of assisting the interpretation of the dictionary sources, and a proportion of the comparisons made are used here and acknowledged in the appendix. I also obtained from Zide another typescript, also apparently from 1963, *Initial Consonant in Proto-Munda-Nicobarese: some tentative correspondences*. This 19 page draft lists approximately 180 Munda-Nicobarese comparisons and tables the apparent segmental correspondences. It is not a reconstruction, but is a demonstration of the genetic relation between Munda and Nicobarese by showing regularities in the correspondences. My assessment is that many of the comparisons in this paper reproduce known etymologies from, e..g.: Pinnow (1959), Schmidt (1904), plus a large proportion that are speculative and not useful, and it is not relied upon in this study.[9]

Shorto (2006) lists some 317 Nicobarese comparisons in his comparative lexicon and many are taken directly from Radhakrishnan's (1981) lexicon of Nancowry roots, and about two thirds of Shorto's lexical comparisons have informed this study.[10] The analyses presented in this study is based on the set of 266 lexical comparisons given in the appendix to this paper, along with the reconstructions I have based on them. The Appendix is organised to group etymologies according to the timbre of the stressed syllable nuclei; this maximises the utility of of the index since these correspondences are the most problematic and this organisational principle maximises the ease of comparing all the relevant data in context. However, since the contoid correspondences do not automatically group by this method, effort is taken to give relevant data examples in the text.

---

[8] Note that orthographic forms are italicized throughout, they are not normalized to IPA because of the limitations associated with the orthographies.

[9] Scans of both papers are available online at: http://sites.google.com/view/paulsidwell/nicobarese-languages-project.

[10] I was gifted Shorto's research collection by his family. and this included a copy of Radhakrishnan (1981) with Shorto's marginal annotations indicating this identification of Austroasiatic roots and his own reanalyses.

2 Phonological Profiles

2.1 Car

The phonological analysis of Car is based primarily on the works of Braine and Das:

3 An 85-page unpublished (1963)[11] lexicon with explanatory notes, given to this writer by Norman Zide. The notes include a guide to the approximate segmental values of Whitehead's orthography. The lexicon includes about 1600 entries after overlapping entries are merged.
4 Braine's 1970 thesis, which this writer had scanned and retyped
5 Das (1977) Car sketch (a work of mixed quality) which conveniently reproduces much of Whitehead's lexical content in pseudo-IPA.

The (1925) dictionary by Whitehead is also an important work, but the lack of phonetic description, and its reliance on a Roman-based orthography, means that it must be used with particular care. It is also worth noting that Braine (1970), for theoretical reasons, strove to represent her data mostly in a strong morphophonemic notation – this approach is eschewed here in favor of achieving a broad segmental representation as far as practicable.

2.1.1 Car word/syllable structure

Car words are built on simple CV(C) syllables. Except for a modest number of onsets with a medial liquid or rhotic in unassimilated loans, onset clusters are not tolerated. Codas are optional, and open syllables tend to lengthen to preserve moraic weight (much like in other Austroasiatic languages). Various sources write syllables with zero onsets but a glottal stop is assumed in such cases. The inventory of coda segments is not quite as rich as the onsets, but on balance syllables are remarkably symmetrical by Austroasiatic standards.

The phonological word is built up of these simple syllables, minimally just one, but sequences up to four syllables are attested and longer words may be possible, given the richness of the morphological system and one's definition of word. Words consist of lexical roots, which can be one- or two-syllable iambs (the rightmost syllable of lexical roots bears the primary word stress), plus various prefixes, infixes, and suffixes. It is also apparent that historical roots with onset clusters and/or sesquisyllables have often been restructured into disyllabic iambs. As a result of such changes, the full range of nuclei are only found in the main syllables of lexical roots, while unstressed syllables are restricted to having a small number of contrastive nuclei.

Descriptions of Car lack mention of syllable-level tones and/or phonation types. Braine (1970) does devote several pages to discussing pitch in the context the context of phrase and sentence level intonation, remarking:

> As with English, Nicobarese may be readily read by a person knowing the language without any indication of pitch. Because of this marginal function of pitch, and because pitch was omitted from much of the data used in making this analysis, pitch is not indicated throughout the rest of the grammar. (Braine 1970:29)

Pitch, intonation, or phonation, are not discussed further here as it appears that they do not distinguish lexical items; the emphasis is on segmental, syllable, and word phonology. However, quantity is phonologically relevant: main syllables of lexical roots have longer nuclei (and somewhat higher pitch) than other syllables, additionally length is contrastive in these nuclei. Nasalization is also contrastive among stressed nuclei. Note that in Braine's (1963) stressed nuclei are marked with a colon (:) while in her (1970) these they are marked with a acute (´).

[11] The copy of the ms. given to me by Norman Zide is undated but it is noted as 1963 in Huffman's (1986) bibliography.

2.1.2 Car segments

Car segments are tabled below. Note that here and elsewhere the terms vowel and *consonant* are only used in reference to the graphemes used in the orthographic data; segmental values are referred to as either vocoid or contoid, and are characterized in terms of their phonotactic positions (onset, coda, nuclei etc.).

Onsets

/	p	t	c	k	ʔ	
	m	n	ɲ	ŋ		
	v	l	ᵈr	ɽ	j	
	f	s			h	/

Codas

/	p	t	c	k	ʔ	
	m	n	ɲ	ŋ		
	v	l	r	s	j	h	/

Nuclei	**Stressed σ**			**Unstressed σ**		
/	i	ɨ	u	i		u
	e	ɤ	o		ə	
	ɛ	ə	ɔ	ɛ	a	
	(ɛɔ)	a				/

± length /ː/ and nasalization /˜/

The onset contoids lack the contrast of voicing and/or implosion characteristic of conservative Austroasiatic languages, reflecting various mergers discussed further below. At the same time, the coda segments are largely unchanged from proto-Austroasiatic; some mergers/restructuring has occurred, but the outcome has been an inventory essentially the same as one finds in conservative Austroasiatic languages.

Stressed syllable nuclei are quite Austroasiatic in their characteristics; long and short monophthongs occur in stressed syllables, although diphthongs are not. This is not unusual in Austroasiatic, but multiple diphthong are indicated for proto-Austroasiatic (Sidwell & Rau 2015) so clearly some mergers have occurred. The apparent presence of four contrastive vocoids in unstressed syllables is paralleled in Munda, and in both cases relates to a combination of factors, principally word-formation and assimilatory processes; we do not find evidence of contrastive timbre in unstressed syllables of disyllabic/sesquisyllabic roots.

The following phonotactic statements are noted in relation to Car:

1. As codas of stressed syllables, nasals are briefly pre-stopped (’affricated’ in Braine’s terms), e.g. [ᵇm, ᵈn, ʲɲ, ᵍŋ].
2. The rhotic /r/ is an alveolar flap, usually pre-stopped, [t] before a voiceless consonant and [d] elsewhere, i.e. [ᵗr, ᵈr].
3. The *ɾ* notation is described as a retroflex by Braine (1970) and transcribed *ɽ* here.
4. Before coda /h/ vowels are exceptionally short in duration – approximately one half a mora.
5. /ɤ/ and /ə/ were formerly allophones of a single segment. Today they contrast only in long and open syllables. Compare: *lɤ́ːkə* ‘by way of’ vs. *ləːkə* ‘to do well’.
6. Epenthetic [i] occurs before the palatal codas /c/ and /ɲ/; this is written as *i* in Whitehead’s orthography (e.g. *söich* ‘to wash’).
7. All nuclei have slightly nasalized allophones: in unstressed closed syllables beginning with /ʔ/ or /h/ ending with a nasal consonant, and after a nasal consonant when neither /ʔ/ or /h/ follows, for example: *ʔĩnɾúj* ‘fly’, *tanɲ̃j* ‘five’, *hũmlúm* ‘gold’, *məl* ‘just now’, etc.
8. [ɛ] and [ɛɔ] are in complementary distribution, [ɛɔ] occurring in stressed syllables with codas *k, ŋ,* and [ɛ] occurring elsewhere.
9. [æ] occur in the data due to the occurrence of English loan words.

2.2 Nancowry

The analysis of Nancowry in this study is based primarily on the works of Radhakrishnan (1970, 1981) and Rajasingh (2016), and the colonial era dictionary and grammar of Man (1889). There are also materials in English by de Röepstroff (1875, 1884) but these are essentially superseded by Man. The 19th century sources suffer from the usual problems of phonetic interpretation of early works, yet still prove to be useful due to the overall lack of good descriptive works.

2.2.1 Nancowry word/syllable structure

Nancowry word and syllable structure follows broadly the same patterns as Car, with some minor differences:

10. Unstressed syllables are restricted to just three contrastive nuclei.
11. There is only one rhotic.
12. The phonetic value of the oral fricative is not quite clear: it is described as "grooved" by Radhakrishnan, and written *sh* by Man suggesting that it may vary to [ɕ] or [ʃ].
13. The labial glide written variously *v* and *w* is written *ʋ* by Rajasingh, and this is assumed to be representative of the real value.
14. Diphthongs are a strong feature of Nancowry, although there is significant allophony in the diphthongs and the phonological analysis relies on assumptions that may be challenged.

2.2.2 Nancowry segments

Nancowry utilizes the following inventory:

Onsets

/	p	t̪	c	k	ʔ	
	m	n̪	ɲ	ŋ		
	ʋ	l r	j			
	f	s			h	/

Codas

/	p	t	c	k	ʔ	
	m	n	ɲ	ŋ		
	ʋ	l r	j			
		s			h	/

Nuclei	**Stressed**			**Unstressed**		
/	i	ɯ	u	i		u
	e		o	(e)	(ə)	
	ɛ	ə	ɔ		a	
	æ	a				
	ia	ɯa	ua			/

± nasalization /˜/

Quantity is not contrastive in its own right, although nuclei do vary in length with stress placement. Lexical roots bear iambic stress and their nuclei are pronounced approximately one and a half mora long in closed syllables and two mora long in open syllables. Rajasingh consistently notes this with (˙, ː) in his works.

The following should be noted:

5 /h/ is treated consistently as [x] by Rajasingh.
6 Apical stops are dental as onsets, alveolar as codas.

7 In unstressed syllables *i ~ e* are in variation and can be treated as /i/, although typically written *e* in older sources.
8 In unstressed syllables *a ~ ə* are in variation and can be treated as /a/.
9 Radhakrishnan (1970:32) says for the nasal codas, "Word finally nasals including [ŋ] have most often a clipped articulation on a par with the inaudible release of the oral stops." This statement hints at weak pre-stopping similar to that reported for Car by Braine.
10 The rhotic /r/ is described as a pre-stopped [ʔr] in onsets of stressed syllables, and as a flap or weak [d] in onsets of unstressed syllables (note that both and *d* and *r* doublets are frequent throughout Man's dictionary). Rajasingh writes /r/ as ɹ throughout.
11 Diphthong-ed nuclei require further study to arrive at a unified synchronic treatment. Radhakrishnan (1970:25) lists Nancowry as having five diphthongs: [iˑə, iaˑ, uˑə, uaˑ, ɯˑə] and reduces these to three contrastive units /ia, ua, ɯa/ invoking a relation between stress and lengthening that, (p.28) "has not been worked out yet." On p.30 Radhakrishnan further explains that these diphthongs render phonetically [iə, ea, uə, oa, ɯə]. Rajasingh (2016:25-26) identifies seven diphthongs [iə, ua, uə, ɯə, ea, eə, oə] in broad phonetic transcription but does not offer a phonological analysis. The diphthongs do vary prosodically, and this is clearly evident in Man's orthographic forms: compare ***et-kōat*** 'to comb' with the nominalised ***kanūat-kōi*** 'a head comb'.

3 Comparative phonology

3.1 Proto-Nicobarese word/syllable structure

Given the similarities between Car and Nancowry, we can propose that the common structures continue a pNicobarese template. On this basis, proto-Nicobarese lexical roots are reconstructed as follows:

Table 1 monosyllabic $\mathbf{C_1V(C_2)}$, or
Table 2 disyllabic iambs $\mathbf{C_{-1}V(C_{-2}).C_1V(C_2)}$, where C_{-2} is one of several infixes which include ***-n-**, ***-m-**, or copies of main-syllable codas in case of coda-copying.[12]

One possible exception to the above pattern is the item #56 *hʊaŋ 'perspiration' for which the *hʊ- onset is posited, but for the present we will treat *hʊ- as a single onset (i.e. equivalent to IPA ʰ̪ or f̤) until such time as more data indicates a reconsideration.

The following proto-Austroasiatic phonological template is reconstructed by Sidwell & Rau (2015):

Proto-Austroasiatic template:

monosyllables	disyllables (including sesquisyllabic forms)
$\mathbf{*C_i(C_m)VC_f}$	$\mathbf{*C_p(n/r/l).C_iVC_f}$

The proto-Austroasiatic mono-syllables permitted onset clusters with rising sonority. This is reflected in the template above in which $\mathbf{C_m}$ stands for a medial segment that could be a liquid or glide (including [h]). Austroasiatic roots are generally of the form C(C)VC, while sesquisyllables/disyllables reflect forms with inflectional and/or derivational morphology. The phonological restructuring that marked the transition from proto-Austroasiatic to proto-Nicobarese resulted in additional sources of disyllabic forms.

With the shift to preferred CVC pattern in pre-proto-Nicobarese, initial clusters syllabified and the resulting syllables acquired new nuclei in unstressed position. Those unstressed nuclei probably only distinguish three contrastive timbres symbolized *a, *i, *u, although phonetically these were likely to have been short lax vocoids [ĭ, ĕ, ŏ]. The timbre of these unstressed nuclei appears to have been variously conditioned by the place of articulation of adjacent contoids. Some examples follow.

[12] See Radhakrishnan (1981) for a discussion of coda-copying in Nicobarese. This is assumed to be an old process in Austroasiatic shared also by at least Aslian and Khmuic. However, no specific proto-forms with copied codas are dealt with in this study.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
49	*tulan	‘python’	*tulân*	tulán		tulan	*tu-lan*	*tla:n #1205.b
88	*kaɹi:	‘road, path’	*kaiyī*	kají	kɑji:			*kraʔ #162.a
235	*pulo:ʔ	‘thigh’	*pulô*	puló?	pulɔˑʔ			*bluːʔ #223.a
68	*ciɹavᵕ[13]	‘deep’	*chiyàu*	cijáw		ʔaṛuːʔ	*a-ṛū*	*ɟruːʔ #172.a
124	*ʔijɯam	‘breathe’	*eyām*	ʔijɯ́am	ʔijɯəˑm			*jhu(ə)m ‘breathe, live’ #1299.a

Disyllables in Nicobarese languages also arise due to affixation. Both Brian (1970) and Radhakrishnan (1970, 1981) discuss affixation in Car and Nancowry respectively, however both these scholars tended to equate the stressed CVC with the lexical root, and thus many prefixes were incorrectly recognised. For example, Radhakrishnan overtly identified the initial syllables in the Nancowry forms in the above table as prefixes, greatly complicating his analysis. At the same time, nasal infixes (primarily derivational) and inflexional prefixation accompanied by coda copying are readily recognised and were essentially handled properly by Brian and Radhakrishnan. Disyllables resulting from affixation also only distinguish three timbres in nuclei of unstressed syllables; some examples follow:

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
Disyllables created by infixation								
2	*sama:	‘jaw’	*shama-lâ-ēshe*			samãːʔ	*sa-mā*	*caʔ #8.a ‘to eat’
62	*kanap	‘tooth’	*kanâp*	kanə́p	kɑnɑˑp	kanap	*ka-nap*	*kap ‘to bite’ #1231.b
205	*kanu:t	‘a comb’	*kanūat-kōi*	kanúat		kanũːt	*ka-nūöt*	*kuət #958b.b
33	*tanaj	‘five’	*tanai*	tanáj	ʈɑnɑˑj	tanij	*ta-neui*	(< pN *taj ‘hand’)
Disyllables created by prefixation								
27	*kapah	‘die’	*kâpâh, kapâh*	kapáh		kapah	*ka-pah*	
158	*ʔinfua	‘dream’	*enfūa*	ʔinfuá				*mp[ə]ʔ #105.a
168	*kaᵈruak	‘to knock’	*komdwâk-hata*	karuák				*dɔ[] k ‘to hammer’ #333.a
255	*ʔuŋ-lɔːŋ	‘neck’	*ong-lônga*	ʔuŋlóŋa	ʔuŋlɔˑŋə			*tluəŋ ‘throat’

Before proceeding to the segmental reconstruction in detail some more remarks are in order:

Quantity: Car is described as having a long-short contrast among syllable nuclei in the main-syllables of lexical roots, and this contrast is substantiated with minimal and sub-minimal pairs by Braine (1970). By contrast, Nancowry (especially as analyzed by Radhakrishnan) has several degrees of length in nuclei but these are not contrastive: nuclei of main-syllables of lexical roots are long and take stress, other syllables are short, and in words of three or more syllables there is a secondary stress which is intermediate between short and long. Consequently, it is necessary that Car is be relied upon to reconstruct proto-Nicobarese quantity distinctions. The comparative data presented here makes it clear that Car quantity distinctions broadly agree with etymological values elsewhere in Austroasiatic, so it is reasonable to treat these as inherited through proto-Nicobarese.

Timbre: The phonetic values represented by the sources (especially the older sources) are often clearly approximate and sometimes obviously unreliable, but there are reasons to have confidence in our interpretations of the better sources. Regarding Nancowry, Radhakrishnan’s work is a well-developed phonological analysis, based on primary data; additionally Rajasingh has worked with native speakers recently, noted substantial phonetic detail, has described extensively the phonetic values of the orthographic materials, presenting his data in a semi-phonetic transcription which is richer in detail that previous phonemic studies. In terms of the Car data, Braine (1963) presents a richly detailed phonetic transcription of

[13] The [ʔa] presyllable in Car reflexes appears to be a replacement of the etymological segment(s).

about 1700 words, and her 1970 thesis discusses the articulation of segments in a systematic fashion.[14] Some of these have been examined and so far they appear to confirm the phonetic values documented by Braine. As mentioned, the description of Car by Das (1977) is mixed in quality, and is utilized here as it witnesses in phonetic script numerous lexical items not recorded by Braine. However, Das is inconsistent in notation of segmental length and of vowel height.

The colonial era sources are extensive, but present only Roman-based orthographic renderings. Their spellings tend of omit glottal stops, divide words into syllables rather than morphemes (using dashes) and employ a plethora of diacritics which are often redundant in terms of what value they add. This being said, we have enough of an understanding of the actual phonology of the languages to allow one to make reasonable guesses as to what values are intended, and we have tried to do so with caution.

For the purposes working out the proto-phonology, a comparative vocabulary of 265 sets was assembled (see the Appendix). The comparative vocabulary utilizes three sources for Nancowry (Man 1889, Radhakrishnan 1981, Rajasingh 2016) and three sources for Car (Whitehead 1925, Das 1977, Braine 1961), plus corresponding items from Shorto's (2006) *Mon-Khmer Comparative Dictionary*.[15] A unified glosses column is used rather than glossing every item, with only divergent glosses included with individual items as appropriate. The criteria for inclusion are as follows:
Table 1 the same root is apparent in both Car and Nancowry, and/or
Table 2 a root is attested in at least Car or Nancowry and elsewhere in Austroasiatic

Phonological correspondences for root onsets, nuclei and codas were compiled, proto-values suggested, and summaries are presented below.

3.2 Proto-Nicobarese onsets

The following pNicobar onset segments can be reconstructed for main-syllables of lexical roots, and are assumed to represent the maximal set of available onsets segments:

Proto-Onsets

*/	p	t		c	k	ʔ
	m	n		ɲ	ŋ	
	ʋ	l, ᵈr,	ɹ	j		
	f	s			h	/

In the Austroasiatic context, the first remarkable feature to note is the lack of a voicing contrast among the obstruents, including a lack of implosive series. It is apparent that there were various mergers and changes in manner of articulation that played out somewhat differently at each place of articulation. The result of these changes was an inventory of onsets that resembles the proto-Austroasiatic set simply without the various voiced stops, but it was more complicated that simple loss of voicing. Note also that the proto-inventory offered here is different to the one offered by Sidwell in Sidwell & Rau (2015:263), which contrasted voiced and voiceless obstruents—on reflection this is better regarded as reflecting an intermediate pre-proto-Nicobarese stage. Below we examine the developments according to places of articulation.

A chain shift among the labial onsets seems to have proceeded as follows:

pAA		**pre-pN**		**pN**
*b, *ɓ	>	*b	>	*p
*p	>	*f	>	*f

[14] Additionally, there are audio recordings of Car reading and singing evangelical Christian materials, and books and pamphlets in orthography, available online at URLs such as: globalrecordings.net/en/language/6401, www.jw.org/caq/l%C4%ABp%C3%B6re/min%C3%AB-l%C4%ABp%C3%B6re/?start=12.

[15] Shorto reconstructions are retranscribed to IPA values, and subscript numbers are deleted from *t, *d, *n as not adding useful information.

Examples demonstrating the fate of Austroasiatic *ɓ have proven to be difficult to find in the data, but we have at least one among the data tabled below:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
*b,*ɓ > *p	127	*ʔupɯaʔ	‘to carry on back’		ʔupúaʔ				*ɓɔʔ #12
	43	*panam	‘place, village, country’				panam	*pa-nam*	*bnəm ‘hill’ #1369.b
	235	*puloːʔ	‘thigh’	*pulô*	puláʔ	pulɔːʔ			*blu:ʔ #223.a
	240	*pok	‘tie, bind’	*pôk-hata*	ʔukpók ‘tether’			*pōk* ‘string together’	*buək #357.c
*p > *f	143	*fəː	‘to blow’	*ifüa*	ʔifúa	fɯðː	fəː	*föö*	*cpiər #1638.a
	151	*fəŋ	‘crossbow’	*fòin*	fə́ŋ	fɔːɲ			*paŋʔ ‘to shoot’ #905.a
	32	*fa/ɔh	‘sweep’	*ifâh*	ʔifáh		fɔh	*fòh*	*tpəs #1916 ‘sweep’

The above pattern of shift, with lenition of *p > *f before devoicing occurred, is indicated since we need to avoid a feeding rule that would have seen all labial obstruents lenite to *f.

Among the dentals the patterning is strikingly different: the implosive remained distinct while the voiced and voiceless plosives merged:

pAA		**pre-pN**		**pN**
*d, *t	>	*d, *t	>	*t
*ɗ	>	*ɗ~^{d}r	>	*^{d}r

The proto-Nicobarese segment reconstructed *^{d}r is reflected variously in the sources:

Nanc. Man	Nanc. Radh.	Nanc. Raja.	Car. Das	Car. Braine	Car.Whit.
r~d	r	ɹ	r̃~r	r	r

Braine (1970:45) describes Car /r/ as a pre-stopped voiced alveolar flap (written *r̃* in the 1961 lexicon without further explanation). This segment regularly corresponds to the proto-Austroasiatic implosive *ɗ, and we can suggest that it lenited directly to a pre-stopped rhotic, rather than via a voiced stop *d, such that there was no feeding rule that would have shifted it to *t in the later devoicing phase. The alternating *r~d* transcription for this segment by Man also clearly hints that a pre-stopped rhotic is indicated for Nancowry.

Some examples:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
*d,*t > *t	175	*toah	‘breast’	*toah*				*teh*	*tɔh #1999.a
	34	*taj	‘hand’	*tai*	táj		ʔukti: ‘palm’	*tī*	*ti:ʔ #66.a
	216	*-tul	‘carry on head/shoulder’	*òl-tōl*				*ha-tul*	*du:l[] #1742.b
	123	*katəː/ɯa	‘stay, dwell’	*kâ-tö*	katɯ́	ʈɯː			*dəːʔ ‘stop, come to rest’ #78.a
*ɗ > *^{d}r	120	*^{d}rɛh	‘first’	*orēh*		ɹɐːx	ranɛh	*ra-neh*	*dĩːs ‘one’ #86.b
	125	*^{d}ruan	‘to perch’	*düan-hata*	rúan	ɹɯːən		*röön*	*ɗ(uː)n #1158
	150	*^{d}rəm	‘to hammer’	*dòm*	rə́m ~ rám			*röm*	*ɗaːm #1361a

The proto-Austroasiatic dorsals *k, *g apparently underwent a simple merger to *k in proto-Nicobarese; examples are tabled below, unfortunately Car reflexes of proto-Austroasiatic *g are scarce:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
*k,*g >*k	22	*kaːʔ	‘fish’	*kâa*	ká	kɑːʔ	kaːʔ	*kāᵏ*	*kaʔ #16
	28	*ʔakah	‘to know’	*akâh*	ʔakáh	ʔɑkɑːx	ʔakah	*a-ka-ha-lōn*	*dk[a]h ‘remember’ #1973.a
	266	*kiː	‘all’	*ki-hēang* ‘each’	kí ‘all’				*geː)ʔ ‘3Ppronoun’ #26
	155	*hakəp	‘to fit, fix’		hakə́p			*ha-köp* ‘hold’	*gap ‘fit, fitting’ #1240

The situation among the laminal onsets is more complex. While *ɟ devoiced to *c, it appears that proto-Austroasiatic *c generally shifted to *s, although there are at least two ambiguous items in the lexical comparisons (*caːt ‘to jump’, *ci(h) ‘who’) with other Austroasiatic forms suggesting an unchanged *c. This could be conditioned by a preceding stop, but more data is required to assess this properly. Also, good examples of *ɟ > *c reflexes in Car are still to be found, but the Nancowry reflexes are regular and can be taken as strongly indicative of the proto-Nicobarese values. Examples:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto#)
*ɟ > *c	51	*canaŋ	‘Clf. for trees, posts, hair etc.’	*chanang*					*ɟəŋ ‘to stand’ #538ii.a
	244	*cɔ(ː)ŋ	‘high’	*chòng*	cóŋ	cɔːŋ			*ɟ[o]ŋ ‘long, high’ #537.a
	166	*cuak	‘step’	*kochōak*	cuák				*ɟuək ‘tread’ #301.c
?*c > *c	20	*caːt	‘jump’	*chat*	cát	cɑːt ‘dance’	caːt lə ‘lift up’		*kcət #988.b
	85	*ci(h)	‘who?’	*chī*		ciː	ʔacih		*[ʔ]ciʔ #46
*c > *s	2	*samaː	‘jaw’	*shama-lâ-ēshe*			samãːʔ	*sa-mā* ‘jaw, chin’	*caʔ ‘to eat’ #8.a
	3	*saː	‘to eat’	*shâ-lare*	sã ‘edibles’				*caʔ ‘to eat’ #8.a
	38	*sak	‘to spear’		sák			*sak*	*cak ‘to prod, pierce’ #292.d
	77	*sian	‘cooked’	*ishīan-hata*					*ciən[] ‘cooked’ #1137.b
	107	*sej	‘lice’	*shēi*	séj	seːj ‘animal’			*ciːʔ #39.a

However, rather than proto-Austroasiatic *s merging with proto-Nicobarese *s, the latter merged with *h. Examples:

pAA >pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto#)
*s > *h	105	*ɹaheʔ	‘vein, nerve’	*ihē*	ʔihé		raheːʔ	*ra-hēᵏ*	*[] rsiiʔ #249.a
	106	*hej	‘fibre’	*heōe*	héj				*ks[i]ʔ #246.a
	19	*hiŋʔaːp	‘yawn’	*hing-âp*	hiŋáp				*sʔaːp #1229.a
*h > *h	35	*pahaj	‘sated’	*pa-hâe*				*u-ha-en*	*bhiːʔ #259.a
	111	*kahɛː	‘moon’	*kâhē*	kahé	kɑxɛ̃ːʊ			*khəjʔ #1542.a
	130	*huʊt	‘sniff’	*höt-hata*	hũt			*hūt* ‘to take soup’	*həʊt #1104.b

Glottal stop onsets are reconstructed, although they are not represented in the orthographies, they are robustly, if inconsistently, noted in the more recent sources. The assumption is that generally there are no zero onsets, consistent with the strong tendency for syllables to have at least CV structure. Examples:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto#)
*ʔ > *ʔ	26	*ʔac	‘feces’	*aiṅk, aiṅ(ch)*	ʔãc		ʔac ‘belly’	*aich*	*ʔ[ə]c #794.c
	29	*ʔah	‘live’	*âṅh*	ʔæ̃h ~ ʔãh	ʔãːx	ʔãh	*aṅ*	
	179	*ʔoal	‘in/inside’	*oal, òl*	ʔuál		ɛl	*el*	

There is a rhotic approximant *ɹ reconstructed for proto-Nicobarese (in addition to the pre-stopped *^{d}r, discussed above). Proto-Nicobarese *ɹ has regularly merged with the palatal approximant *j in Nancowry, while it is retained as a rhotic in Car. Braine (1961) writes this Car rhotic segment with under-dot (ṛ), and (1970:45) describes it as, “a voiced apico-dorsal fricative. The dominant phonetic quality of this phoneme is the retroflexion.” This suggests Car [ɽ] although the original typescript *r* notation is retained here. The reconstruction as an approximant in proto-Nicobarese is indicated by the merger wth *j in Nancowry. Austroasiatic onset *j appears to be unchanged; although I have not found good examples reflecting this in Car there are multiples examples of Car [j] corresponding to Nancowry [j]. Examples demonstrating these developments are tabled here:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
*ɾ > *ɹ	64	*ɹat	‘cut (with knife)’	*yât*	ját		ṛat	*ṛat*	*rat ‘reap’ #1058.e
	174	*ɹoac	‘overflow’	*yuait-nga*	juácŋa		taṛeːci		*ruəc ‘fall, drip’ #843.b
	187	*ɹuːj	‘fly (n.)’	*yūe*	júaj		inṛuːj		*ruəj #1504.c
*j > *j	17	*jaːŋ	‘to hear’	*yâng*	jáŋ	jɑːŋ			*kj[ə]ŋ #649.a
	124	*ʔijɯam	‘breathe’	*eyām*	ʔijɯ́am	ʔijɯəːm			*jhu(ə)m #1299.a
	228	*hVjoːj	‘drunk’	*huyòie*	hujój		hijoːj	*hi-yōi*	
	128	*sajɯh	‘year’	*shaiyūh*	sajɯ́h	sɑjɯːx		*sam-yeu-hö*	

The remaining onsets—nasals and approximants—effectively continue unchanged from proto-Austroasiatic without significant restructuring.

3.3 Proto-Nicobarese codas

The proto-Nicobarese coda inventory is effectively the same as the proto-Austroasiatic inventory and the typologically ‘normal’ inventory for conservative Austroasiatic languages:

Codas

*/	p	t	c	k	ʔ	ø
	m	n	ɲ	ŋ		
	ʋ	l	j			
		s		h		/

The correspondences between Nancowry and Car codas show little significant change once notational differences are taken into account. The colonial era sources, and Das, consistently write an *i* before palatal stop and nasal codas, but this a non-contrastive transition rather than an independent segment. Glottal stops are not reliably indicated in older sources: Man does not write them at all, and Whitehead indicates them much of the time—although not always—with an italic *k*, re-transcribed with a superscript ᵏ here. Das is inconsistent in his notation of glottals, and Rajasingh appears to hypercorrect, occasionally noting glottals where they are not expected on the basis of other sources.

The fricative codas broadly show a general merger of proto-Austroasiatic *s, *h > *h, although in the light of Rajasingh’s consistent transcription of this segment in Nancowry as [x] we might reconsider the proto-Nicobarese value. Proto-Nicobarese coda *s reemerged in new coinages. Examples:

pAA > pN	No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)		Car (Braine)	Car (Whit.)	pAA (Shorto#)
	64	*ɹat	'cut (with knife)'	*yât*	ját			ṛat	*ṛat*	*rat 'reap' #1058.e
	176	*moah	'nose'	*moah*	muáh	muɑˑx	mɛh	ʔɛlmeh	*el-meh*	*muh #2045.a
	30	*ʔah	'swell'		jə́h			aha	*aṅ-haṅ*	*ʔəs #1871.b
*h,s > *h	32	*fa/ɔh	'sweep'	*ifâh*	ʔifáh		fɔh	fɔh	*fòh*	*tpəs #1916 'sweep'
	72	*ɹiah	'root'	*yiah*		jiəˑx				*riəs root #1927.b
	137	*ɹəːh	'dry'	*heash*	hijə́h ~ hijə́s	xɛˑəx				*srəs 'dry' #160.b
new *s	204	*fuːs	'smoke'	*fush*	fús	fuˑs	fanuːsa 'steam'	hanũːsŋə 'steam'	*ha-nus-ngö* 'vapour'	

3.4 Proto-Nicobarese nuclei

The proto-Nicobarese main-syllable nuclei inventory is typologically unremarkable for a conservative Austroasiatic language; three degrees of height, backness, contrastive length, and several falling diphthongs, are reconstructed as follows:

Nuclei – stressed syllables

*/	iː, i		uː, u	ia	ɯa	ua
	e(ː)	əː, ə	oː, o	ea		oa
	ɛ(ː)	aː, a	ɔː, ɔ			

/

The range of nuclei in main syllables contrasts with the dearth of contrasts in unstressed syllables: both Car and Nancowry only contrast three or four nuclei in non-primary-stress positions, with considerable allophony. Based on the forms found in the comparative vocabulary, the following set of nuclei is reconstructed:

Nuclei – unstressed syllables

*/	i [ɪ]	u [ʊ]	
	a [ɐ]		/

The timbre of unstressed vocoids is often determined morphologically; when they occur in affixes the timbre is important, while those in pre-syllables created in the restructuring of earlier onset clusters the timbre appears to be conditioned by adjacent segments, but this requires a larger data compilation to investigate fully.

Returning to main-syllables, the monophthong correspondences are relatively straightforward; the most complex and interesting aspect of the vocalism is the treatment of diphthongs. Previous commentary on the issue (e.g. Diffloth 2008, Sidwell 2015) has suggested that the five Nancowry diphthongs: [iˑə, iaˑ, uˑə, uaˑ, ɯˑə] directly continue five proto-Nicobarese diphthongs. The analysis presented in this paper still reconstructs five proto-Nicobarese diphthongs, but the problem is revealed to be more complex than previously presented.

The most remarkable correspondences emerging when comparing Car and Nancowry is the occurrence of a Car front vowel [ɛː ~ ɛ ~ eː ~ e] in etyma which Radhakrishnan consistently records with Nancowry /uá/ ([oa]), Man mostly writes with *oa* (with or without various diacritics) and Rajasingh writes [uɑˑ]. Compare:

gloss	Nanc. Man	Nanc. Rhad.	Nanc. Raja.	Car. Das	Car. Braine	Car. Whit.	pAA (Shorto #)
'overflow'	*yuait-nga*	juácŋa		tareːci	taɾeːci		*ruəc 'fall, drip' #843.b
'nose'	*moah*	muáh	muɑˑx	mɛh	ʔɛlmeh	*el-meh*	*muh #2045.a
'breast'[16]	*toah*			tɛh		*teh*	*tɔh #1999.a
'to cough'	*oōàh*	ʔuʔuáh		ʔɛhɛ	ʔɛhɛ		
'arm'[17]	*koâl*	kuál	kuɑˑl	kɛl	kɛːl	*këll*	*[ɻ]guːl 'finger' #1717
'in, inside'	*oal, òl*	ʔuál		ʔɛl	ɛl	*el*	
'four'	*fõan*		fuɑˑn	feːn	fɛːn	*fën*	*puən #1166.b

Note also the footnotes in the Appendix showing that Great Nicobar (GN), Teressa-Bomboka (TB), and Chowra (Ch) evidently have reflexes similar to Nancowry, pointing to the marked character of the Car reflexes. Additionally, the external comparisons indicated with reference to Shorto (2006) consistently indicate that a back monophthong and/or diphthong reflexes are found in cognates farther afield, confirming that Car is the most innovating language in respect of this correspondence.

The fact that very solid Austroasiatic etymologies 'nose', 'breast' and 'four' are represented in this correspondence testifies to its reality, and gives us important clues to interpreting how we come to have a low fronted reflex in Car. The external comparisons, and the Nicobarese comparisons, clearly show that the front vowel reflexes are restricted to Car, and logically a distinct proto-Nicobarese back vowel (monophthong or diphthong), which de-rounded in Car, needs to be reconstructed for this correspondence.

The above diphthong correspondence also apparently has a parallel with a front diphthong in Nancowry and a low back vowel in Car, although only one example has been found so far:[18]

gloss	Nanc. Man	Nanc. Rhad.	Nanc. Raja.	Car. Das	Car. Braine	Car. Whit.	pAA (Shorto #)
'elbow'	*det-ongkēang*	rétʔuŋkiáŋ		sikɔŋ	sikɔŋ	*si-kòng*	*kiəŋ[] #891.b

Whether or not the Nancowry /ua/ is synchronically one contrastive unit, there is evidence that it has multiple historical origins. If we examine the apparent cognates with Car back vowels, a clear split emerges. The broad pattern is as follows:

Nanc. Man	Nanc. Radh.	Nanc. Raja.	Car.Das	Car.Braine	Car.Whit.
oa~ua	uá	uɑˑ	ɔ(ː)	ɔː	*o*
oa~ua	úa~ú	uˑə	u(ː)	uː	*ūö*

Man's Nancowry transcriptions are ambivalent, while those of Radhakrishnan and Rajasingh more consistently show split reflexes corresponding to both low and high long back monophthongs in Car. Radhakrishnan occasionally vacillates between úa ~ ú for items corresponding to Car /uː/ (e.g. 'twist' ʔúaŋ ~ ʔú) or also records ú forms (e.g. cúk 'place', fús 'smoke'). The Car sources are consistent in indicating a high back monophthong (disregarding Das' inconsistent length notation).

How are we to interpret these data? Superficially it seems that we have several distinct correspondences involving diphthongs of some sort, but it is not clear that these were all diphthongs in proto-Nicobarese, so it is necessary to include the other back vowel correspondences to see the wider context. When this is done, a clear pattern emerges: Nancowry /úa/ corresponds consistently to Car /uː/ while Nancowry /uá/ has split

[16] GN toâh, TB tòh, Ch tòh.

[17] GN koâl, TB kôr.

[18] Note the different word-formational strategies in these forms. The *det/rét* in Nancowry is historically the word of 'tail' and is attached to body parts at extremities or articulations, while the *ʔuŋ-* with an inflectional prefix with copy-infixation of the dorsal nasal coda. Literally the Nancowry form can be read as a nainalization which is conceotually the "bending extremity". The initial *si* syllable of the Car form is not identified as a prefix or separate morpheme in any of the literature, in fact the apparent Pearic cognates (e.g. Chong *cʰkɛːŋ* 'elbow') clearly suggest that *si* is etymologically part of the stem, indicating proto-Nicobarese *cikeaŋ.

reflexes between Car /ɔː/ and /ɛː, ɛ, eː, e/. Clearly Nancowry /uá/ is innovative, coming from proto-Nicobarese *uː, while /uá/ actually reflects the merger of two older proto-vowels, provisionally reconstructed *oa and *ua. Simplifying the presentation to just the Radhakrishnan and Braine forms to make it easier to process visually, the pattern is follows:

Nanc.	**Car**	**pN**	**Examples in Appendix**
uá	ɛ/eː	*oa	171, 174-181
uá	ɔː	*ua	158-170, 172-173
úa~ú	uː	*uː	182-208
ú~ó	u	*u	209-225
ó	oː	*oː	226-235
ó	o	*o	236-243
ó	ɔː	*ɔː	244-256
ɔ́~ó	ɔ	*ɔ	257-265

We would predict this result to be paralleled among the front vowels, but regrettably only one relevant comparison—noted above—has been found at this stage, suggesting proto-Nicobarese *ea. Assuming that this comparison is valid, we can summarize the front vowel correspondences and reconstructions as follows (again restricting data to just Radhakrishnan and Braine forms):

Nancowry	**Car**	**pN**	**Examples in Appendix**
iá	ɔ	*ea	81
iá	eː	*ia	71-80, 82-84, 103
ía	iː	*iː	86-88
í	iː	*I	85, 90-102
é~ɛ́	e	*e(ː)	104, 106-110
é~ɛ́	ɛ(ː)~e(ː)	*ɛ(ː)	111-122

The main ambiguity lies with length values for the [-high] front vowels – reflexes are so few that it is difficult to determine clearly whether length is contrastive, though it would be expected on general grounds. The data compiled suggests the following pattern:

1) [ɛ] with /h/ codas and one example with –j (no other closed syllables)
2) [ɛː] elsewhere.
3) [e] in closed syllables
4) [eː] only one example with -ʔ

Coda [h] tends to shorten nuclei anyway in Nicobarese, and coda glides and glottal stop tend to lengthen preceding nuclei, so we really need more examples to determine if there are specific syllable types in which length is important for these. For the moment proto-forms are marked long or short according to the phonetic values indicated by the sources.

There appear to be fewer central proto-vowels, but also relatively few comparisons, so the reconstructions are somewhat under-determined. There is no Nancowry /ɯá/, so there is no parallel issue to the /uá, iá/ problem, which simplifies that aspect of the problem. However, since we have reconstructed /úa/ and /ía/ as coming from *uː and *iː respectively, we would expect /ɯ́a/ by analogy to come from *ɯː, yet the evidence for *ɯː is weak. Nancowry /ɯ́a/ regularly corresponds to Car /əː/, and it is not immediately clear if these reflect a proto-Nicobarese monophthong or diphthong. I have found only one example of Nancowry /ɯ́/ (katɯ́ ‘stay, dwell’) with a possible Austroasiatic etymology (cf. Shorto #78.a *dəːʔ ‘stop, brought up short, come to rest’) but Car cognates are missing, and the same word is written ***kâ-tö*** by Man, suggesting /əː/ or /ɯː/ as the correct value and *ɯa, and the ‘stay, dwell’ etymon is reconstructed ambiguously as **katəː/ɯa*. Consequently, there is little to suggest a contrast between *ɯː and *ɯa at the proto-level, and provisionally only *ɯa is reconstructed but more evidence could compel *ɯː or potentially a contrast between *ɯː and *ɯa.

At the same time there is plenty of evidence of robust contrasts between *əː and *ə, and *aː and *a, at the proto-Nicobarese level. Braine treats Car /ɤ/ and /ə/ (long and short) as only marginally in contrast, and likely to represent a historical split. This seems to be the case as both are found corresponding to Nancowry /ə́/. Nancowry also shows some variation, with etyma that have clear antecedents with *ə showing both or either ə́~á (e.g. cák ~ cə́k 'pain' cf. Car cɤk), and etyma that have clear antecedents with *a showing both or either á~ǽ (e.g. ʔæ̃h ~ ʔãh 'live' cf. Car ʔãh).

The preceding discussion is summarized in the following (again restricting data to just Radhakrishnan and Braine forms):

Nanc.	**Car**	**pN**	**Examples in Appendix**
ɯ́a	əː	*ɯa	123-127
ə́	ɤː~əː	*əː	131-137
ə́~á	ɤ~ə	*ə	138-157
á	aː	*aː	1-22
á~ǽ	a	*a	23-70

There remain some residual issues concerning the proto-nuclei. Nasalization is a feature synchronically in both Car and Nancowry, and on that basis it would be logical to reconstruct this a feature at the proto-Nicobarese level. However, contrastive nasalization is not a typical feature of Mainland Austroasiatic languages, so it is likely secondary within Nicobarese. Additionally, paying regard to the comparative data assembled here, two factors are evident: 1) the languages often disagree in nasalization (e.g. Car *ʔac* 'belly', Nancowry *ʔãc* 'feces'; Car *ɲɔ̃ːk* 'suck', Nancowry *ɲuák* 'breathe', etc.), and 2) nasalization is clearly associated with strongly nasalizing environments, such as adjacent nasal or glottal segments (so called rhinoglottophilia, see Matisoff 1975, Sprigg 1987). Consequently, it does not appear that contrastive nasalization can be coherently reconstructed for proto-Nicobarese, and proto-forms are unmarked for nasalization.

One will also notice some tabled comparisons involve one-off correspondences that have not been discussed here, yet provisional reconstructions are offered. In these cases, no reconstructed segments rely solely on these items, and the proto forms have been posited as best estimates based on the totality of the available facts.

4 Conclusion

The results of the present paper confirm some points that have been made in previous studies, but also represent new solutions that contradict some previous claims. It is clear that the proto-Nicobarese syllable shapes and segmental values are readily derived from proto-Austroasiatic, however a modest number of changes that, combined with a general loss of voicing contrast, amount to a phonological restructuring away from a typical Mainland Austroasiatic profile. Nonetheless, the character of the language is revealed to be strongly Austroasiatic.

Of particular interest is the question of proto-Nicobarese diphthongs, which potentially has impact on the wider question of the history of Austroasiatic vocalism. At the 18th Southeast Asian Linguistics Society meeting (Bangi Malaysia, May 22nd 2008) Gerard Diffloth presented a handout listing seven lexical comparisons indicating correspondences between the Nancowry /uə, oa/ and pAslian *uə, *uɔ respectively (equivalent to Phillips (2012) pAslian *uə, *ua). These seem to lend external support to the reconstruction of two levels of diphthongs in proto-Nicobarese and a hypothetical proto-NicoAslian. However, the present bottom-up reconstruction of proto-Nicobarese vocalism presents a very different picture. It is argued here that Nancowry /uə/ is a reflex of *uː, while Nancowry /oa/ represents a merger of proto-Nicobarese *ua and *oa (and analogically Nancowry /ea/ < *ia and *ea). There were two levels of diphthong in proto-Nicobarese, but these do not neatly equate to cognate Aslian diphthongs, nor apparently to the distribution of proto-Austroasiatic diphthongs reconstructed by Shorto, and the matter requires a thorough reappraisal by extending the set of Nicobarese etymologies and careful comparison to Aslian.

Looking at the external comparisons, it is apparent proto-Nicobarese *oa and *ea have wider cognates with a range of monophthong and diphthong values, as do proto-Nicobarese *ua and *ia, and it is not at all clear that the two levels of diphthongs are an ancient feature. On general grounds, it would seem likely that there were various vocalic splits and mergers in the early history of Nicobarese at the same time that syllable

structure was simplifying and words were restructuring to take longer strings of syllables. In that context, one can suggest that the kind of prosodic alternation in the phonetics of diphthongs noted for modern Nancowry was also in effect and changing the timbre of vowel sequences in pre-proto-Nicobarese. As already noted, a much better comparative data set would be needed to demonstrate this.

This paper has focused on the phonology of the main-syllables of proto-Nicobarese lexical roots. These are found to be simple CV(C) syllables which utilized the following segments:

Proto-Onsets:

*/	p	t		c	k	ʔ
	m	n		ɲ	ŋ	
	ʋ	l, ᵈr,	ɹ	j		
	f	s			h	
	/					

Proto-Codas:

p	t	c	k	ʔ	ø
m	n	ɲ	ŋ		
ʋ	l	j			
	s			h	

Proto-Nuclei:

*/	iː, i	uː, u	ia	ɯa	ua
	e(ː)	əː, ə	oː, o	ea	oa
	ɛ(ː)	aː, a	ɔː, ɔ		

/

A reduced inventory of segments was evidently utilized on other syllables, but a significantly richer set of comparisons will be necessary to adequately work out the full details (segments and restrictions) of these other syllables (pre-syllables of lexical roots and affixes). This may be inherently difficult to achieve; even if it is possible to obtain more extensive lexical data from other Nicobarese lects, the combined of effects of word tabooing and morphological restructuring may have gravely reduced the absolute amount of Austroasiatic vocabulary retained on the Nicobars. It is clear from examining the Car and Nancowry lexicons as represented in the available dictionaries that both languages only utilize a modest set of roots which are combined with affixes to achieve a high level of lexical productivity. This effect is particularly evident in the etymological vocabulary of Radhakrishnan (1981:84-158). That lexicon extracts some 778 numbered roots, very many of which lack apparent Austroasiatic etymologies.

The Appendix to this paper sets out the evidence on which the present results are based. The comparisons are numbered and are ordered into groups according to the proto-rimes. The abbreviations GN, TB, Ch in the footnotes refer to Great Nicobar ("Coastal Inhabitants"), Teressa and Bampoka, and Chowra forms provided in the Appendix C in Man (1889).

References:

Blench, Roger and Paul Sidwell. 2011. Is Shom Pen a Distinct Branch of Austroasiatic? In *Austroasiatic Studies: papers from the ICAAL4: Mon-Khmer Studies Journal Special Issue No. 3, ed. by* Sophana Srichampa, Paul Sidwell, and Kenneth Gregerson, 90–101. Dallas, SIL International; Canberra, Pacific Linguistics; Salaya, Mahidol University.

Braine, Jean Critchfield. 1970. *Nicobarese Grammar (Car Dialect)*. PhD dissertation, University of California, Berkeley.

Braine, Jean Critchfield. 1963. Ms. *Car Nicobarese Vocabulary* (87 pages typescript).

Braine, Jean Critchfield. 1976. Numeration in Car Nicobarese. *Linguistics* 174:21–30.

Chattopadhyay, Subhash Chandra, and Asok Kumar Mukhopadhyay. 2003. *The language of the Shompen of Great Nicobar: a preliminary appraisal*. Kolkata, Anthropological Survey of India.

Das, A.R. 1977. *A Study of the Nicobarese Language*. Calcutta: Superintendent of Government Press.

De Röepstroff, Frederik. 1875. *Vocabulary of Dialects Spoken in the Nicobar and Andaman Isles*. (2nd edition) Calcutta: Superintendent of Government Press.

De Röepstroff, Frederik. 1884. *Dictionary of the Nancowry Dialect of the Nicobarese Language, in Two Parts: Nicobarese—English, and English—Nicobarese*. Ed. by Mrs. De Röepstrofff. Calcutta: Superintendent of Government Press.

Elangaiyan, Rathinasabapathy et. al., 1995. *Shompen–Hindi Bilingual Primer Śompen Bhāratī 1.* Port Blair and Mysore.

Jenny, Mathias. 2015. Syntactic diversity and change in Austroasiatic languages. In *Perspectives on historical syntax*, ed. by Carlotta Viti, 317–340. Amsterdam: John Benjamins.

Man, E.H. 1889. *A Dictionary of the Central Nicobarese Language.* Reprint: Delhi: Sanskaran Prakashak, 1975.

Matisoff, James, A. 1975. "Rhinoglottophilia: The Mysterious Connection between Nasality and Glottality." In *Nasálfest: Papers from a Symposium on Nasals and Nasalization, Universals Language Project*, ed. by C.A. Ferguson, L.M. Hyman, and J.J. Ohala, 265–287. Stanford University, Stanford.

Murthy, R. V. R. 2005. *Andaman and Nicobar Islands Development and Decentralization*. New Delhi, Mittal Publications.

Norman Zide and D.P.S. Dwarikesh. 1962 (ms.) The Comparative Phonology of Proto-Nicobarese from Kar Nicobarese and Central Nicobarese: preliminary version. (copy provided by Zide).

Radhakrishnan, R. 1970. *A preliminary descriptive analysis of Nancowry*. PhD dissertation. Department of Linguistics, University of Chicago.

Radhakrishnan, R. 1981. *The Nancowry word: phonology, affixual morphology and roots of a Nicobarese language*. Current inquiry into language and linguistics, 37, Edmonton, Alberta: Linguistic Research Inc.

Rajasingh V. R. 2016. Mūöt (Nicobarese). *Mon–Khmer Studies* 45:14–52

Rizvi, S. N. H. 1990. *The Shompen: a vanishing tribe of the Great Nicobar Island*. Calcutta: Seagull Books (on behalf of the Anthropological Survey of India).

Shorto, Harry L. 2006. *A Mon-Khmer Comparative Dictionary.* Canberra: Pacific Linguistics.

Sidwell, Paul and Felix Rau. 2015. Austroasiatic Comparative-Historical Reconstruction: an overview. In *The handbook of Austroasiatic languages*, ed. by Mathias Jenny and Paul Sidwell, 221–362. Leiden: Brill.

Sidwell, Paul. 2015. *Austroasiatic Classification*. In Mathias Jenny & Paul Sidwell (eds.) The handbook of Austroasiatic languages. Leiden, Boston: Brill. pp. 144–220.

Sidwell, Paul. 2015. Car Nicobarese. In *The handbook of Austroasiatic languages*, ed. by Mathias Jenny and Paul Sidwell, 221–362. Leiden: Brill.

In *The handbook of Austroasiatic languages*, ed. by Mathias Jenny and Paul Sidwell, 1229-265. Leiden: Brill.

Sprigg, R. K. 1987. "Rhinoglottophilia" re-visited : observations on "the mysterious connection between nasality and glottality", *Linguistics of the Tibeto-Burman Area* 10(1):44–62 (1 p. 1/2)

Temple, R.C. 1902. *A Grammar of the Nicobarese Language*. Chapter IV, Part II, The Census Report on the Andaman and Nicobar Islands. Port Blair: Superintendent's Press

Whitehead, George. 1925. *Dictionary of the Car-Nicobarese language.* Rangoon: American Baptist Mission Press.

Wurm, S.A. and S. Hattori (eds.) 1981, 1983. *Language atlas of the Pacific area*. (Pacific Linguistics C-66, C-67). Canberra: Australian Academy of the Humanities in collaboration with the Japan Academy.

Appendix: Nicobarese comparisons and reconstructions

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
1	*ɹaː	‘abandon’	*yâ-she*	jáʔ		ra ɲə	ɾaːŋə	*ɾ̣ā-ngö* ‘leave, renounce’	*raʔ ‘fall, be shed’ #2051.q
2	*samaː	‘jaw’[1]	*shama-lâ-ēshe*			samaː	samãːʔ	*sa-mā ‘jaw, chin’*	*caʔ #8.a
3	*saː	‘to eat’	*shâ-lare*	sã ‘edibles’					*caʔ #8.a
4	*ʔaː	‘two’[2]	*âṅ*	ʔã	ʔãː				*ʔaːr #1562
5	*kaᵈraːc	‘knead’	*kendech-hata*	karéc ~ karɛ́c		kiraːcə		*ki-rāich*	
6	*ŋaːc	‘oil’	*ngai(ch), ngai(j)*	ŋác					*[ ]ŋaic #805a.a
7	*faːj	‘cloud’	*mifaiṅya*	mifãja	mifãːjə				*paːj #1479.a
8	*litaːk	‘tongue’[3]	*kale-tâk*	kaliták	kɑliʇɑːk	litak	litaːk	*li-tāk*	*l(n)taːk #320.a
9	*ᵈraːk	‘water’[4]	*dâk, râk*	riák	ɹeɑːk	(mak)	(mak)	*(mak)*	*ɗaːk #274
10	*ŋaːl	‘ripe’	*īñgol*			haŋaːw	təŋãːv	*ngāṅv*	
11	*mahaːm	‘blood’	*(wâ)*	(wá)	(ʊɑː)	mahaːm	mahaːm	*ma-hām*	*ɟhaːm #1460
12	*kaːn	‘female, wife’	*kân*	kán	kɑːn	kan	kaːn	*kan*	*kan #1126.a
13	*taːɲ	‘weave’	*en-tain-ya*	ʔitáɲ					*taːɲ #898.a
14	*kalaːŋ	‘eagle’	*kalâng*	kaláŋ ‘vulture’					*laːŋ ‘large raptor’ #714.b
15	*ʔoal-faːŋ	‘mouth’[5]	*oal-fâng*	ʔuálfáŋ	ʔɔɑːl fɑːŋ	ʔɛlwaŋ	ʔɛlvaːŋ	*el-vāng*	*paːŋ #605.a
16	*ʔaːŋ	‘open’	*âng*	ʔáŋ					*ʔaːŋ #1229.a
17	*jaːŋ	‘to hear’	*yâng*	jáŋ	jɑːŋ				*kj[ə]ŋ #649.a
18	*haːŋ	‘to hear’[6]				haŋ	haŋ	*hang*	
19	*hiŋʔaːp	‘yawn’[7]	*hing-âp*	hiŋáp					*sʔaːp #1229.a
20	*caːt	‘jump’	*chat*	cát	cɑːt ‘dance’	cat lə ‘lift up’	caːt lə ‘lift up’		*kcət #988.b
21	*ʔuɹaːʊ	‘coconut’	*oyàu*	ʔujáw					*braːw ‘coconut palm’ #1852.a

1 TB *shaːma*, Ch *shaːmâ*.
2 GN, TB, Ch *âṅ*.
3 GN *kale-tâk*, TB *kali-tâk*, Ch *kalitâk*.
4 GN dâk, TB râk, Ch râk.
5 TB *a-fâng*, Ch *oal-fâng*.
6 GN *hâng*, TB *heäng*, Ch *hēːang*.
7 GN *angâp*, TB *hing-âp*, Ch *hing-âp*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
22	*kaːʔ	‘fish’	*kâa*	ká	kɑːʔ	kaː	kaːʔ	*kāᵏ*	*kaʔ #16
23	*taʊ (?)	‘wasp’	*tâo*				kilamtao ‘gnat’	*ki-lam-ta-ō ‘bee’*	
24	*ʊac	‘miss target’	*wait*	wác		wac		*vaich ‘forget’*	*[r]wəc ‘inattentive, forget’ #1094.c
25	*ɪac	‘wash’	*et-yait*	ʔitjác					*rac ‘sprinkle, scatter’ #837.a
26	*ʔac	‘feces’	*aiṅk, aiṅ(ch)*	ʔãc		ʔac ‘belly’	ʔac ‘belly’	*aich*	*ʔ[ə]c #794.c
27	*kapah	‘die’[8]	*kâpâh, kapâh*	kapáh		kapah	kapah	*ka-pah*	
28	*ʔakah	‘to know’	*akâh*	ʔakáh	ʔɑkɑːx	ʔakah	ʔakah	*a-ka-ha-lōn*	*dk[a]h ‘remember’ #1973.a
29	*ʔah	‘live’[9]	*âṅh*	ʔæ̃h ~ ʔãh	ʔãːx	ʔãh	ʔãh	*aṅ*	
30	*ʔah	‘swell’		jə́h			aha	*aṅ-haṅ*	*ʔəs #1871.b
31	*fana/ɔh	‘broom’	*hannâh-oal-ñī*	fanáhʔuálɲi		fanɔh	fanɔh	*fa-nòh*	*tpəs #1916 ‘sweep’
32	*fa/ɔh	‘sweep’	*ifâh*	ʔifáh		fɔh	fɔh	*fòh*	*tpəs #1916 ‘sweep’
33	*tanaj	‘five’[10]	*tanai*	tanáj	ʈɑnɑːj	tanɯi	tanij	*ta-neui*	
34	*taj	‘hand’[11]	*tai*	táj		ʔukti ‘back of hand’	ʔuktiː ‘palm’	*tī*	*tiːʔ #66.a
35	*pahaj	‘sated’	*pa-hâe*					*u-ha-en*	*bhiːʔ #259.a
36	*ᵈrak	‘break, split’	*dà(k)*	rə̃k ~ rãk	ɹɑːk				*ɗak #331
37	tak	‘to drip’	*patâk-shu* ‘drop (fruit)’	ták					*[k]tək ‘drip, drop’ #314.b
38	*sak	‘to spear’		sák		sak		*sak*	*cak ‘to prod, pierce’’ #292.d
39	*kaʊal	‘throw away’[12]	*ka-wâl*	kawál	kɑʊɑːl			*ka-val* ‘cast’	
40	*ʔam	‘dog’	*âm*	ʔám	ʔɑːm	ʔam	ʔam	*am*	
41	*hilam	‘leech’	*helam*						*tlam #1104.b
42	*ᵈram	‘night’	*râm*	rám ~ rə́m					*ɗəm #1360.c
43	*panam	‘place, village, country’				panam	panam	*pa-nam*	*bnəm ‘hill’ #1369.b
44	*kɪam	‘rim, edge’	*kēam*	kiám					*riəm/*rəm ‘edge, rim’ #1383.c/d

8 GN *kâpâh*, TB *kâ-päh*.

9 GN *hari-âṅh*, TB *âṅh*, Ch *eṅh*. Cf. àh Kammu-Yuan, **ʔəh* ‘to stay, to be’ proto Vietic.

10 GN, TB *tanī*.

11 TB *mòh-tī*.

12 *kʰwalˡ* ‘throw stone’ Riang Sak (Luc1964:C:RS-1403)

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
45	*katam	'roe'	*katam-kâa*						*ktəm 'egg' #1348.a
46	*tam	'to spear'[13]	*omtâm-hata*			tam	tam	*tam*	*tam 'hit repeatedly' #1340.d
47	*ʊan	'coil'	*en-wan-hala*						*wan #1208.e
48	*laŋan	'heavy'[14]		laŋãn	lɑŋã˙n	laŋan	laŋan	*la-ngan*	
49	*tulan	'python'[15]	*tulân*	tulán		tulan	tulan	*tu-lan*	*tla:n #1205.b
50	*taŋ	'hot'	*taiñ*	táŋ	ʈɑ˙ŋ	jal taŋa 'gluttony'	taŋ 'angry'	*tainy* 'savage'	*taŋ #897
51	*canaŋ	'Clf. for trees, posts, hair etc.'	*chanang*						*ɟəŋ 'to stand' #538ii.a
52	*faŋ	'cut'	*fânga* 'cut sticks'	fáŋa 'that which is cut'		faŋ 'sword'	faŋ 'cut with big knife'	*fang* 'to cut, lop off'	--
53	*naŋ	'ear'[16]	*nâng*	náŋ ~ nəŋ	nɑ˙ŋ	naŋ	naŋ	*nang*	*ktaŋ 'to hear' #555.b
54	*pintaŋ	'gall'	*pin-tang*	pintáŋ					*ktaŋ 'bitter' #554.a
55	*haŋ	'hot, spicy'		háŋ				*hang* 'to smart, hot (chillies)'	*haŋ #783.a
56	*hʊaŋ	'perspiration'[17]	*hoâng*					*vang*	
57	*taŋ	'to fence'	*ka-tâng*	katáŋ		tanaŋ 'wall'		*ta-nang-tö* 'curtain, screen, palings'	*bdaŋ 'walling material' #580.a
58	*sap	'answer'	*op-shâp*	sáp		sap		*sap-ṛô*	
59	*kap	'bite'	*kâpa*		kɑ˙p	kap	kap	*kap*	*kap #1231.b
60	*tap	'parasite'	*tamâp* 'tick'	muptáp 'flea'		tamap 'leech'			*[k]t[ə]p 'cockroach, vermin' #1252.a
61	cap	'pick up'	*op-châp*	ʔupcáp	ʔupcɑ˙p	cap lə			
62	*kanap	'tooth'[18]	*kanâp*	kanə́p	kɑnɑ˙p	kanap	kanap	*ka-nap*	*kap 'to bite' #1231.b
63	*kap	'turtle'[19]	*kâp*	káp		kap	kap	*kap*	*ka:p #1235.c
64	*ɹat	'cut (with knife)'	*yât*	ját		rat	ṛat	*ṛat*	*rat 'reap' #1058.e
65	*mat	'eyeball'	*oalmât, oalmat*	ʔuálmát	ʔɔɑ˙l mɑt	mat	mat	*mat*	*mat #1045

13 TB *umtom*.
14 Khmer *tŋʊən*, Vietnamese *nặng*.
15 Ch *tulân*.
16 GN *nâng*, TB *a-nang*, Ch *nâng*.
17 GN *henâng*, TB *hō*, Ch *hoâ˙ng* .
18 GN, TB, Ch *kanâp*.
19 GN, TB, Ch *kâp*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
66	*sat	'seven'[20]	*issât*		ʔisɑ·t	sat		*sat*	
67	*haɪat	'sour'	*haiyöt*						*sra:t #1074.A
68	*ciɪaʊ	'deep'	*chiyàu*	cijáw		ʔaru:	ʔaɻu:ʔ	*a-ɻū*	*ɟru:ʔ #172.a
69	*pacaʊ	'sour'	*pachau*				--		*ɟuʔ #50.a
70	*-ʋaʊ	'vomit'	*oàu*	ʔuʔʔów	xuʔo·ʊ	kuwao	kuvav	*kū-ö-vö*	*cʔ[au]ʔ #11
71	*ɲia	'child'		nĩa 'smaller'		ɲi:ə	ɲĩ:ʔ	*nyĩö*	*ɲ[e:]ʔ 'small' #59.a
72	*ɹiah	'root'	*yiah*		jiə·x				*riəs root #1927.b
73	*ʔitiak	'sleep'	*iteak*	ʔitiák	ʔiʇeɑ·k				*tiək #305.b
74	*kial	'brinjal'	*kēal*	kíal					*[t]kiəl 'cucumber' #1710.a
75	*ʋial	'turn'	*wīal*		ʋiə·l	cuwil 'spin'		*chu-vī* 'round'	*wiəl(ʔ) #1794.c
76	*kanial	'tusk'	*kaneäl*	kaniãl	kɑneã·l	kanel			(Katu kial 'to bite off')
77	*sian	'cooked'	*ishīan-hata* 'cooked ready for eating'						*ciən[ ] cooked #1137.b
78	*hiaŋ	'one'[21]	*hēang*	hĩaŋ	xi·əŋ	heŋ	heŋ	*hĕng*	
79	*ʋiaŋ	'stomach'[22]	*wīang*	wíaŋ					*rwiə[ŋ] #776.a
80	*liap	'can, able'[23]	*lēap*	líap					*liəp 'know, used to' #1286.b
81	*kaʔiap	'centipede'	*kaēap*	kaʔiáp		kaiep			*kʔiəp #1226.c
82	*ʔiat	'squeak'		ʔiãt			ʔẽ:t 'sound (as a cradle does)'		
83	*miaʊ	'cat'	*meàu*			kumiao	kumea:v	*ku-mëňv*	*miəw #1838.a
84	*ciaʔ	'tree'	*chīa*	cíaʔ					*ɟhe:ʔ #254.c
85	ci(h)	'who?'	*chī*		ci:	ʔacih	ʔacih		*[ʔ]ciʔ #46
86	*ᵈri:ᵈri	'all, whole'	*dī-re*	ríri					*dî:ʔ 'one' #816.a
87	*ɲi:	'house'[24]	*ñī*	ɲí	ɲi:				*sŋiʔ #37.a
88	*kaɪi:	'road, path'	*kaiyī*	kají	kɑji:				*kraʔ #162.a
89	*ʋi:(ʔ)	'make, do'	*wī*	wíʔ	ʋi·ʔ	wi:	vi:	*vī*	
90	lic	'to peel'		ʔitlíc	li·c				
91	*sih	'blow nose'		hĩh			sih		*ksi:r #1680.A

[20] GN *ishât*, TB *isseät*, Ch *ishât*.
[21] GN *heg*, TB *hēang*, Ch *hēang*.
[22] GN, TB, Ch *wī·ang*.
[23] TB *līap*.
[24] GN, TB, Ch *ñī*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
92	*ŋih	‘clam’	*ingeh*			til ŋĩh		*til-ngi-iṅ*	
93	*(ʔ)jih	‘come, arrive’		ʔíh ‘near’	ʔĩ˙x ‘come near’	jih	jih	*yih*	
94	*ɲih	‘to sell’	*iñīh*	ɲíh	ɲi˙x	iɲih ha	ʔiɲiha	*i-nyih*	
95	*ʔicih	‘to stitch’	*ichīh*	ʔicíh	ʔici˙x				*[ ]ɟi:s #1897.a
96	*ʔin	‘in (prep.)’	*en*				in	*ʔin* ‘in, from, to, with, by, on’	
97	*-ʔiŋ	‘bone’	*ong-eng*	ʔuŋʔíŋ	ʔuŋʔi˙ŋ				*cʔi[ ]ŋ #488.c
98	*ɹiŋ	‘hard’		ʔíŋ					*ri:ŋ ‘hard, savage, harsh’ #657.b
99	*caliŋ	‘long’	*chaling*	calíŋ	cɑli˙ŋ				*ɟli:ŋ ‘long’
100	*ʔamis	‘rain’[25]	*amiṅh*	ʔamís	ʔɑmi˙s				*mih #127.b
101	*kiɹit	‘grind’				kirit		*ki-ṛiöt*	*ri:t ‘rotate, go round, grind’ #1056.a
102	*ᵈrit	‘tail’	*det*	rét	ɹe˙t	lam̲rit	lamrit	*rit, lamrit*	(*kti:t #1007)
103	*cikeaŋ	‘elbow’[26]	*det-ongkēang*	rétʔuŋkiáŋ		sikɔŋ	sikɔŋ	*si-kòng*	*kiəŋ[ ] #891.b
104	*me(h)	‘you’	*me*		mɛ̃:	meeh		*meh*	*mi[:]ʔ #128.a
105	*ɹaheʔ	‘vein, nerve’	*ihē*	ʔihɛ́		ɾahe:	rahe:ʔ	*ra-hē*^{k}	*[ ]rsiiʔ #249.a
106	*hej	‘fibre’	*heōe*	héj					*ks[i]ʔ #246.a
107	*sej	‘lice’	*shēi*	séj	se˙j ‘animal’				*ci:ʔ #39.a
108	*ɹeh	‘begin’	*orēha*			ɾɛ tɛn ɾɛ	rehtɛnrɛ	*rĕh*	
109	*ʋet	‘duck’[27]	*wet*	wét	ʋe˙t				
110	*ɹeʋ	‘crocodile’[28]	*yēo*	ʔijáw		rew	ɾe:vu	*ṛēv*	
111	*kahɛ:	‘moon’[29]	*kâhē*	kahɛ́	kɑxɛ̃˙ʋ				*khəjʔ #1542.a
112	*ᵈrɛ:	‘reflexive’	*dē-de ‘self’*			-ɾɛ	-rɛ	*-re*	*dɛʔ #87.a
113	*kamalɛ:	‘sea’	*kamalē*	kamalɛ́	kɑmɑlɛ˙ʔ				*d[n]liʔ #210.a
114	*pumpɛ:	‘to roll’	*pomlē*	pumlɛ́					*ple[:]ʔ ‘rotate’ #213.a
115	*ŋɛ:	‘voice’	*ngē*	ŋɛ́ ~ ŋé					*[s]ŋə:j ‘speak’ #1457
116	*fɛ:	‘you’	*ifē*	ʔife:					*piʔ #905.a
117	*pɛ:c	‘snake’[30]	*pai(ch); pai(j)*	pác	pɑ˙c	pe:c	pe:c	*pēc*	

[25] TB, Ch *amiṅh*.

[26] GN *gut-ong-kē˙ang*, TB *det-ong-kē˙ang*, Ch *det-ong-kē˙ang*.

[27] **vi:t* proto-Vietic (Ferlus 2007).

[28] GN *yēo*, TB *ē˙o*, Ch *ē˙o*.

[29] GN *kâhē*, TB *ka-hai*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
118	*kahɛːm	'younger brother'[31]				kahaem	kahɛːm	*ka-hem*	
119	*ɹikɛ̃j	'thin'		ʔikɛ̃j					*[ ]rgəj #1451.a
120	*ᵈrɛh	'first'	*orēh*		ɹeˑx	raneh	ranɛh	*ra-neh*	*dìːs 'one' #86.b
121	*ʔɛ̃h	'this (near)'	*eṅh*	ʔɛ̃h ~ ʔɛ̃h	ʔɛ̃ˑx				*ʔ[əj]h ' deictic' #1435a.b
122	*hɛh	'to fly'	*hēh-hanga*	hɛ́h	xeˑx				*hə(ː)r #1683.a
123	*katəː/ɯa	'stay, dwell'	*kâ-tö*	katɯ́	t̪ɯː				*dəːʔ 'stop, come to rest' #78.a
124	*ʔijɯam	'breathe'	*eyām*	ʔijɯ́am	ʔijɯəˑm				*jhu(ə)m 'breathe, live' #1299.a
125	*ᵈrɯan	'to perch'	*düan-hata*	rɯ́an	ɹɯˑən	rɔːn		*röön*	*dʃuːn #1158
126	*cɯap	'to wear (skirt)'	*op-chiap, op-chuap*	ʔucɯ́ap					*[ʔ]cuəp/[ʔ]ciəp #1244.b/c
127	*ʔupɯaʔ	'to carry on back'		ʔupɯ́aʔ					*ɓɔʔ (#121)
128	*sajɯh	'year'[32]	*shaiyūh*	sajɯ́h	sɑjɯˑx	samijul		*sam-yeu-hö*	
129	*ʔɯj	'smell'	*öi*	ʔɯ́j	ʔɯ̃ː				*sʔuj 'rotten, to stink' #1441
130	*hɯt	'sniff'	*höt-hata*	hɯ̃t		huːt hət		*hūt* 'to take soup'*;* *höt* 'into'	*hət #1104.b
131	*ɹəːc	'sparrow'				ɾ̪əcə	rɤːcə		*rac 'sparrow' #838.a
132	*səːc	'wash'	*et-shēch-hanga*	séc	seˑc		səːc, səc	*söich*	(Mnong caːc 'to splash water')
133	*fəːj	'wind'	*fũi*	fúj			lamfɨj 'whirlwind'		*cpiər #1638.a
134	*kamləːk	'worm'	*kamilök*	kamilək	kɑmiləˑk	kamləːkə	kamlɤːkə	*kam-löö-kö*	
135	*hatəːm	'night'	*hatòm*	hatə́m	xɑt̪ɔˑm	hataːm	hatəːm	*ha-töom*	*btɔm #1352.a
136	*ɲəːp	'wink, blink'	*ñap-oal-mât*			paɲəːpa		*pa-nyöö-pa*	*rʔiəp 'close [eyes]' #1228.c
137	*ɹəːh	'dry'	*heash* 'wither'	hijə́h ~ hijə́s	xeˑəx				*srɔs 'dry' #160.b
138	*taləh	'fall, slip'	*talöah-hi-lâh*			tələh		*ta-leū-si*	(Jahai tlĩh 'fall to the ground')
139	*təh	'float'[33]		tə́h	t̪əˑx	təh			

[30] GN *paich*, TB *paich*, Ch *pē(d)*.
[31] Khmu *hɛːm* 'younger sibling'
[32] GN *shâü*, TB *samen-nēoh*, Ch *samâīˑha*.
[33] *tɔeh* 'float (stationary)' Khmer

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
140	*ᵈrəh	‘near’				ɾəhtə ‘near’	rəhtə ‘near’	*röh-ta ‘near’*	*tdə̂h ‘near’ #2014.b
141	*təː	‘level’	*tā*	tá ‘level’, tə́ ‘plain’	t̪əː				*taʔ ‘level’
142	*fəː	‘swell/abcess’	*füa*	fúa	fɯðː				*poːʔ #101.a
143	*fəː	‘to blow’	*ifüa*	ʔifúa	fɯðː	fəː	fəː	*föö*	*cpiər #1638.a
144	*haᵈrək	‘hiccup’	*hīdā*	hirə́ʔ		harək ~ haɾ̪ək	harɤk	*ha-rök*	
145	*cək	‘pain’	*chòk*	cák ~ cə́k		cək	cɤk	*chök*	
146	*sək	‘stand up’[34]		sə́k		sək	asɤklə	*sök*	
147	*kintəl	‘heel’[35]	*kentöla-lâh*						*kdə̂l ‘heel’ #1748.c
148	*kəl	‘to break’	*kâṅl-hanga* ‘to halve’			likkəl	likɤl	*li-köl*	*kal #1702.a
149	*səm	‘ten’	*shòm*		sɔːm	səm	sɤm	*söm*	
150	*ᵈrəm	‘to hammer’	*dòm*	rə́m ~ rám				*röm*	*də̂ːm #1361a
151	*fə̀ŋ	‘crossbow’	*fòin*	fə́ŋ	fɔːŋ				*paŋʔ ‘to shoot’ #905.a
152	*təŋ	‘arrive’[36]	*tang-la*			tɯŋ lə			
153	*ᵈrəŋ	‘play instrument’	*ong-dang* ‘music’		ɹəːŋ	ɾ̪oŋ	rəŋ	*röng*	*kndaŋ ‘listen to’
154	*ᵈrəp	‘cover’	*danap, danâp*					*ra-nup* ‘awning’	*də̂p #1261.a
155	*ha-kəp	‘to fit, fix’		hakə́p		hakap		*ha-köp ‘hold’*	*kap ‘to fasten’ #1232.a
156	*ʔət	‘not’				ʔət	ʔət	*öt*	*ʔət ‘used up, lacking’ #943
157	*kət	‘to cut’					kɤt	*köt*	*kat #958.a
158	*ʔinfua	‘dream’[37]	*enfũa*	ʔinfuá					*mp[ɔ]ʔ #105.a
159	*tuac	‘to husk coconut’	*kentòit* ‘coconut husk’	hatuác ‘tear out’			tɔːc	*tôich*	
160	*kinsuah	‘nail, claw’[38]	*ke-shuah*	kisuáh	kisuɑːx	kinsɔh, kunsɔh	kinsɔh, kasɔh	*kin-sòh*	(Jahai cənrɔs ‘claw, nail’)
161	*fuah	‘open, uncover’	*ofoah*	ʔufuáh					*puəh #2029.c
162	*kuah	‘to shave’	*ikōah*	ʔikúah				*ku-a-ha ‘scrape’*	*kuəs #1881.b

[34] TB *shòk-le*.
[35] TB, Ch *kentöːla-lâh*.
[36] *təːŋ* ‘arrive’ Khmu.
[37] GN *enfũːa*, TB *enfô*, Ch *om-fē*.
[38] GN *kishuâːh*, TB *ke-shòh*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
163	*ᵈruaj	'beckon'	*iruai-hata*	ʔiruáj	ɹuɑˑj				*dûəj 'swing, dangle' #1473.b
164	*huaj	'between'		huãj			ihɔːj		
165	*ŋuak	'snore'	*hi-ngôak*	hiŋuák		kin ŋɔːkə	kiŋɔ̃ːk	*kin-ngô-kö*	*srŋɔːk #932.a
166	*cuak	'step'	*kochōak*	cuák					*ɟuək 'tread' #301.c
167	*ɲuak	'suck'[39]	*ngyūàk*	ɲuák 'breathe'		ɲɔːk	ɲɔ̃ːk	*nyôk*	
168	*kaᵈruak	'to knock'	*komdwâk-hata*	karuák					*dɔ[ ]k 'to hammer' #333.a
169	*ɹuaŋ	'fruit'	*yūang*	juáŋ	juɑˑŋ	rəŋ coːn		*ṛòng*	
170	*kanuaŋ	'knee'[40]	*kōi-kanōng*	kújkanuáŋ	kuˑj kɑnuɑˑŋ				
171	*foaŋ	'window'	*foàng*	fuáŋ					*pɔːŋ #608a.a
172	*tuap	'grasp with tongs'	*toâpa 'tongs'*	tuáp 'tongs'		(tenəp 'tongs')	tɔːp 'grasp with tongs', tanɔ̃ːp 'tongs'	*(ta-nap 'clamp')*	
173	*kajuaʔ	'birth'	*kaiyūa*	kajuáʔ			kajɔː	*kai-yōᵏ*	
174	*ɹoac	'overflow'	*yuait-nga*	juácŋa		tareːci	taṛeːci		*ruəc 'fall, drip' #843.b
175	*toah	'breast'[41]	*toah*			tɛh		*teh*	*tɔh #1999.a
176	*moah	'nose'	*moah*	muáh	muɑˑx	mɛh	ʔɛlmeh	*el-meh*	*muh #2045.a
177	*ʔoah	'to cough'	*oōàh*	ʔuʔuáh		ʔɛhɛ	ʔɛhɛ		
178	*koal	'arm'[42]	*koâl*	kuál	kuɑˑl	kɛl	kɛːl	*kël*	*[ɻ]guːl 'finger' #1717
179	*ʔoal	'in, inside'	*oal, òl*	ʔuál		ʔɛl	ɛl	*el*	
180	*foan	'four'	*fōan*		fuɑˑn	feːn	fɛːn	*fën*	*puən #1166.b
181	*pahoaʔ	'fear, afraid'[43]	*pahôa*	pahuáʔ	pɑxuɑˑʔ				*bhaʔ #261.a
182	*ʔaluː	'bat'	*a-lōâa*						*klwaʔ #237.a
183	*ʊuː	'current, flow'	*wuâ*		ʊuɑː				*huər[ ] #1686.b
184	*ŋuː(ʔ)	'green/blue'	*chu-ngôa*	ŋuá		liːŋu	liŋuːʔ	*lī-ngu*	*ɻŋ[ɔː]r #1585.b
185	*luːc	'to shed'[44]	*et-lōi(ch)*	ʔitlúc		luːc 'scrape (hide)'		*leūich*	

[39] Semelai *ʔŋuk* 'suck; smoke'

[40] Khmer *cʊəŋkʊəŋ* 'knee'.

[41] GN *toâh*, TB *tòh*, Ch *tòh*.

[42] Gn *koâl*, TB *kôr*.

[43] GN *pahönˑa*, TB *pahô*, Ch *pahôṅ*.

[44] Khmu *luəc* 'peel off, skinned off'.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
186	*ɻuːc/*ɻəc	‘melt, dissolve’	*yūit-nga*	júacŋa		rəc ŋə	ɻɤc	*ɻ̣öich*	*ruəc ‘retreat, withdraw’ #842.c
187	*ɹuːj	‘fly’	*yūe*	júaj		ʔinruːəj	inɻuːj		*ruəj ‘fly’ #1504.c
188	*luːj	‘three’	*lōe*		luˑəj	luːj	luːj	*lūöi*	*ʔuːj #1437a.b
189	*ʔuːj	‘warm’[45]	*ōe*	ʔúaj	xiʔoˑəj ‘fire’				*[c]ʔuər #1559a.c
190	*muːk	‘appear’		múak					*məːk ‘emerge’ #378.b
191	*ɹuːk	‘load’	*ok-yūak*	ʔukjúak					*ruək ‘force in, cram in’ #395.c
192	*cuːk	‘place, location’	*chuk*	cúk		cuːək	cuːk	*cūök*	
193	*tuːk	‘pull, draw’	*tuak*	túak		tuːk	tuːk	*tūök*	*[ʔ]tuːk ‘scoop up, root up’ #315.a
194	*muːl	‘gather, collect’	*tomōl-hata*	mól, múl		hamuːl	hamul	*ha-mūl*	
195	*tuːl	‘prop, support’	*hen-tôl*						*duəl ‘prop, support’ #1744.b
196	*tafuːl	‘six’	*tafūal*	tafúal	t̪afuəˑl	tafuːl		*ta-fūöl*	*tuəl #1734a.a
197	*ʔuːl	‘to dig’[46]	*oal-ōl ‘bury’*	ʔúl		ʔul	ul	*ul*	
198	*kuːn	‘child’	*kōan*	kúan, kón		kuɔn	kuːn	*kūön*	*kuən #1127.b
199	*manuːɲ	‘lip’[47]	*manōin*	manúɲ					*muːɲ[ ] ‘mouth’ #911.c
200	*ʔuːɲ	‘twist’		ʔúaɲ, ʔúɲ					*wəɲ #931.c
201	*ᵈɻ̣uːŋ	‘ladder/bridge’	*hen-dūanga*	hinrúaŋa					*rtuəŋ #565.b
202	*kuːp	‘door’				ʔinkuːp	ʔinkuːp	*in-kūp*	*ckuːp ‘to cover’ #1237.b
203	*tuːs-a	‘cotton’	*itōsha*	ʔitúsa		tuːsa		*tū-sa ‘wool’*	
204	*fuːs	‘smoke’	*fush*	fús	fuˑs	fanuːsa ‘steam’	hanũːsŋə ‘steam’	*ha-nus-ngö ‘vapour’*	
205	*kanuːt	‘a comb’	*kanūat-kōi*	kanúat		kanuːt	kanũːt	*ka-nūöt*	*kuət #958b.b
206	*muːt	‘hide, conceal’	*hamūt-hanga*	mút					*muːt ‘enter’ #1046.b
207	*kuːt	‘to comb’	*et-kōat*	ʔitkúat		kut	kuːt	*kūöt*	*kuət #958b.b
208	*luːt	‘to swallow’	*chin-lūat-hashe*						*tluət #1088.a
209	*ɹuh	‘continue’	*yūh-hata*	ʔujúhta		ʔiruhɛn	iʔɻ̣uhə	*in-ɻ̣u-hö* ‘continuance’	
210	cuh	‘go, return’	*chūh*	cúh		cu		*chuh*	

[45] TB *heōˑe* ‘fire’.
[46] GN *eul*.
[47] GN *paṅ-nōˑin*, TB *manôˑin*, Ch *manòiˑi*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
211	*paɹuh	'person'[48]	*paiyūh*	pajúh	paju·x				
212	*luh	'untie'	*lōh* 'remove headgear'			luhɲə	luhɲə	*luh-nyö* 'to unloose'	*loh 'unravel, unfold' #2067.a
213	*kuj	'head'[49]	*kōi*	kúj	ku·j	kuj	kuj	*kui*	*kuj #1443.a
214	*tuj	'next'		tój					*tuj 'to follow, accompany' #1463.a
215	*ʔuk	'back/skin'	*ok*	ʔók		ʔuk	ʔuk	*uk*	
216	*-tul	'carry on head/shoulder'	*òl-tōl*			hatul		*ha-tul*	*du:l[ ] #1742.b
217	*tum	'bunch'	*tōm*		mumn̪u·mə 'all'		tum 'fruit cl.'	*tum 'numeral coefficient'*	*tum #1344.a
218	*kiɲum	'child'[50]	*kenyūm*	kiɲóm	kiɲɔ̃·m				*kɲum #1339.a
219	*fun	'navel'[51]	*fun*	fún		fun			
220	*tuɲ	'fern'	*la-tōin*						*k[l]tu:ɲ #899a.a
221	*puŋ	'group, collection'		púŋ					*bu:ŋ #625.b
222	*ᵈruŋ	'hill'				ɾuŋ	ruŋ	*rung*	*ru:ŋ[] #667.b
223	*ᵈrup	'cover'				rup		*rup*	*ɗup #1261.e
224	*put	'pull'	*pôt*	pót					*pu:t 'strip off' #1024a.a
225	*sut (?)	'to kick'		ʔisút		sut			
226	*po:	'suckle'				hapo:	po:	*pō*	*ʔbu:ʔ #114.a
227	*fo:	'thatch grass'	*fo*			ʔafo:	ʔafo:	*a-fō*	*spuʔ #106.a
228	*hVjo:j	'drunk'[52]	*huyòie*	hujój		hijo:j	hijo:j	*hi-yōi*	
229	*so:k	'to point'		sõk 'index finger'				*sōk* 'to point'	
230	*kiɲo:m	'child'[53]	*kenyūm*	kiɲóm	kiɲɔ̃·m				*kɲum #1339.a
231	*so:n	'to bend'		sónsiri		so:n		*sōn 'bend'*	*[ʔ]cu:n 'to walk bent over' #1142.a
232	*ko:ɲ	'male'	*kòin*	kóɲ	kɔ·ɲ	kiko:ɲə	ko:ɲ	*kōiny*	*[ ]ku:ɲ 'father, mother's brother' #893.a
233	*lo:ŋ	'to hole'	*ong-lòng*	ʔuŋlóŋ		lo:ŋ ti 'deep'			*luŋ[h] #724.a

48 Khmer *proh* 'man, male'.
49 GN, TB, Ch *kō·i*.
50 TB, Ch *kenyūm*.
51 Proto-Katuic **puon, *pun* 'navel'.
52 TB *hōē·òie*.
53 TB *kenyūm*, Ch *kenyūm*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
234	*loːŋ	‘to sink’	*pomlòng-shire*	lóŋsi		loːŋti	loːŋti	*lo-ong-ti*	*lɔŋ ‘immersed’ #721.a
235	*puloːʔ	‘thigh’	*pulô*	puló?	pulɔˑʔ				*blu:ʔ #223.a
236	*ʔoh	‘broken’		ʔõh		laoh ‘break as stick’		*oh*	
237	*ŋoh	‘chest’[54]				ʔɛl ŋoh	ʔɛl ŋoh	*el-ngòh*	
238	*foh	‘to strike’	*ofõh*	ʔufóh	ʔufɔˑx	foh ‘whip’	foh	*foh*	*puh ‘slap, hit’
239	*ɲok	‘jerk, pull up’	*ñuk-hanga*	ɲúkhaŋa		ɲok		*nyòk*	
240	*pok	‘tie, bind’	*pôk-hata*	ʔukpók ‘tether’				*pōk* ‘string together’	*buək #357.c
241	*hol	‘friend’[55]	*homnòl*	hól		hol	hol	*hol*	
242	*poŋ	‘cucumber’	*yūang-pong*						*tpuŋ #614.a
243	*kinpoŋ	‘kidney’	*kenpòng*	kinpóŋ		kilpoŋ	ʔɛl kinpoŋ	*el-kil-pong*	
244	*cɔ(ː)ŋ	‘high’	*chòng*	cóŋ	cɔˑŋ				*ɻ[o]ŋ ‘long, high’ #537.a
245	*ɹɔː(k)	‘word, speech’				rɔː	ɻɔːk	*ɻô*	*roːʔ ‘make an inarticulate noise’ #161
246	*ʔinmɔː	‘snot’	*môṅ*			ʔinmɔːrə	inmɔːɻə	*in-môṅ-ɻö*	*smuər ‘nose, beak’ #1655.a
247	*kulɔːc	‘penis’				kulɔl kulɔːc	kulɔːc	*ku-löich*	*lɔːc #855.b
248	*jɔːk	‘hair’[56]	*yôk*						*suək #467.c
249	*ʊɔːk	‘hook’[57]				wok	vɔːk	*vòk*	
250	*sɔːk	‘pick the teeth’	*ok-shòk-kanâp*						*cuək ‘prod, pierce’ #292.c
251	*ʔuhɔːm	‘breathe’				ʔuhɔːm	ʔuhɔːm	*u-hôm*	*jhu(ə)m #1299.b
252	*ᵈrɔːn	‘bend’	*dôn*	rón					*dʉər ‘curve, arch’
253	*mɔːn	‘pimple’	*môn*						*muən #1186.c
254	*ᵈrɔːŋ	‘mountian’				rɔːŋə	rɔːŋə	*rô-ngö* ‘ridge’	*ruəŋ[] #667.c
255	*ʔuŋ-lɔːŋ	‘neck’[58]	*ong-lônga*	ʔuŋlóŋa	ʔuŋlɔˑŋə				*tluəŋ ‘throat’
256	*kapɔːt	‘wrestle’	*kapôt*	kapót					*kpət ‘struggle’ #1025.a
257	*kɔh	‘beat, kill’		ʔukóh ‘to murder’	kɔˑx ‘beat with club’				*kɔh ‘cut (down)’ #1969.a

54 Kui *ŋɤh* ‘to breath’, Chứt *taŋəh* ‘breath’, etc.
55 GN *holchu*, Ch *ho-lōˑang*.
56 GN *yôk*, TB *hē-òk*, Ch *hē-òk*.
57 Khmer *kaŋvak* ‘hooked’.
58 TB *en-lôˑnga*, Ch *ang-lôˑnga*.

No.	pN	gloss	Nanc. (Man)	Nanc. (Radh.)	Nanc. (Raja.)	Car (Das)	Car (Braine)	Car (Whit.)	pAA (Shorto #)
258	*-ʊɔh	‘fell, slash’	*iwâsha* ‘fell’			wanoh ‘sickle’	ʔuvɔh ‘mow’	*u-vòh* ‘reap’	
259	*ʔɔh	‘firewood’	*òn̂h*		ʔɔ̃ːx				*[ ]ʔuəs #1872.c
260	*tɔk	‘break (as rope)’	*tôk-nga*	tɔ́kŋa		latɔk	latɔk	*latòk*	
261	*lɔk	‘pierce/prick’		kalók		lɔk	lɔk	*lòk*	*luk ‘have a hole’ #430.a
262	*ʔɔk	‘to drink’				ʔɔk	ʔɔk	*òk*	*ʔuək #268
263	*ᵈrVːj	‘leaf’	*rai, dai*	ráj	ɹɑːj	roːj	roːj	*rōi*	*daːj ‘calyx’ #1469
264	*lVːŋ	‘good’			ləːŋ	ləŋ ‘correct’	lɛːŋ		
265	*kVnVŋ	‘crab’	*kinòng*	kinóŋ		kanaŋ	kanaŋ	*ka-nang*	
266	*kiː	‘all’	*ki-hēang* ‘each’	kí					*ge(ː)ʔ ‘3Ppronoun’ #26

KATUIC PRESYLLABLES AND DERIVATIONAL MORPHOLOGY IN DIACHRONIC PERSPECTIVE

Ryan Gehrmann
Payap University
ryangehrmann@gmail.com

Abstract

The phonological history of the Katuic language family, an Austroasiatic sub-group of mainland Southeast Asia, is fairly well understood today (Ferlus 1971, 1974b, 1979; Huffman 1976; Diffloth 1982, Sidwell 2005, Gehrmann 2015, 2016). However, there are two topics which have received relatively little attention in the historical linguistic literature on Katuic: 1) the diachrony of the unstressed, penultimate syllable (presyllable) of Proto-Katuic and 2) the morphophonology of Proto-Katuic. This paper aims to make a contribution in both of these areas by discussing the various structural and sound changes which have affected the presyllables of modern Katuic languages and by reconstructing the morphophonological template and derivational affixes of Proto-Katuic. Changes to the modern Katuic presyllables include the development of presyllable vowel quality contrasts, reanalysis of coda nasals as main syllable onset prenasalization, simplification of geminates or their reanalysis as long consonants, metathesis of coda liquids and simple deletions of coda consonants. Three formal affix types (prefixes, rime-onset infixes and rime infixes) and four morphological processes (nominalization, reciprocation, anticausation and causation) are reconstructed for Proto-Katuic.

Keywords: Austroasiatic, Katuic, Presyllable, Morphology, Derivation

ISO 639-3 codes: bru, brv, ncq, sct, sss, kdt, pac, tth, tto, ngt, ktv, kuf

1 Introduction

The Katuic languages are an Austroasiatic language family of southern Laos, central Vietnam, northeastern Thailand and north-central Cambodia. The language family comprises six sub-groups, two of which, Bru and Kuay, are definitively nested into an intermediate node called West Katuic (Ferlus 1974a, Diffloth 1982, Sidwell 2005, Gehrmann 2016). Table 1 lists the Katuic sub-groups along with their primary glossonyms and corresponding ISO 639-3 codes.

Table 1: *The Katuic language family*[1]

Katuic	**West Katuic**	**Bru**	*Bru [bru, brv], Katang [ncq, sct], So [sss], Khua [xhv], Makong*
		Kuay	*Kuay/Kuy/Suay [kdt], Nyeu [nyl]*
		Pacoh	*Pacoh [pac], Ta'oi/Ta'oih, Cado, Bahi/Pahi*
		Ta'oi	*Ta'oiq [tth], Ta'oih/Ta'uaih/Ta'uas [tto], Ong [oog], Ir [irr]*
		Kriang	*Kriang/Ngkriang [ngt], Ngeq/Nheq/Nyeq, Khlor, Chatong,*
		Katu	*Katu [ktv, kuf], Kantu, Phuong [phg], Triw, Dak Kang*

Phonologically, the Katuic languages are noteworthy among the Austroasiatic languages for their large vowel inventories, especially those of West Katuic languages which are doubled for both length contrast and register contrast (cf. Diffloth 1982, Gehrmann 2016). Also notable are the vowel quality contrasts present in the unstressed, penultimate syllables (presyllables) of geographically eastern Katuic languages (Sidwell

[1] Note that Sidwell (2005, 2009, 2015) also nests Ta'oi and Kriang into one sub-group called Ta'oi.

2005, Gehrmann 2017a). Apart from these features, Katuic languages are phonologically typical Austroasiatic languages of the lower Mekong region with prosodic words of a maximally disyllabic, iambic template and vowel inventories with nine to eleven vowel qualities (Gehrmann and Conver 2015, Jenny et al. 2015). Katuic languages are almost entirely lacking in inflectional morphology but the remnants of an earlier derivational morphological paradigm are readily identifiable.[2] These derivational morphemes are exclusively prefixing and infixing and as such, they are closely linked to synchronic issues of presyllable structure and diachronic issues of presyllable reduction in the Katuic languages.

This paper is part of a larger comparative investigation of presyllable structure and segmental inventory across the Katuic language family. This project was inspired by an outstanding question in the historical study of the Katuic languages – where did the presyllable vowel contrasts of Bru, Katu and especially Pacoh come from? That question has now been addressed to my satisfaction (cf. Gehrmann 2017a), but in order to investigate the issue, it was necessary to undertake a general comparative study of the Katuic presyllable and Katuic derivational morphology first. This paper presents the results of that comparative study.

In this paper, I will describe the diversity of presyllable structures found in the modern Katuic languages, introducing various language-specific pathways of diachronic presyllable reduction. A discussion of Katuic derivational morphology in synchronic and diachronic perspective follows, in which a derivational morphological template for Proto-Katuic (PK) is proposed and individual nominalizing, reciprocal, anticausative and causative affixes are reconstructed. Only the non-reduplicative morphology of PK is discussed here.[3]

2 The Proto-Katuic Phonological Word

Sidwell (2005) reconstructs the following word template for PK:

$$(c_i\{v/c_f\})C_i(C_m)V(C_f)$$

PK words may be monosyllables or disyllables. The final syllable or *main syllable* of a disyllabic word is always prosodically prominent and phonotactically unrestricted while the penultimate syllable or *presyllable* is always unstressed and phonotactically restricted with respect to the main syllable. A monosyllable is by definition a main syllable without a presyllable and, therefore, phonotactically equivalent to the main syllable of a disyllabic word. The underlying structure of the presyllable is deficient compared with that of the main syllable in two significant respects: 1) no medial consonant is permitted between the presyllable onset (c_i) and the presyllable rime ($\{v/c_f\}$) and 2) the presyllable rime itself may contain no more than one sonorant segment. This presyllable rime segment may be either a vowel (v) or a consonant (c_f), unlike the main syllable rime, in which both a vowel and a coda consonant may occur together (VC_f).

The PK word cannon does not permit *sesquisyllables*. The PK phonological word is a *maximally disyllabic iamb* (cf. Butler 2015a) but not a sesquisyllable in the original sense of the word as coined by Matisoff (1973).[4] True sesquisyllables are monosyllables with two prevocalic consonants, between which an excrescent vocalic transition is inserted in order to facilitate the articulation of the consonant cluster (e.g. /CCVC/ [C̥əCVC]). Sesquisyllables come about due to a reduction in the maximal syllable template of a language like PK with maximally disyllabic iambs. In this process, etymological presyllables atrophy to the point that they lose their status as syllables in their own right and become absorbed into the main syllable. Eventually, the onset clusters of the sesquisyllable will further reduce to either simplex onsets or complex onset clusters which adhere to the sonority sequencing principle, completing the monosyllabicization process. Many languages of East and Southeast Asia have completed or are currently undergoing such a reduction towards monosyllabicity.[5]

Sesquisyllables are therefore a transitional word shape and sesquisyllabicity occupies the middle ground on a continuum between monosyllabicity and disyllabicity. Since this diachronic progression operates along

---

[2] Note, however, case-marking of pronouns in the Pacoh sub-group, which is a relatively recent innovation (Solntseva 1996; Alves 2006, 2007).

[3] Sizeable works on reduplication in Katuic are available for Pacoh (R. Watson 1966) and Kriang (Smith 1973).

[4] The term *sesquisyllable* is often applied broadly, becoming essentially synonymous with *prosodically iambic*.

[5] A clear example of this process has been described in Nyaheun (Ferlus 1971, Sidwell & Jacq 2003, Sidwell 2012).

a continuum, there is a point during the transition from disyllable to sesquisyllable at which the two syllable shapes are essentially indistinguishable using only phonetic criteria. For that reason, when analyzing the syllable canon of a language synchronically, a phonological test of syllabicity provides a more decisive answer. In this paper, we will consider disyllables to have restructured into sesquisyllables in a language only when the quality and presence of the erstwhile vocalic nucleus of the presyllable has become entirely predictable. Strictly speaking, the syllable canon of PK may not be called sesquisyllabic on account of the phonologically real, vocalic syllable nucleus which separates c_i and C_i. The justification for this presyllable vowel is discussed in Section 3.

The minimal PK word is an open monosyllable with an onset consonant (C_i) and a long vowel or diphthong (V)[6]. Main syllables may be open or closed by a coda consonant (C_f). An optional medial liquid (C_m) may stand between the main syllable onset (C_i) and the main syllable vowel (V) but medial liquids are permissible only after oral stops or *s in the C_i position. The presyllable is optional but when it does occur it must contain both an onset (c_i) and a rime comprised of either a vowel (v) or a sonorant coda consonant (c_f) - never both. Presyllable structure is discussed in greater detail in Section 3.

Pacoh is a Katuic language which has preserved the PK word and syllable template intact. The Pacoh examples in Table 2 are taken from Watson et al. (2013) and demonstrate the twelve possible PK word shapes.

Table 2: *Examples from modern Pacoh demonstrating the twelve possible PK word shapes*[7]

C_iV	/caː/	to eat	**C_iVC_f**	/sɔːj/	tail
C_iC_mV	/truː/	deep	**$C_iC_mVC_f$**	/klɔːŋ/	trail (animal)
c_ivC_iV	/kaɟəː/	crab	**$c_ivC_iVC_f$**	/tapat/	six
$c_ivC_iC_mV$	/katruː/	spotted dove	**$c_ivC_iC_mVC_f$**	/kaplḭh/	blink, wink
$c_ic_fC_iV$	/krnaː/	road	**$c_ic_fC_iVC_f$**	/pntu̯ːr/	star
$c_ic_fC_iC_mV$	/pŋkraː/	to repair	**$c_ic_fC_iC_mVC_f$**	/tmpraːŋ/	clf. for crossbow

3 Proto-Katuic Presyllables

In addition to being phonotactically restricted, the inventory of phonemes which may occur in the PK presyllable is also impoverished when compared with that of the main syllable. The implosive, nasal and semivowel glide consonants which are permissible in the main syllable onset do not occur in the presyllable onset. The presyllable coda may only be filled by sonorant consonants, namely the liquids *r and *l and a nasal *N which assimilates its place of articulation features from the main syllable onset in all cases. The presyllable vowel is entirely underspecified for place features and this vowel was realized as a mid to open central vowel in PK as it is in modern Katuic languages.[8] Note these deficiencies in the segmental inventory of the PK presyllable vis-à-vis the main syllable segmental inventory in Table 3, which lays out Sidwell's (2005) reconstruction of the segmental inventory of PK.

The presyllable rime may either be open or closed. Open presyllable rimes contain only a vocalic nucleus (v) and closed presyllable rimes contain only a sonorant coda consonant (c_f). There is no justification for positing a phonological vowel in closed presyllables. When the presyllable is closed, a short, epenthetic, schwa-like vocalic transition is inserted between c_i and c_f (e.g. PK *sŋkaɲ 'catfish' [sᵊŋˈkaⁱɲ]).

[6] Short vowels do not occur in open syllables in PK.

[7] Throughout this paper, a regularized transcription system for Katuic languages is employed in order to aid in crosslinguistic comparison.

[8] Note, however, that we cannot at this time rule out the possibility that presyllable vowel quality contrasts, developed under Chamic influence, were actually present in PK (Gehrmann 2017a).

Table 3: *PK phoneme inventory*

Presyllable						Main Syllable												
C_i					**v/C_f**	**C_i**					**C_m**	V			**C_f**			
						*ˀb	*ˀd	*ˀɟ										
*b	*d	*ɟ	*g			*b	*d	*ɟ	*g									
*p	*t	*c	*k	*ʔ		*p	*t	*c	*k	*ʔ		*ia	*i̯a	*ua	*p	*t	*c	*k	*ʔ
					*N	*m	*n	*ɲ	*ŋ			*ie	*i̯ə	*uo	*m	*n	*ɲ	*ŋ	
	*l				*l	*w	*l	*j			*l	*i(:)	*i̯(:)	*u(:)	*w	*l	*j		
	*r				*r		*r				*r	*e(:)	*ə(:)	*o(:)		*r			
	*s			*h	*a		*s			*h		*ɛ(:)	*a(:)	*ɔ(:)		*s			*h

Open presyllables in PK were very nearly analyzable as sesquisyllables, as is still the case in many modern Katuic languages. If a language is to be described as permitting sesquisyllables, the appearance and quality of the phonetic presyllable vowel must be entirely predictable from the environment. In PK, any two consonants preceding the main syllable vowel were predictably syllabified into a sesquisyllabic structure ([c_iəC_i]) with one exception: sequences of stop+liquid are unpredictably syllabified into either 1) a monosyllable with a tautosyllabic consonant cluster onset (e.g. PK *klɔːk ‘cowardly’, *tri̯a ‘mushroom’) or 2) a sesquisyllabic shape (e.g. PK *kalaːŋ ‘hawk’, *tariːk ‘buffalo’). In a language with true sesquisyllables, only the tautosyllabic realization is typically permissible, as in modern Khmer (Henderson 1952, Huffman 1972, Thomas 1992, Butler 2015a). All modern Katuic languages outside of the Kuay sub-branch demonstrate this contrast of syllabicity for prevocalic stop+liquid consonant sequences, as demonstrated in the examples in Table 4.

Table 4: *Examples of unpredictable monosyllabicity/sesquisyllabicity across Katuic languages*

Sub-Group	Language	Monosyllabic stop+liquid		Disyllabic stop+liquid		
Katu	Western Katu	/plɛːŋ/	*power to cure*	/palɛːŋ/	*truly*	(Sulavan et al 1998)
Pacoh	Pacoh	/ploː/	*head*	/paloː/	*to kindle a fire*	(Watson et al. 2013)
Ta’oi	Ta’oiq	/plɔː/	*head*	/palɔː/	*to ignite*	(Conver et al. 2017)
	Ta’oi	/plaj/	*to overflow*	/palaj/	*fruit*	(L-Thongkum 2001)
Kriang	Kriang	/pla̤ːʔ/	*fish*	/pala̤ːʔ/	*elephant tusk*	(Gehrmann et al. 2016)
	Chatong	/plaː/	*knife*	/palɛː/	*fruit*	(L-Thongkum 2001)
	Eastern Bru	/plaw/	*infertile*	/palaw/	*to harm*	(Miller & Miller 2017)
Bru	Western Bru	/plah/	*to shatter*	/palih/	*to pick fruit*	(Gehrmann 2016)
	Southern Katang	/praːj/	*to heal*	/para̤ːj/	*thread*	(Gehrmann 2016)

Pittayaporn (2015) has termed this particular type of unpredictable syllabification in languages of Mainland Southeast Asia the ‘contrastivity of sesquisyllabicity’ and it amounts to smoking gun evidence for a phonologically real presyllable vowel.[9] As Pittayaporn points out, syllabification is not included in a word’s underlying representation but rather, the syllabification of a word must be predictable based on language-specific rules which are applied to the string of segments found in the underlying representation of a word.

In light of this, we are obliged to include a phonological vowel between the stop and liquid consonants of PK words such as PK *kalaːŋ ‘hawk’ in order to prevent their incorrect syllabification into monosyllables (e.g. PK *kalaːŋ ‘hawk’ ≠ PK *klaːŋ ‘to pipe water’). Because this vowel is phonologically real in this one environment, we may analyze it as real in the underlying representation of all open presyllables in PK, thus allowing the simple word structure template presented above to correctly describe all monosyllables and disyllables in PK. Note, however, that if the contrast of syllabification of stop+liquid sequences becomes neutralized, as it has in some Kuay languages, the language then gains the sesquisyllable as one of three

[9] This analysis was anticipated in Thomas’s (1992) *type ii* sesquisyllables.

word archetypes – monosyllables Ci(Cm)V(Cf), sesquisyllables CpCi(Cm)V(Cf) and disyllables cicf.Ci(Cm)V(Cf) – complicating the word canon of the language.

4 Pathways of Presyllable Reduction in Modern Katuic

The description of the PK word above does not accurately describe the prosodic word of any one modern Katuic language. Various phonological changes both big and small have affected the presyllable structure of the modern Katuic languages. The majority of these changes are phonological reductions, which is to say simplifications of the PK structure. Note however, that the presyllable vowel quality contrasts developed in Katuic were a rare case of presyllable strengthening in a linguistic area which is otherwise known for presyllable weakening (cf. Gehrmann 2017a).

A description of various pathways of structural reduction and segment reanalysis in Katuic presyllables is presented below. These pathways may be divided into two categories: those which affect presyllable vowels and those which affect presyllable coda consonants.

4.1 Presyllable Vowel Changes

There is a three-way distinction to be drawn between Katuic languages in terms of the status of the presyllable vowel. The three categories are determined by the presence or absence of presyllable vowel quality contrasts (PVCs) and the presence or absence of the contrastivity of sesquisyllabicity (CoS) (cf. Section 3). The Katuic languages with the most robust and unambiguously phonologically real presyllable vowels are those which have both PVCs and CoS. A second category with more reduced presyllable vowels is comprised of languages which have no PVCs but do maintain CoS. Finally, there are at least two Katuic languages from the Kuay sub-group which have neither of these features. So far, no Katuic language has been described as having PVCs while not having CoS though this remains theoretically possible. Table 5 summarizes these categories.

Table 5: *Presyllable vowel typology in modern Katuic*

	PVCs	**CoS**	**Example Languages**
Robust Presyllables	✓	✓	Pacoh, E. Bru, N. Katang, W. Katu
Marginal Presyllables	✓	x	Kriang, W. Bru, S. Katang, E. Katu, Ta'oi
Sesquisyllables	x	x	Kuay

Only Pacoh, Bru and Katu languages have PVCs among the modern Katuic languages. In all three of these languages, /a/ is certainly the default presyllable vowel quality but /i/ and /u/ are also found. Pacoh maintains a robust opposition of the three vowel qualities (/i, a, u/) in many environments but PVCs are more restricted in their distributions in Bru and Katu, as demonstrated in Table 6. The data in Table 6 is based on an analysis of the substantial dictionaries available for Eastern Bru (EBru) (Miller and Miller 2017), Western Katu (WKatu) (Sulavan et al. 1998) and Pacoh (Watson et al. 2013). The numbers represent absolute counts for non-open presyllable vowels occurring in unique etyma which appear in the dictionaries. These figures are included here to demonstrate the cline in the distribution of the close presyllable vowels relative to presyllable onsets across the Katuic languages which have PVCs.

Table 6: *Distribution of non-open presyllable vowel qualities in Katuic languages*

		Main Syllable Onset										
		Labial		**Alveolar**		**Palatal**		**Velar**		**Glottal**		
		/u/	**/i/**	**/u/**	**/i/**	**/u/**	**/i/**	**/u/**	**/i/**	**/u/**	**/i/**	
Presyllable Onset	**p**	-	-	-	-	-	-	-	-	-	-	**EBru**
	t	-	-	-	-	-	-	-	-	-	-	
	k	69	-	108	-	27	-	2	-	12	-	
	ʔ	-	-	-	-	-	-	-	-	-	-	
	p	-	-	-	12	-	-	-	5	-	2	**WKatu**
	t	-	8	-	21	-	3	-	7	-	1	
	k	-	-	3	11	-	-	-	-	-	-	
	ʔ	6	10	2	13	-	6	1	4	-	1	
	p	1	4	-	79	-	15	-	22	-	18	**Pacoh**
	t	57	1	15	36	3	3	1	44	3	8	
	k	31	-	38	37	16	15	3	-	6	-	
	ʔ	1	13	11	27	3	15	3	13	-	2	

As discussed in Gehrmann (2017a), there is good evidence that PVCs were previously more widespread in Bru and Katu but are now being leveled off due in large part to contact with prestige languages which lack PVCs such as Vietnamese and Southwestern Tai languages. For example, in modern EBru, PVCs are extremely limited in distribution with /a/ and /u/ only in contrast following presyllable /k/ onsets (e.g. /katop/ 'unexpectedly' vs. /kutop/ 'to rout', /katɜw/ 'to heat something up' vs. /kutɜw/ 'to be hot'). Even within the modern Bru and Katu languages, we find some conservative languages which maintain marginal PVCs and others which have already lost them completely.[10]

Other Katuic languages have CoS but the presyllable vowel is nevertheless completely underspecified for vowel quality features. In these languages, the presyllable vowel is phonologically real but its phonetic realization is predictable as a mid to open central vowel or another allophonic vowel quality predictably conditioned by surrounding consonants. Languages at this second stage have very nearly developed sesquisyllables, since the syllabification of two prevocalic consonants is predictable for all possible combinations except the aforementioned obstruent + liquid clusters. Table 7 demonstrates this.

Table 7: *The presyllable vowel is contrastive/phonologically real only for obstruent+liquid clusters*

		Main Syllable Onset				
		obstruent	**nasal**	**liquid**	**glide**	**ʔ**
Presyllable Onset	**obstruent**	[ə]	[ə]	**/a/ ≠ ø**	[ə]	[ə]
	nasal	[ə]	[ə]	[ə]	[ə]	[ə]
	liquid	[ə]	[ə]	[ə]	[ə]	[ə]
	glide	[ə]	[ə]	[ə]	[ə]	[ə]
	ʔ	[ə]	[ə]	[ə]	[ə]	[ə]

A third category for the status of the presyllable vowel in Katuic languages is filled by languages which have neither PVCs nor CoS. At least two Kuay languages have lost CoS and consequently reanalyzed open presyllables into phonological sesquisyllables (Prasert 1978, Bos personal communication). This development can surely be explained by Kuay's close historical contact with Khmer, which has undergone the same development.

[10] For Bru, see Gehrmann (2016:76). For Katu, see Costello (1998).

4.2 Presyllable Consonant Changes

Changes to the presyllable coda consonants across the Katuic languages generally affect the two liquid codas (*-r-, *-l-) and the nasal coda (*-N-) differently.

*4.2.1 Status of Reflexes of *ʔN*

All modern Katuic languages have been described as having either *syllabic nasal presyllables* or *consonant prenasalization*, both of which are reflexes of PK presyllables of the shape *ʔN. A conservative phonetic realization of these *ʔN presyllables, one which follows the regular rules for the vocalization of CC presyllables, is found in WKatu. As data from Costello (1998), Sulavan et al. (1998) and L-Thongkum (2001) indicate, the same short, epenthetic schwa which is found between the consonants in all other CC presyllables is found in the reflexes of *ʔN in WKatu (i.e. /ʔN/ [ʔᵊN]). In many other Katuic languages, Pacoh being a well documented example (Alves 2006:20–21), no vowel epenthesis occurs in the reflexes of *ʔN, but rather the nasal coda itself becomes the sonorant peak of the presyllable resulting in a syllabic nasal presyllable (i.e. [ʔN̩]). In yet others, the glottal stop onset in reflexes of *ʔN has been lost.

When this happens, we are left with three options for how to analyze the prenasal synchronically: 1) define a new syllable archetype for syllabic nasals (/N.CVC/, e.g. /m.pat/), (2) allow for onset clusters which violate the sonority sequencing principle (/NCVC/ e.g. /mpat/) or (3) posit a series of prenasalized consonants (/ᴺCVC/ e.g. /ᵐpat/). The most parsimonious solution is to accept the third option, which though resulting in a phonemic split in into plain and prenasalized consonant series, removes the necessity for positing new syllable shapes or unnatural segment sequences within a syllable. For more details on this process, including phonetic measurements, see Gehrmann (2017b). Table 8 summarizes these categories.

Table 8: *Presyllable coda nasal typology in modern Katuic*

	***ʔN.C**	**Example Languages**
Conservative	**/ʔN.C/ [ʔᵊN.C]**	W. Katu [kuf]
Transitional	**/ʔN.C/ [ʔN̩.C]**	Pacoh [pac], E. Bru [bru], N. Katang [ncq]
Reanalyzed	**/ᴺC/** [ᴺC]	Kriang [ngt], Ta'oiq [tth], Kuay [kdt]

4.2.3 Gemination

Languages from four different Katuic sub-branches (Kriang, Pacoh, Ta'oi, and Katu) show a contrast between sequences of identical sonorants (geminates) across the presyllable: main syllable boundary (/rr/, /ll/, /NN/) and sequences of vowel+consonant across the syllable boundary (/ar/, /al/, /an/). This gemination contrast is reconstructable back to PK based on its retention in this broad subset of modern languages. [11]

The geminates developed through infixation, especially infixation in monosyllabic roots with clustered onsets (see Section 5.3). Gemination phonologized into contrastively long main syllable sonorant onsets in some languages, such as Kriang (Gehrmann 2017b), but in other languages, sonorant length is a predictable phonetic realization of a sequence of two identical consonants. In still other languages, gemination has been lost in a complete merger with corresponding VC type sequences. This results in the vocalization of the presyllable coda sonorants to /a/ (e.g. *rr, *ll, *NN > /ar/, /al/, /aN/). Languages which have undergone this merger and lost the gemination contrast include all Bru languages, most Kuay languages, most Ta'oi languages, and some Katu languages.

4.2.3 Coda Liquid Metathesis

In almost all Katuic languages, presyllables with a liquid in the coda are realized with an excrescent vocalic transition between the onset and the coda ([Cᵊr], [Cᵊl]) and this may be considered the default articulation of such presyllables, inherited from PK. In certain Katuic languages, however, presyllables of the shape stop+liquid have undergone metathesis of the liquid and the epenthetic schwa vowel. For example, in these

[11] Note that the phonetic realization of PK geminates has shifted from increased duration to preglottalization in the Ta'oi variety spoken near Tha Taeng described by L-Thongkum (2001).

languages, the expected [kᵊr] realization of /kr/ presyllables shifts to [krᵊ], with the liquid realized as a medial consonant following the presyllable onset rather than as a final consonant. In Ta'oiq, this is a general shift affecting all sequences of stop + liquid. In Kuay Ntua, metathesis of the stop and liquid is common but optional, with both [CᵊC] and [CCᵊ] realizations being acceptable. In Bru languages, this metathesis seems to have primarily affected sequences of *tr, modern reflexes of which are pronounced /ra/ [ra], presumably through an intermediary step of /tr/ [tra]. Modern reflexes of *kr are also occasionally /ra/ in modern Bru, but this does not apply regularly unlike the reflexes of *tr (Gehrmann 2016:86–87).

4.2.4 Coda Consonant Deletion

Some modern Katuic languages have simply deleted coda consonants. To take an extreme example, Kuy no longer permits presyllable coda consonants at all, though reflexes of nasal codas persist as prenasalization of monosyllable onsets in many cases (e.g. PK *ʔmpaːŋ 'maggot' > Kuy /ᵐpaːŋ/, PK *lmpaːk 'shoulder' > Kuy /ᵐpaːʔ/) (Prasert 1978). EKatu has deleted all presyllable coda liquids but retains presyllable coda nasals (Costello 1971). WKatu has done the opposite, retaining liquid codas but deleting all nasal codas except for those in following a glottal stop presyllable onset (/ʔN/) (Sulavan et al. 1998).

5 Proto-Katuic Derivational Morphology

It is recognized that in Austroasiatic as a whole, verbal roots are generally monosyllabic and disyllabic verbs are generally morphologically complex, having been derived from those roots (Sidwell 2008). Katuic is no exception to this general trend in Austroasiatic verbal typology but, of course, there are a great many exceptions to this. Katuic languages also have *many* disyllabic verb roots, which may themselves be manipulated through morphological marking.

The derivational morphological processes reconstructable to PK all involve affixes which attach to verbal roots. These processes fall into two broad categories: those that turn verbs into nouns (*nominalization*) and those which alter the argument structure of the verbal root. This latter category may increase the verb root's valency by adding a causer/outside agent (*causative*) or they may decrease the verb root's valency either by removing an agent (*anticausative*[12]) or by removing a patient (*reciprocal*). It is unsurprising to find that these four derivational morphological processes are reconstructable to PK because they are the most common processes found throughout the Austroasiatic family (cf. Alves's (2014) overview). In fact, these four processes are reconstructed for Proto-Austroasiatic (PAA) as well (Sidwell 2008, Sidwell and Rau, 2015:234–37). Sidwell's reconstruction of the derivational affixes of PAA are shown in Table 9.

Table 9: *Sidwell's (2008) reconstruction of the derivational affixes of Proto-Austroasiatic*[13]

Nominalizing	*-n-	**Causative**	*p-, *pC-
	*-m-	**Reciprocal**	*t-, *tN-
	*-r- / *-l-	**Stative**	*h-, *hN-
	*-p-		

Different types of affixes attach to roots of different shapes in Katuic. There are three types of derivational affixes, *prefixes*, *rime-onset infixes*, and *rime infixes*. Prefixes consist of two segments, a presyllable onset and rime (c_iv- or c_ic_f-), which attach to monosyllables as presyllables turning the monosyllables into disyllables (e.g. PK *lɔh 'to exit' + *pr- 'NOM' > *prlɔh 'doorway'). Rime-onset infixes consists of two segments, a presyllable rime and a main syllable onset (-vC_i- or -c_fC_i-), and are inserted following the onset consonant in simplex onset monosyllables. This pushes the original main syllable onset out into the presyllable onset position while material from the infix fills in the presyllable rime and main syllable onset positions (e.g. PK *piːh 'to sweep' + *-rn- 'NOM' > PK *prniːh 'broom'). Finally, rime infixes consist of only one segment, a presyllable rime (-v- or -c_f-), which may attach to either a complex onset monosyllable or to a disyllable. When a rime infix attaches to a complex monosyllable, the main syllable onset of the root becomes the presyllable onset, the medial consonant of the root becomes the new main syllable onset and the

[12] Note that what I call here *anticausatives* have often been referred to as *statives*, *resultatives* or even *passives* in the literature on Austroasiatic morphology.

[13] The capital C in Sidwell's reconstructed affixes represents an "unspecified coda".

infix takes its place in the presyllable rime position (e.g. EBru /kluːm/ 'to urinate' + /-r-/ 'NOM' > /krluːm/ 'urine' (Miller and Miller 2017)). When a rime infix attaches to a disyllable, it simply supplants the presyllable rime of the root (e.g. Kriang /taŋaw/ 'to sit down' + /-r-/ 'NOM' > /trŋaw/ 'a seat' (Gehrmann et al. 2016)).

In some Katuic languages, it is possible to combine derivational affixes either by fusing prefixes and infixes together or by attaching more than one prefix to the left of the root. For example, S. Watson (1965) analyzes a complex *causative-reciprocal* prefix /pr-/ in Pacoh which is a combination of the /pa-/ 'CAUS' prefix and /-r-/ 'RECIP' infix. The word /prdu̯c/ 'to make each other angry' would thus be doubly derived from /du̯c/ 'to be angry' + /pa-/ 'CAUS' > /padu̯c/ 'to anger someone' with the subsequent addition of /-r-/ 'RECIP' to the derived stem resulting in /prdu̯c/.

Only one Katuic language, Eastern Katu, permits multiple prefixes. For example, Costello (1966) also analyzes a *causative-reciprocal* form for Eastern Katu, but this one is formed by the simple concatenation of /ta-/ 'RECIP' and /pa-/ 'CAUS' (e.g. /ɲəːp/ 'to be dirty' + /pa-/ 'CAUS' + /ta-/ 'RECIP' > /tapaɲəːp/ 'to make each other dirty'). Gehrmann (2017a) suggests that the presence of stacking prefixes in Eastern Katu alone among the Katuic languages is a legacy of language contact with Old Northern Chamic. This is based on the similar use of stacking prefixes in Northern Roglai, a descendent of the Chamic variety formerly spoken at parallel latitudes to Katu in what is today Vietnam.

In the following sections, the reconstruction of specific PK derivational affixes is discussed. For each affix, we compare the same set of languages; Eastern Katu (EKatu) (Costello 1966), Western Katu (WKatu) (Costello 1998, 2001; Sulavan et al. 1998), Kriang (Gehrmann et al. 2016, Smith 1970), Ta'oiq (Conver et al. 2016), Pacoh (R. Watson 1966, S. Watson 1966, Watson et al. 2013) and Eastern Bru (EBru) (Hoàng and Tạ 1998, Miller and Miller 2017, Vương Hữu Lễ 1997). For brevity's sake, at most three examples supporting the reconstruction of each PK affix is given for each of these six modern languages but a great many more examples are collected in my database, which is available upon request.

Note that certain Katuic languages have undergone sound changes reducing the number of permissible consonants in the presyllable onset position. These sound changes have affected both roots and derived etyma. The most relevant such changes for this paper are 1) the shift of presyllable onset *c to /t/ in Pacoh, /s/ in EBru and /h/ in Ta'oiq and 2) the shift of presyllable *tr to /ra/ in EBru and /hr/ in Ta'oiq (Gehrmann 2016, Gehrmann and Conver 2015). The modern reflexes of PK *crnoh 'crops' (PK *coh 'to plant' + *-rn- 'NOM') and PK *trləːŋ 'walking stick' in Table 10 demonstrate these shifts.

Table 10: *Sound changes affecting presyllable onset consonants*

	PK	**Kriang**	**Pacoh**	**EBru**	**Ta'oiq**
crops	*crnoh	crnoh	trnu̯h	srnɔh	hrnoh[14]
walking stick	*trləːŋ	trlɐ̯ːŋ	trloːŋ	ralɜːŋ	hrlɔːŋ

*5.1 PK Nominalizing Prefix *pr-*

The nominalizing prefix /pr-/ attaches to monosyllables and is primarily, though not exclusively, used to derive verbal nouns in the modern languages. This prefix is stable across the Katuic languages with the exception of languages like EKatu, which have deleted liquids from the presyllable coda. In these languages, the PK coda *r has vocalized to /a/ (see Table 11). Curiously, the initial labial stop is aspirated in WKatu and rather than creating verbal nouns, /pʰr-/ usually derives objects of the verbal action in the examples given by Costello (1998). It is possible that the aspirated form in WKatu is not cognate with PK *pr-. Sidwell (2008) does not reconstruct any such nominalizing prefix for PAA, but various AA families and individual languages have innovated nominalizing prefixes and Katuic is no exception (Jenny et al. 2015, 48-50).

Table 11: *Table of correspondences for PK nominalizing prefix *pr-*

PK	**EKatu**	**WKatu**	**Kriang**	**Ta'oiq**	**Pacoh**	**EBru**
*pr-	/pa-/	/pʰr-/	/pr-/	/pr-/	/pr-/	/pr-/

[14] Note that Ta'oiq /hrnoh/ means 'fields' rather than 'crops' as in the other languages.

Examples of PK *pr- 'NOM' in modern Katuic languages are presented in Table 12.

Table 12: *Examples supporting the reconstruction of PK *pr- 'NOM'*

EKatu				**WKatu**			
jaːl	*to be long*	pajaːl	*length*	lɔh	*to go outside*	pʰrlɔh	*a doorway*
ʔeːp	*to be short*	paʔeːl	*shortness*	caː	*to eat*	pʰrcaː	*sth. eaten*
dil	*to be smooth*	padil	*smoothness*	rap	*to perform a ritual*	pʰrrap	*a ritual*
Kriang				**Ta'oiq**			
lɔh	*to exit*	prlɔh	*door*	lo̱ːh	*to exit*	prlo̱ːh	*door, window*
juʔ	*to be afraid*	prjuʔ	*fear*	ŋal	*to think*	prŋal	*thoughts*
ŋɔːc	*to drink*	prŋɔːc	*feast, celebration*	ʔaj	*to be sick*	prʔaj	*illness*
Pacoh				**EBru**			
ŋoh	*to exit*	prŋoh	*door*	lo̱ːh	*to exit*	prlo̱ːh	*door*
laː	*to split*	prlaː	*wedge*	ciːn	*to reconcile*	prciːn	*a judge*
jɛh	*to turn*	prjɛh	*fork in rd.*	doːl	*to carry on shoulder*	prdoːl	*burden*

*5.2 PK Nominalizing Rime-Onset Infixes *-an-, *-rn-, *-nn- and *-mp*

Four nominalizing rime-onset infixes are reconstructable to PK, three of which are thematically linked by an alveolar nasal in the main syllable onset (*-an-, *-rn-, *-nn-). The fourth inserts a bilabial oral stop into the presyllable onset (*-mp-). The connection between these PK infixes and Sidwell's (2008) reconstructed nominalizing infixes for PAA is obvious (see Table 9). Table 13 demonstrates the correspondences supporting the reconstruction of these four infixes.

Table 13: *Table of correspondences for PK nominalizing rime-onset infixes *-an-, *-rn-, *-nn- and *-mp-*

PK	**EKatu**	**WKatu**	**Kriang**	**Ta'oiq**	**Pacoh**	**EBru**
*-rn-	/-an-/	/-rn-/	/-rn-/	/-rn-/	/-rn-/	/-rn-/
*-an-		/-an-/	/-an-/	/-an-/	/-an-/	/-an-/
*-nn-			/-nn-/		/-nn-/	
*-mp-	-	-	/-mp-/	/-mp-/	/-mp-/	/-mp-/

The variability between *a, *r and *n in the presyllable rime part of the alveolar nasal rime-onset infixes clearly shows that the alveolar nasal displacing the root onset and occupying the main syllable onset slot is the most salient characteristic of etyma derived by infixation from simplex onset monosyllables. However, the presyllable rime must nevertheless be occupied by some segment in order to produce a well-formed disyllable in Katuic. These three presyllable rime segments were all available to use, if not actually interchangeable, in PK, based on the persistence of all three forms in at least some of the modern languages. Note that there is one other theoretically possible rime-onset infix, *-ln-, but it is not attested in the modern languages and is therefore not reconstructable to PK.

The nominalizing infix /-an-/ attaches to simple monosyllables only and it is primarily, though not exclusively, used to derive the instrument or the object/result of the verbal action. This infix is found throughout Katuic, though it is indistinguishable from PK *-rn- 'NOM' and PK *-nn- 'NOM' in languages like EKatu which retain neither the PK gemination contrast nor PK liquid presyllable coda consonants. In EKatu then, all three alveolar nasal rime-onset nominalizing infixes are merged as /-an-/. In languages like

Ta'oiq and EBru which have lost the PK gemination contrast but have retained coda /r/, only *-an- and *-nn- are merged. Examples of PK *-an- 'NOM' in modern Katuic languages are presented in Table 14.

Table 14: *Examples supporting the reconstruction of PK *-an- 'NOM'*

EKatu				**WKatu**			
ciam	*to feed*	caniam	*food given*	ciəm	*to feed animals*	caniəm	*animal food*
teːŋ	*to work*	taneːŋ	*work*	teːŋ	*to work*	taneːŋ	*work*
ʔbec	*to sleep*	banec	*bed*	sok	*to be rich*	sanok	*possessions*
Kriang				**Ta'oiq**			
pəŋ	*to cast fishing net*	panəŋ	*fishing net*	caː	*to eat*	hanaː	*food*
pi̯ʔ	*to sleep*	pani̯ʔ	*bed*	pəːh	*to be wide*	panəːh	*width*
cɔh	*to make a hole*	canɔh	*a hole*	kaːnʔ	*to slice*	kanaːnʔ	*a slice*
Pacoh				**EBru**			
ku̯ːr	*to row a boat*	kanu̯ːr	*oar*	caː	*to eat*	sanaː	*food*
ku̯ah	*to shave*	kanu̯ah	*razor*	kaŋ	*to block*	kanaŋ	*an obstruction*
paŋ	*to cast a fish net*	pinaŋ	*fish net*	sɒk	*to scoop up*	sanɒk	*ladle, spoon*

The nominalizing infix /-rn-/ attaches to simple monosyllables only and it is primarily, though not exclusively, used to derive the instrument of the verbal action. Similar /-rn-/ infixes are found in languages from other Austroasiatic families such as Sre (Bahnaric) (Olsen 2015), Kri (Vietic) (Enfield and Diffloth 2009), Nyah Kur (Monic) (Diffloth 1984) and Kammu (Khmuic) (Svantesson 1983). Examples of PK *-rn- 'NOM' in modern Katuic languages are presented in Table 15.

Table 15: *Examples supporting the reconstruction of PK *-rn- 'NOM'*

EKatu				**WKatu**			
				tɛh	*to hammer*	trnɛh	*hammer*
				kuːk	*to wear a necklace*	krnuːk	*necklace*
				toŋ	*to tie*	trnoŋ	*rope, vine for tying*
Kriang				**Ta'oiq**			
tɛ̯h	*to hammer*	trnɛ̯h	*a hammer*	pi̯ːh	*to sweep*	prni̯ːh	*broom*
piːh	*to sweep*	prniːh	*broom*	taʔ	*to do*	ranaʔ	*work*
coh	*to plant*	crnoh	*crops*	coh	*to plant*	hrnoh	*field, land*
Pacoh				**EBru**			
tɛh	*to forge metal*	trnɛh	*hammer*	taʔ	*to do*	ranaʔ	*work*
cu̯h	*to plant*	trnu̯h	*all crops*	paːj	*to say*	prnaːj	*word, language*
tu̯ŋ	*to tie to*	trnu̯ŋ	*string*	cɔh	*to plant*	srnɔh	*crops*

The geminate infix /-nn-/ is primarily though not exclusively used to derive objects. It is found only in Pacoh, Kriang and WKatu, all of which continue to permit gemination of sonorants across the presyllable: main syllable boundary (see Section 4.2.2). However, the /-nn-/ infix is encountered far less frequently than the /-an-/ and /-rn-/ nominalizers even in those languages which retain gemination. External comparison of Katuic /-nn-/ with the nominalizing infix /-mn-/ which occurs in both Khmer (Jenner and Pou 1980–81:l–li) and Khmu (Svantesson 1983:98) tempts one to reconstruct this infix to PK. A pre-PK *-mn- nominalizer would have become PK *-nn- due to the constraint on PK presyllable nasal codas which required them to be

homorganic with the main syllable onset.[15] Given the paucity of examples, it remains quite possible that this infix was either borrowed or locally innovated, the latter option being perhaps more likely. Nevertheless, at this point I tentatively reconstruct *-nn- back to PK with the caveat that if it was a part of PK, it was used infrequently and was likely unproductive. Examples of PK *-nn- 'NOM' in modern Katuic languages are presented in Table 16.

Table 16: *Examples supporting the reconstruction of PK *-nn- 'NOM'*

EKatu				**WKatu**			
				cual	*to be strong*	tnnual	*strength*
Kriang				**Ta'oiq**			
cual	*to be strong*	cnnual	*strength*				
caː	*eat*	cnnaː	*food*				
mpɔː	*to dream*	pnnɔː	*a dream*				
Pacoh				**EBru**			
kaj	*to plow*	knnaj	*plowed ground*				
caː	*to eat*	tnnaː	*food*				
kaːŋ	*to shackle*	knnaːŋ	*pig-yoke*				

The labial nominalizing infix /-mp-/ occurs infrequently but it is well distributed outside of Katu. External comparison with other labial nominalizing infixes such as Khmer /-b-/ (Jenner and Pou 1980–81:xlvi) and Nyah Kur (Monic) /-w-/ (Diffloth 1984:263) strengthen the case for reconstructing *-mp- for PK.[16] I do so while acknowledging that, like *-nn-, if *-mp- existed in PK it was likely uncommon and unproductive. PK *-mp- is best preserved in Pacoh and EBru. Examples of PK *-mp- 'NOM' in modern Katuic languages are presented in Table 17.

Table 17: *Examples supporting the reconstruction of PK *-mp- 'NOM'*

EKatu				**WKatu**			
Kriang				**Ta'oiq**			
tuc	*to steal*	tmpuc	*thief*	taʔ	*to do*	tmpaʔ	*work*
Pacoh				**EBru**			
tan	*to observe taboo*	tmpan	*a taboo*	tɒk	*to wear clothes*	tmpɒk	*clothes*
taʔ	*to work, do, make*	tmpaʔ	*doings, work*	cɔːj	*to plant rice, dibble*	smpɔːj	*dibble stick*
tuk	*to bump against*	tmpuk	*trap*	rə̯ːh	*to pour out*	rmpə̯ːh	*foam, bubbles*

[15] In Watson et al.'s (2013) dictionary of Pacoh, we find two examples of an apparent /-mm-/ 'NOM' infix; /tuːn/ 'to follow' vs. /tmmuːn/ 'followers' and /toːŋ/ 'to say' vs. /tmmoːŋ/ 'language, conversation'. This perhaps points to a partial retention of a labial place of articulation for pre-PK *-mn- between alveolar /t/ in the presyllable onset and back, rounded vowels. Alternatively, /-mm-/ 'NOM' could be a reflex of PK *-mp- with weakening of the oral stop.

[16] The Khmer infix /-b-/ reconstructs to *p since prevocalic *p became /b/ [ɓ] in modern Khmer. Likewise, the Nyah Kur /-w-/ infix reconstructs to *p, since intervocalic *p become /w/ in Monic.

*5.3 PK Nominalizing Rime Infixes *-r- and *-N-*

Two nominalizing rime infixes, *-r- and *-N-, are reconstructable to PK. Both of these may occur in either complex onset monosyllables or disyllables. When a rime infix attaches to a complex onset monosyllable, the most salient marker of the derived etyma is the movement of the original main syllable onset out to the presyllable onset and the promotion of the root medial to the main syllable onset in the derived word. The infix itself constitutes the presyllable rime and we find variation between a vocalic infix /-a-/, a rhotic infix /-r-/, a geminate infix which assimilates to the root medial liquid and a nasal infix which always becomes alveolar since both of the Katuic medial consonants, /r/ and/l/, are alveolar. This nasal infix developed an excrescent alveolar stop in EBru before the rhotic medial but does not occur with the lateral medial. It is possible to condense these four synchronic affixes down to two nominalizing rime infixes for PK, a rhotic *-r- and a nasal *-N, based on the correspondences in Table 18.

Table 18: *Table of correspondences for PK nominalizing rime infixes *-r- and *-N- in complex onset monosyllables*

PK Infix	Root Medial	EKatu	WKatu	Kriang	Ta'oiq	Pacoh	EBru
*-r-	Rhotic	/-a-/	/-a-/	/-r-/	/-a-/	/-r-/	/-a-/
	Lateral	/-a-/	/-r-/	/-r-/	/-r-/	/-r-/	/-r-/
*-N-	Rhotic	/-a-/	/-a-/	/-r-/	/-a-/	/-r-/	/-nt-/
	Lateral	/-a-/	/-a-/	/-l-/	/-a-/	/-l-/	/-a-/

The *-r- infix before a rhotic medial remains as such in Kriang and Pacoh, forming a geminate. The other languages have all lost the PK gemination contrast (see Section 4.2.2) leading to vocalization of the PK *-r- infix to /-a-/ before the rhotic medial. Before medial /l/, the *-r- infix is preserved in all of the languages investigated here except EKatu, which has deleted presyllable coda liquids. In EKatu, PK *-r- vocalizes to /-a-/ in all environments. Examples of *-r- before lateral medials are presented in Table 19 but no examples of PK *-r- before rhotic medials are presented here. This is because the modern reflexes of PK *-r- are merged with and indistinguishable from reflexes of PK *-N- before rhotic medials in all languages except for EBru. As a result, we cannot be sure if the infix standing before a rhotic medial is a descendent of PK *-r- or PK *-N- without cognate EBru etyma to compare against. At this point, no such clear examples are available.

Table 19: *Examples supporting the reconstruction of PK *-r- 'NOM' before lateral medials*

EKatu				WKatu			
				pʰlɜːj	*to buy*	pʰrlɜːj	*something bought*
				kliaŋ	*to lock door with piece of wood*	krliaŋ	*piece of wood to lock door*
Kriang				**Ta'oiq**			
pluh	*to blow*	prluh	*blowpipe*	plaːs	*to make use of someone*	prlaːs	*orders, instructions*
klo̤ːs	*to chop up bamboo*	krlo̤ːs	*chopped bamboo pieces*	pləŋʔ	*to make a hole*	prləŋʔ	*a hole*
Pacoh				**EBru**			
klən	*to partition*	krlən	*boundary*	kluːm	*to urinate*	krluːm	*urine*
				ploŋ	*to blow to heal (shaman)*	prloŋ	*ceremony of blowing to heal*
				kliəŋ	*to tie, bind*	raliəŋ (<*krliəŋ)	*fetters*

The *-N- infix before a rhotic medial remains a nasal in EBru. In EBru and in West Katuic generally, sequences of *Nr developed an epenthetic stop consonant [t]. The epenthetic stop has since phonologized as a segment occupying the main syllable onset slot and forming a new tautosyllabic clustered onset with the root rhotic medial (*C-n-r > /Cntr/). This is an extension of a process that was current in PK, whereby syllabic nasals preceding rhotic main syllable onsets regularly developed an excrescent stop consonant (Sidwell 2005:32). Similar epenthetic processes triggered by nominalizing infixes have been described for Kammu (Svantesson 1983:97) and Bahnar (Banker 1964:104). Nasal infixation of onset clusters is no longer productive in West Katuic, but many frozen forms are evident, as the EBru examples in Table 20 demonstrate.

In the other languages, reflexes of PK *-N- before a rhotic medial have assimilated fully to the medial consonant. Kriang and Pacoh preserve the geminate sequence /rr/ across the syllable boundary but the other languages have subsequently vocalized these reflexes of the PK *-N- infix to /-a-/. This results in a total merger of PK *-r- and *-N- nominalizing infixes before medial /r/ outside of West Katuic. The examples in Table 20 demonstrate the preservation of PK *-N- 'NOM' in West Katuic taking examples from EBru. Note the rhoticization of the nasal in Pacoh and the subsequent vocalization of the rhotic to /a/ in WKatu.

Table 20: *Examples supporting the reconstruction of *-N- before rhotic medials*

PK Root	PK Derivation	WKatu	Pacoh	EBru
*grɔːŋ *to fence*	*gnrɔːŋ *a fence*	/karɔːŋ/ *a fence*	/krrɔːŋ/ *a fence*	/kntrụaŋ/ *a fence*
*kraːŋ *to carry between two people on a pole*	*knraːŋ *stretcher, pallet*	/karaːŋ/ *stretcher*	/krraːŋ/ *stretcher*	/kntraːŋ/ *pallet*
**krias *to scratch*[17]	*knrias *fingernail*	/karias/ *fingernail*	/krrias/ *fingernail*	/kntrę̣ːh/ *fingernail*

I have found no evidence of the *-N- infix being preserved before the lateral medial in EBru. Instead, it would appear that in all languages the *-N- infix assimilated to the lateral medial, after which point it became geminate as it remains in the languages which allow it (Kriang and Pacoh) or vocalized to /-a-/ in the languages which do not. Table 21 presents examples of PK *-N- before lateral medials.

Table 21: *Examples supporting the reconstruction of PK *-N- 'NOM' before lateral medials*

EKatu				WKatu			
				klɔːs	*to exchange*	kalɔːs	*an exchange*
				klɜn	*to block a road*	kalɜn	*a blockade*
				kliəp	*to patch*	kaliəp	*a patch*
Kriang				**Ta'oiq**			
klaːk	*to diverge*	kllaːk	*fork in trail*	klaːr	*to dig*	kalaːr	*scar*
kliaŋ	*to bar a door*	klliaŋ	*bar to close door*				
plɛːw	*to puncture*	pllɛːw	*hole*				
Pacoh				**EBru**			
klaːŋ	*to pipe water*	kllaːŋ	*bamboo water pipe*	plụak	*to be gray*	palụak	*tusks, ivory*
klu̱ːn	*to play*	kllu̱ːn	*sports, games*	kla̱ːn	*to separate, divide*	kala̱ːn	*border between rice fields*
plu̱ːt	*to set (sun)*	pllu̱ːt	*west*				

[17] From an unattested, Pre-Katuic root **krias (cf. Bahnar /krɛh/ 'to scratch' (Banker 1979))

Moving on from complex onset monosyllables to disyllables, reflexes of PK *-N- 'NOM' are found in disyllables in EBru, Pacoh and Ta'oiq (see examples in Table 23) but this infix was apparently lost in disyllables in Kriang and WKatu, both of which use only /-r-/ 'NOM' in this environment. As for PK *-r- 'NOM', we find reflexes of this infix in disyllables in all the languages which retain presyllable liquid codas, as the examples in Table 22 demonstrate. Note that because sequences of nasal + rhotic become geminate rhotics in Pacoh and Kriang, it is not possible to determine whether *-r- or *-N- are being used in disyllables with /r/ in the main syllable onset such as Pacoh /taroː/ 'to shine' - /trroː/ 'light' (< *trroː or *tnroː).

Table 22: *Examples supporting the reconstruction of PK *-r- 'NOM' in disyllables*

EKatu				**WKatu**			
				katas	*to name*	krtas	*a name*
				ʔacia	*to give*	ʔrcia	*a gift*
				tapɨːŋ	*to put a roof on*	trpɨːŋ	*a roof*
Kriang				**Ta'oiq**			
kawaːŋ	*to surround*	krwaːŋ	*area, region*	kajoːm	*to wrap smth up*	krjoːm	*package*
patah	*to abandon*	prtah	*divorce*	kawa̱ːŋ	*to envelop, surround*	krwa̱ːŋ	*territory*
taŋaw	*to sit down*	trŋaw	*a seat*	pali̱ːh	*to treat, take care of*	prli̱ːh	*medicine*
Pacoh				**EBru**			
taŋɯh	*to breathe*	trŋɯh	*breath*	kahaːk	*to spit*	krhaːk	*saliva*
takɤwʔ	*to cut with scissors*	trkɤwʔ	*scissors*	palaːj	*to treat, cure*	prlaːj	*medicine*
kataw	*to warm (food)*	krtaw	*warmed food*	sapɔː	*to make a roof*	srpɔː	*a roof*

Table 23: *Examples supporting the reconstruction of PK *-N- 'NOM' in disyllables*

EKatu				**WKatu**			
Kriang				**Ta'oiq**			
				takaːmˀ	*to eat with chopsticks*	tŋkaːmˀ	*chopsticks*
				kacik	*to comb*	kɲcik	*comb*
				kataŋˀ	*to close*	kntaŋˀ	*door*
Pacoh				**EBru**			
tikap	*to pinch with chopsticks*	tŋkap	*chopsticks*	takap	*to grasp with tongs*	tŋkap	*tweezers*
kanaj	*to plow*	knnaj	*plowed ground*	kasaŋ	*to spin thread*	knsaŋ	*thread*
ʔaniə	*to be extra*	ʔnniə	*surplus*	kajɜm	*to rest one's head*	knjɜm	*pillow*

*5.4 PK Reciprocal Prefix *tr- and Rime Infix *-r-*

Table 24: *Correspondences for PK reciprocal prefix *tr- and infix *-r-*

PK	EKatu	WKatu	Kriang	Ta'oiq	Pacoh	EBru
*tr- (+/- Redup.)	/ta-/	/tr-/ (+/- Redup.)	/tr-/ (+/- Redup.)	/tr-/ (+/- Redup.)	/tr-/ (+/- Redup.)	/ra-/
*-r-	-	-	/-r-/	/-r-/	/-r-/	/-r-/

The Katuic reciprocal construction decreases the root verb's valency by one by moving the object of the verb into a compound subject. The plural subjects then perform the verbal action either on each other or simply simultaneously, so that "SVO" becomes "{S&O}V each other" or "{S&O}V together".

A reciprocal prefix *tr- is clearly reconstructable to PK and a reciprocal infix *-r- is also indicated. The PK form with a coda *r is cognate with Kammu *tr- 'RECIP' (Svantesson 1983:111–12) and with reciprocal /ta-/ prefixes in the Bahnaric languages Bahnar (Banker 1964), Sedang (Smith and Sidwell 2015) and Jeh (Gradin 1976). Two other Bahnaric languages have a reciprocal marker /təm/ which Butler (2015b) describes as being either a prefix or a free standing, preverbal particle in Bunong and Olsen (2015) describes as a proclitic in Kơho-Sre. The /təm/ reciprocal may be unrelated to the /tr-/ prefixes but the shared /t/ initial is suggestive of a connection. All of these forms with initial /t/ would appear to be cognate with Sidwell's (2008) PAA *t-, *tN- 'RECIP'. A connection with Khmer /pr-/ [pra-] 'RECIP', which is attested as far back as Old Khmer, is also likely given the presence of the thematic /r/ coda (Jenner and Pou 1980–81:xxxv). An ultimate connection with Proto-Austronesian prefix *paRi- 'reciprocal/collective action' is possible (Blust 2013:380).

Outside of EKatu and EBru, modern reflexes of PK *tr- 'RECIP' are marked by an optionally reduplicated root verb. Kriang, Ta'oiq and Pacoh have the construction (/tr+ROOT ROOT/) with the prefix on the first utterance of the verb root whereas WKatu reverses this, putting the prefix on the second utterance (/ROOT tr+ROOT/). Examples of this are provided in Table 25.

Table 25: *Examples supporting the reconstruction of PK *tr- 'RECIP' ("e.o." = "each other")*

EKatu				**WKatu**			
ɲər	*to love*	taɲər	*to love e.o.*	heːl	*to love*	heːl trheːl	*to love e.o.*
nal	*to know*	tanal	*to know e.o.*	waʔ	*to borrow*	waʔ trwaʔ	*to borrow from e.o.*
lej	*to see*	talej	*to see e.o.*	lət	*to do wrong*	trlət	*to wrong e.o.*
Kriang				**Ta'oiq**			
cɔːm	*to know*	trcɔːm cɔːm	*to know e.o.*	kɔːmˀ	*to grasp*	rakɔːmˀ tiː	*to shake hand*
hɛːl	*to love*	trhɛːl hɛːl	*to love e.o.*	cəl	*to fight, attack*	racəl cəl	*to fight e.o.*
jɔːʔ	*to go*	trjɔːʔ jɔːʔ	*to go together*	təm	*to punch*	trtəm təm	*to box e.o.*
Pacoh				**EBru**			
cɔːm	*to know*	trcɔːm cɔːm	*to know e.o.*	cu̯aj	*to help*	racu̯aj	*to help e.o.*
ʔat	*to stay*	trʔat ʔat	*to stay with e.o.*	pu̯aj	*to chase*	rapu̯aj	*to chase e.o.*
pok	*to go*	trpok pok	*to go to e.o.*	huːn	*to kiss*	rahuːn	*to kiss e.o.*

A reciprocal rime infix which may attach to complex onset monosyllables or disyllables is found in Kriang, Ta'oiq, Pacoh and EBru but not in either of the Katu varieties researched here. Examples of the reciprocal infix are presented in Table 26.

Table 26: *Examples supporting the reconstruction of PK *-r- 'RECIP' ("e.o." = "each other")*

EKatu				**WKatu**			
Kriang				**Ta'oiq**			
prɻ̰ːm	*to agree w/ smb*	*prrɻ̰ːm*	*to agree w/ e.o.*	kadą̰ːh	*to be ashamed*	krdą̰ːh	*to be ashamed of e.o.*
krɔŋ	*to be angry*	*krrɔŋ*	*to argue*	kalə̰ːs	*to say goodbye*	krlə̰ːs lə̰ːs	*to bid e.o. farewell*
tapuːn	*to follow (intr)*	*trpuːn*	*to follow (tr)*				
Pacoh				**EBru**			
trɯm	*to wrestle*	trrɯm	*to wrestle e.o.*	saɲɨː	*to miss*	srɲɨː	*to miss e.o.*
patam	*to advise, remind*	prtam	*to tell e.o.*	kaciːt	*to kill*	krciːt	*to kill e.o.*
ʔawoːj	*to give, bring*	ʔrwoːj	*to receive from e.o.*	tamɔh	*to meet*	ramɔh	*to meet e.o.*

*5.5 PK Anticausative Prefixes *tr- & *sr-*

Table 27: *Correspondences for PK anticausative prefixes *tr- and *sr-*

PK	EKatu	WKatu	Kriang	Ta'oiq	Pacoh	EBru
*tr-	/ta-/	/tr-/	/tr-/	/ra-/	/tr-/, /tV-/	/ra-/
*sr-	/ha-/	/sr-/	-	/hr-/, /ha-/	-	/sr-/, /sa-/

The Katuic anticausative construction reduces the valency of the root verb by removing the syntactic subject NP (prototypically a semantic agent or experiencer) and moving the syntactic object NP into the subject position. This typically has the effect of deriving an intransitive, stative verb from a transitive, active verb root. Two anticausative prefixes are reconstructable for PK, *tr- and *sr-. Note that the rhotic in the presyllable coda position is thematic for the Katuic anticausative and an anticausative rime infix /-r-/ appears in certain Katuic languages though there is not enough evidence to reconstruct this infix back to PK (see Section 5.7.4).

The PK *tr- 'ANTICAUS' prefix is formally equivalent to PK *tr- 'RECIP'. These two prefixes also serve similar functions in that both prefixes move the root verb's original syntactic object into the subject position. Similar anticausative prefixes have been described in Bahnaric languages including, for example, the *passive prefix* /tə-/ in Bahnar (Banker 1964) and the *intransitive prefix* /ta-/ in Jeh (Gradin 1976).[18] All of these also bear resemblance to the so called *involuntary prefix* /ta-/, which has been described for EKatu (Costello 1966) and Pacoh (S. Watson 1966) but also appears in Chamic and Bahnaric languages. This involuntary prefix is formally similar to the reciprocal and anticausative prefixes and also performs a similar role in that it diminishes the agency involved in the verbal action. As Thurgood (1999:239–41) discusses, all three of these prefix categories in /t/ in Katuic, Bahnaric and Chamic may ultimately be related to the Proto-Austronesian prefix *taR- 'sudden, unexpected or accidental action' (Blust 2013, 382). The inadvertent prefixes of Katu and Pacoh at any rate have clear parallels in Chamic indicating that they were borrowed from Austronesian but the original source of *tr- 'ANTICAUS' and *tr- 'RECIP' is perhaps more likely Austroasiatic given the parallels with Kammu *tr- 'RECIP' (Svantesson 1983).

[18] A connection with the Old Khmer passivizing particle /tiː/ is also possible (Jenny et al. 2015:106).

S. Watson (1966) reports the Pacoh anticausative prefix as /ti-/.[19] In my own analysis of the lexical data available in Watson et al.'s (2013) Pacoh dictionary, I find examples of /ti-/, /ta-/, /tu-/ and /tr-/ all marking anticausatives and see no reason so give special prominence to the /ti-/ form. That being said, it is true that the *r coda is no longer seen in many of modern Pacoh's anticausatives due to the innovation of a new opposition of presyllable vowel qualities in which /a/ has become thematic for causative and/or transitive while the other two qualities, /i, u/, have become thematic for anticausative and/or intransitive (Gehrmann 2017a) (see Section 5.7.2). This new apophonic strategy for marking transitivity seems to have allowed Pacoh speakers to disambiguate /tr-/ 'RECIP' from /tr-/ 'ANTICAUS', though examples of /tr-/ 'ANTICAUS' do remain, as can be seen in the examples in Table 28.

Table 28: *Examples supporting the reconstruction of PK *tr- 'ANTICAUS'*

EKatu				**WKatu**			
rɔh	*to burn (tr)*	tarɔh	*burnt*	pʰoc	*to pull*	trpʰoc	*pulled out*
ʔih	*to sew*	taʔih	*sewn*	wət	*to get rid of*	trwət	*got rid of*
palaːŋ	*to turn smth over*	tapalaːŋ	*turned over*	ʔboːn	*to acquire*	trʔboːn	*acquired*
Kriang				**Ta'oiq**			
klah	*to break (tr)*	trklah	*to break (intr)*	pɔʔ	*to strip, peel (tr)*	rapɔʔ	*to peel off (intr), come off*
Pacoh				**EBru**			
piːl	*pull out hair, feathers*	trpiːl	*shed; molt*	kluʔ	*crush, squash*	rakluʔ	*crushed*
wiəʔ	*twist, wring*	tiwiəʔ	*twisted, wrung*	haʔ	*tear*	rahaʔ	*torn*
poh	*to open (tr)*	tupoh	*to be open, uncovered*	ciːk	*write (lit. scratch)*	raciːk	*scratched*

Table 29: *Examples supporting the reconstruction of PK *sr- 'ANTICAUS'*

EKatu				**WKatu**			
ʔul	*to hunger*	haʔul	*to be made hungry*	teːk	*to tear*	srteːk	*already torn*
				lɔː	*to ruin*	srlɔː	*already ruined, destroyed*
				klɔːc	*to finish*	srklɔːc	*already finished (work, life)*
Kriang				**Ta'oiq**			
				poːc	*to pluck, pull out*	hrpoːc	*to fall out*
				pəːh	*to open (tr)*	hrpəːh	*to open (intr)*
				lih	*to untie, undo*	hrlih	*to come untied, undone*
Pacoh				**EBru**			
				ʔabaːl	*to illuminate*	srbaːl	*to be blinded by light*
				ʔaloʔ	*to dip into*	srloʔ	*to fall into*
				piːh	*to break*	sapiːh	*to be broken*

[19] S. Watson (1966) calls this a *resultant state* prefix.

A second PK anticausative prefix is also indicated (Table 29), though the evidence is a bit weaker. This prefix would have had the form *sr- and featured the same thematic /r/ presyllable coda found in PK *tr- 'ANTICAUS'. Reflexes of *sr have not been forthcoming in Kriang or Pacoh but this marker is quite common in WKatu, Ta'oiq and EBru. I have been unable to find any phonetically similar anticausative prefixes outside of Katuic, which indicates that PK *sr- was probably a Katuic innovation.

*5.6 PK Causative Prefixes *pa-, *ta- & *sa-*

The Katuic causative is used with both intransitive and transitive verbs to add an agent or causer argument for the verbal action. Valency increasing morphology is also found throughout the Austroasiatic language family and beyond (cf. discussion in Reid (1994), Sidwell (2008)). There is good evidence for the reconstruction of three causative prefixes (*pa-, *ta-, *sa-) in PK.

The famous labial causative prefix, which is known throughout Austroasiatic and Austronesian, is also found in the modern Katuic languages and was clearly a part of the morphological system of PK. The PK causative prefix *pa- and its modern reflexes attach to both simplex and complex onset monosyllables. It is found with a vocalic presyllable rime in all languages, but both Kriang and Ta'oiq frequently have a variant form /pN-/ with a nasal presyllable coda. This is an innovation, which probably developed by analogy with a causative infix, /-N-/, which is commonly employed in Kriang and Ta'oiq. Pacoh reflexes of *pa- are commonly found with an /i/ presyllable vowel, which is the typical vowel quality associated with /p/ presyllable onsets in that language (Gehrmann 2017a). This demonstrates that the causative forms built from *pa- in Pacoh had already been derived at the time when Pacoh developed presyllable vowel contrasts. Note that there are very few modern reflexes of *pa- in EBru, which has innovated another prefix, /ʔa-/, for marking causatives. Examples of PK *pa- 'CAUS' in modern Katuic languages are presented in Table 30.

Table 30: *Examples supporting the reconstruction of PK *pa- 'CAUS'*

EKatu				**WKatu**			
moːp	*to be bad*	pamoːp	*to make bad*	coː	*to return home*	pacoː	*to give sth back*
tam	*to be black*	patam	*to make black*	luəs	*to be free*	paluəs	*to set free*
cariat	*to be cold*	pacariat	*to make cold*	krɛː	*to be right*	pikrɛː	*to make right*
Kriang				**Ta'oiq**			
jn̥ːʔ	*to go*	pajn̥ːʔ	*to send*	sn̥ːh	*to ascend*	pasn̥ːh	*to lift up*
tṵː	*to move house*	patṵː	*to chase away*	hum	*to bathe (intr)*	pahum	*to bathe (tr)*
sək	*to ascend*	pasək	*to take sth up*	toʔ	*to be upside down*	patoʔ	*to turn sth upside down*
Pacoh				**EBru**			
tḭh	*to be tight, taut*	pitḭh	*to tighten*	dɒh	*to explode*	padɒh	*to cause to explode*
cḭːn	*to reconcile*	picḭːn	*to mediate*	rn̥ːm	*to meet*	parn̥ːm	*to collect*
ʔoːjˀ	*to crave, desire*	paʔoːjˀ	*to tempt*	tn̥ːʔ	*to arrive*	patn̥ːʔ	*to make arrive*

PK *pa- 'CAUS' did not attach to roots with labial main syllable onsets. This is clearly demonstrated in the modern languages, in which one of two alternative causative prefixes occurs before labial onset roots in almost all cases. Table 31 summarizes the distribution of the causative prefixes in the Katuic languages analyzed in this paper. Parentheses around a prefix indicate that that prefix is found only rarely.

Table 31: *Correspondences for PK causative prefixes *pa-, *ta- and *sa-*

Root Onset	PK	EKatu	WKatu	Kriang	Ta'oiq	Pacoh	EBru
Labial	*pa-	/pa-/	/pa-/, /pi-/	/pa-/, /pN-/	/pa-/, /pN-/	/pi-/, /pa-/	(/pa-/)
	-	-	-	-	-	-	/ʔa-/
Non-Labial	*ta-	/ta-/	/ta-/	(/tN-/)	(/tN-/)	/ta-/	(/ta-/)
	*sa-	(/ha-/)	(/sa-/)	/sN-/	/hN-/	-	/sa-/

The causative prefix with the /t/ onset appears to be the original PK causative prefix for labial onset roots, as demonstrated by its presence in all six languages examined here. Examples of *ta- 'CAUS' are presented in Table 32.

Table 32: *Examples supporting the reconstruction of PK *ta- 'CAUS'*

EKatu				**WKatu**			
bral	*to arrive*	tabral	*to cause to arrive*	ʔboːn	*to get, catch*	taʔboːn	*to cause to get, catch*
plak	*to break (intr)*	taplak	*to cause to break*	mɔt	*to enter*	tamɔt	*to cause to enter*
mut	*to run*	tamut	*to cause to run*	pɔŋ	*to go down*	tapɔŋ	*to cause to go down*
Kriang				**Ta'oiq**			
mo̤ːt	*to enter*	tmmo̤ːt	*to bring smth in*	pɨn	*to follow (intr)*	tmpɨn	*to follow (tr)*
puːn	*to follow (intr)*	tmpuːn	*to follow (tr)*				
Pacoh				**EBru**			
boːn	*to have*	taboːn	*to get*	kap	*to bite*	takap	*to grasp (as with tongs)*
mṳːt	*to enter*	tamṳːt	*to take smth in*				
maʔ	*to bear*	tamaʔ	*to put smth in*				

Though reflexes of PK *ta- 'CAUS' are still found in Kriang, Ta'oiq and EBru, another causative prefix with an /s/ onset is more common in these three languages before labial main syllable onsets. Similar causatives in /s/ are also found in both WKatu and EKatu.[20] This prefix appears to be reconstructable to PK as well, though it is admittedly less common. Note that the *s presyllable onset is reduced to /h-/ in some modern Katuic languages (Gehrmann and Conver 2015). No trace of this prefix is found in Pacoh, in which PK presyllable initial *s has been deleted. Examples of PK *sa- in modern Katuic languages are presented in Table 33.

[20] Costello (1966) and Nguyễn (1995) both describe valency-altering /ha-/ prefixes in Eastern Katu. Both authors attempt to identify a unifying, unitary purpose for the /ha-/ prefix but based on their examples, the prefix encodes causatives in certain examples and anticausatives in others. This is a result of both PK *sa- 'CAUS' and PK *sr- 'ANTICAUS' prefixes having merged to /ha-/ in Eastern Katu.

Table 33: *Examples supporting the reconstruction of PK *sa- 'CAUS'*

EKatu				**WKatu**			
cɛn	*to cook*	hacɛn	*to cause to be cooked*	paːɲ	*to divide out*	sapaːɲ	*to order to divide out*
jur	*to rise*	hajur	*to cause to be raised*				
lɔːʔ	*to peel*	halɔːʔ	*to cause to be peeled*				
Kriang				**Ta'oiq**			
bɔːk	*to be white*	smbɔːk	*to whiten*	par	*to fly*	hmpar	*to fly a kite*
mo̤ŋ	*to be alive*	smmo̤ŋ	*to bring smb back to life*	bak	*to be injured, hurt*	hmbak	*to injure, hurt smb*
pat	*to go out (fire)*	smpat	*to extinguish smth*	moːt	*to enter*	hamoːt	*to insert*
Pacoh				**EBru**			
				pɜr	*to fly*	sapɜr	*to cause to fly*
				pah rapa̤ːŋ	*to view from a distance*	sapah	*to show*

5.7 Other Notable Derivational Affixes

A number of other derivational affixes in modern Katuic languages are worth describing briefly here, even though they do not reconstruct back to PK. These include the Bru causative prefix /ʔa-/, the Pacoh and Bru causative rime infix /-a-/, the Kriang and Ta'oiq causative rime infix /-N-/ and the anticausative rime infix /-r-/.

5.7.1 Bru Causative Prefix /ʔa-/

As mentioned above, /ʔa-/ is the most common causative prefix in modern Bru languages and it has mostly replaced the original PK causative prefix *pa-. Unlike other affixes in Bru, /ʔa-/ is largely productive in the modern language. It has the unusual property of being able to supplant the etymological presyllable of a disyllabic root as can be seen in one of the examples in Table 34. This prefix is clearly not reconstructable to PK but can be securely reconstructed to Proto-Bru based on its appearance in all modern Bru varieties. It would seem to have developed based on its phonetic parallel with the apophonic /-a-/ causative infix of Bru and Pacoh, but Pacoh did not develop this prefix and it remains a uniquely Bru feature.

Table 34: *Examples of /ʔa-/ 'CAUS' in EBru*

EBru			
coːn	*to go up*	ʔacoːn	*to lift up*
lo̤ːh	*to exit*	ʔalo̤ːh	*to take out*
padeh	*to lie down*	ʔadeh	*to cause to lie down*

5.7.2 Pacoh and Bru Causative Rime Infix /-a-/

Bru and Pacoh have developed a presyllable vowel quality opposition in disyllabic etyma in which /a/ is associated with causatives and/or transitive verbs and /i, u/ are associated with anticausatives and/or intransitive verbs (Gehrmann 2017a). This opposition is more robust in Pacoh, which has retained a contrast of presyllable vowels /a, i, u/ in various environments but has become quite marginal in Bru, where only

/ka-/ and /ku-/ remain contrastive in most varieties and even these contrasts have disappeared in some. Table 35 provides examples of causative vowel infix /-a-/ in Pacoh and EBru.

Table 35: *Examples of /-a-/ 'CAUS' infix in Pacoh and EBru*

Pacoh				**EBru**			
kucn̰ːt	*to die*	kacn̰ːt	*kill*	kuciːt	*to die*	kaciːt	*to kill*
tupujʔ	*to lack, be insufficient*	tapujʔ	*to use up something*	kuluh	*to flood (intr)*	kaluh	*to let out water*
pitoʔ	*to come towards*	patoʔ	*to go up to*	kutɜw	*to be hot*	katɜw	*to heat smth up*

Note that when a rime infixes takes the place of the root presyllable rime in a disyllable, the directionality of the affixation as I have analyzed it here could potentially be reversed. For example, what I have labeled as a causative /-a-/ infix in Bru /kaciːt/ 'to kill' could instead be conceived of as part of the root form. In that case, an anticausative /-u-/ could be posited to explain /kuciːt/ 'to die'. Therefore, in cases such as this, determining which is the root form and which is the derived form is a historical question rather than a morphological one.

*5.7.3 Kriang and Ta'oiq Causative Rime Infix *-N-*

Kriang and Ta'oiq often use a nasal rime infix /-N-/ to derive causatives from disyllabic roots. This infix is not found in the other modern Katuic languages and would appear to be an innovation. Table 36 provides examples of the /-N-/ causative.

Table 36: *Examples of /-N-/ 'CAUS' infix in Kriang and Ta'oiq*

Kriang				**Ta'oiq**			
hatn̰h	*to fall*	hntn̰h	*to drop*	padn̰ːh	*to fall over*	pndn̰ːh	*to push, knock over*
takn̰h	*to break (intr)*	tŋkn̰h	*to break (tr)*	takn̰ːs	*to break (intr)*	tŋkn̰ːs	*to break (tr)*
kaciːt	*to die*	kɲciːt	*to kill*	kaceːt	*to die*	kɲceːt	*to kill*

5.7.4 Anticausative Rime Infix /-r-/

Outside of Katu, we find a modest number of examples of anticausatives formed by a rhotic infix /-r-/. It remains possible that this infix was a part of PK but with so few examples, it cannot be determined with any certainty at this time. Some examples are presented in Table 37.

Table 37: *Examples of /-r-/ 'ANTICAUS' infix*

EKatu				**WKatu**			
Kriang				**Ta'oiq**			
craːn	*to mark, write*	crraːn	*to be striped*	takɨːl	*to rest head (on pillow)*	trkɨːl	*to support a head (pillow)*
pl̰ak	*to turn smth over*	prl̰ak	*to flip over (intr)*				
Pacoh				**EBru**			
paŋaː	*to cause to become rich*	prŋaː	*to be rich and esteemed*	katn̰ːʔ	*to hide*	krtn̰ːʔ	*to be hidden*
				palət	*to turn upside down*	prlət	*flip over, turn over*
				patɜp	*to appoint*	prtɜp	*to be appointed*

6 Conclusions

In conclusion, modern Katuic languages have by and large preserved the disyllabic character of the PK word canon. The contrastivity of sesquisyllabicity is maintained outside of Kuay, where Khmer contact has influenced the loss of this contrast and the development of phonological sesquisyllables in at least some varieties. Disyllabicity was even reinforced in the eastern range of Katuic by the development of presyllable vowel quality contrasts. Nevertheless, there is evidence that the Katuic languages have not been immune to the inexorable typological drift towards sesquisyllabicity and eventual monosyllabicity in Southeast Asia. This can be seen in the near loss of these same presyllable vowel quality contrasts in Bru and Katu and in the general erosion of the presyllable rime which has affected every modern Katuic language, except perhaps for Pacoh.

The derivational morphological system of PK was almost without a doubt more productive than those of its modern descendants, some of which retain only fossilized evidence of such a system. Nevertheless, morphological derivation remains common and partially productive in the central, mountainous areas where Pacoh, WKatu and Kriang are spoken even as contact with more strictly isolating prestige languages has led to a leveling of the PK morphophonological paradigm in favor of syntactic strategies in the more peripheral areas. Certain PK affixes were retained from PAA, including nominalizing infixes (*-an-, *-rn-, *-nn-, *-mp-, *-r-, *-N-) the reciprocal prefix (*tr-) and the causative prefix (*pa-). Others were either PK internal innovations or were borrowed in through contact with other prestige languages such as Chamic or Khmer.

The findings of this paper have implications for the morphological reconstruction of PAA and for our understanding of the historical contact between the Katuic languages and their geographical neighbors in the Bahnaric and Vietic sub-groups of AA, in Khmer and in Chamic. It is hoped that this paper will make a small contribution in these areas of ongoing linguistic research.

References

Alves, Mark. 2006. *A grammar of Pacoh: A Mon-Khmer language of the Central Highlands of Vietnam*. Canberra: Pacific Linguistics.

Alves, Mark. 2007. Pacoh pronouns and grammaticalization clines. In *SEALSXIII: Papers from the Thirteenth Meeting of the Southeast Asian Linguistics Society 2003*, ed. by Shoichi Iwasaki, Andrew Simpson, Karen Adams, and Paul Sidwell, 1–12. Canberra: Pacific Linguistics.

Alves, Mark. 2014. A survey of derivational morphology in the Mon-Khmer language family. In *The Oxford Handbook of Derivational Morphology*, ed. by Rochelle Lieber and Pavel Stekauer, 520–44. Oxford: Oxford University Press.

Banker, Elizabeth. Bahnar affixation. *Mon-Khmer Studies* 1:99–117.

Blust, Robert. 2013. *The Austronesian languages: Revised edition*. Canberra: Asia-Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Butler, Becky. 2015a. Approaching a phonological understanding of the sesquisyllable with phonetic evidence from Khmer and Bunong. In *The Languages of Mainland Southeast Asia: The State of the Art*, ed. by N.J. Enfield and Bernard Comrie 443–99. Berlin: De Gruyter Mouton.

Butler, Becky. 2015b. Bunong. *The Handbook of Austroasiatic Languages*, ed. by In Matthias Jenny and Paul Sidwell, 719–45. Leiden: Brill.

Conver, Johanna, Mackenzie Conver, and Jonathan Schmutz. 2017. *Lexicon of Ta'oiq*. Unpublished.

Costello, Nancy. 1966. Affixes in Katu. *Mon-Khmer Studies* 1:63–86.

Costello, Nancy. 1971. *Ngữ-Vựng Katu*. [Katu Vocabulary]. Saigon: Trung Tâm Học-Liệu: Bộ Giáo-Dục.

Costello, Nancy. 1998. Affixes in Katu of the Lao P.D.R. *Mon-Khmer Studies* 28:31–42.

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

Diffloth, Gérard. 1984. *The Dvaravati-Old Mon language and Nyah Kur*. Monic Language Studies 1. Bangkok: Chulalongkorn University Printing House.

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

Ferlus, Michel. 1971. Simplification des groupes consonantiques dans deux dialectes Austroasiens du Sud-Laos. [The simplification of consonant clusters in two Austroasiatic dialects of Southern Laos]. *Bulletin de la Societé de Linguistique de Paris* 66:389–403.

Ferlus, Michel. 1971. La langue Souei: Mutations consonantiques et bipartition du système vocalique. [The Souei language: Consonant mutations and the splitting of its vowel system]. *Bulletin de la Société de Linguistique de Paris* 66:379–88.

Ferlus, Michel. 1974a. Délimitation des groupes linguistiques Austroasiatiques dans le centre Indochinois. [Determining the Austroasiatic language groups in central Indochina]. *Asie du Sud-Est et Monde Insulindien* 5:15–23.

Ferlus, Michel. 1974b. La langue Ong, mutations consonantiques et transphonologisations. [The Ong language, consonant mutations and trasphonologizations]. *Asie du Sud-Est et 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

Gehrmann, Ryan and Johanna Conver. 2015. Katuic phonological features. *Mon-Khmer Studies* 44:lv–lxvii.

Gehrmann, Ryan, Feikje van der Haak, and Jennifer Engelkemier. 2016. *Lexicon of Kriang Tha Taeng*. Unpublished.

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*. MA thesis. Chiang Mai: Payap University.

Gehrmann, Ryan. 2017a. *Katuic presyllable vowel contrasts*. Paper presented at the 7th International Conference on Austroasiatic Linguistics, Christian-Albrechts Universität, Kiel, Sept. 29 – Oct. 1 2017.

Gehrmann, Ryan. 2017b. The historical phonology of Kriang, a Katuic language. *Journal of the Southeast Asian Linguistics Society* 10(1):114–139.

Gradin, Dwight. 1976. Word affixation in Jeh. *Mon-Khmer Studies* 5:25–42.

Henderson, Eugenie. 1952. The main features of Cambodia pronunciation. *Bulletin of the School of Oriental and African Studies* 14(1):149–174.

Hoàng Văn Hành and Tạ Văn Thông. 1998. *Tiếng Bru-Vân Kiêu*. [The Bru-Van Kieu Language]. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.

Huffman, Franklin. 1972. The boundary between the monosyllable and the disyllable in Cambodian. *Lingua* 29:54–66.

Huffman, Franklin. 1976. The register problem in fifteen Mon-Khmer languages. In *Austroasiatic Studies*, ed. by Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta, 575–90. Honolulu: The University Press of Hawaii.

Jenner, Philip and Saveros Pou. 1980–81. A lexicon of Khmer morphology. *Mon-Khmer Studies* 9–10.

Jenny, Matthias, Tobias Weber and Rachel Weymuth. 2015. The Austroasiatic languages: A typological overview. In *The Handbook of Austroasiatic Languages*, ed. by Matthias Jenny and Paul Sidwell, 13–143. Leiden: Brill.

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

Matisoff, James. 1973. Tonogenesis in Southeast Asia. In *Consonant Types and Tone*, ed. by L.M. Hyman, 71–96. Southern California Occasional Papers in Linguistics 1. Los Angeles: University of Southern California.

Miller, John and Carolyn Miller. 2017. *Bru-English-Vietnamese-Lao Dictionary*. SIL International. Online: http://bru.webonary.org/.

Olsen, Niel H. 2015. Kơho-Sre. Austroasiatic comparative-historical reconstruction. In *The Handbook of Austroasiatic Languages*, ed. by Matthias Jenny and Paul Sidwell, 746–88. Leiden: Brill.

Pittayaporn, Pittayawat. 2015. Typologizing sesquisyllabicity. In *The Languages of Mainland Southeast Asia: The State of the Art.* ed. by N.J. Enfield and Bernard Comrie, 500–28. Berlin: De Gruyter Mouton.

Reid, Lawrence. 1994. Morphological evidence for Austric. *Oceanic Linguistics* 33(2):323–44.

Sidwell, Paul and Felix Rau. 2015. Austroasiatic comparative-historical reconstruction. In *The Handbook of Austroasiatic Languages*, ed. by Matthias Jenny and Paul Sidwell, 221–363. Leiden: Brill.

Sidwell, Paul and Pascale Jacq. 2003. *A Handbook of Comparative Bahnaric, Vol. 1: West Bahnaric.* Pacific Linguistics 551. Canberra: Pacific Linguistics.

Sidwell, Paul. 2005. *The Katuic languages: Classification, reconstruction and comparative lexicon.* Studies in Asian Linguistics 58. Munich: Lincom Europa.

Sidwell, Paul. 2008. Issues in the morphological reconstruction of Proto-Mon-Khmer. In *Morphology and Language History: In Honor of Harold Koch*, ed. by Claire Bowern, Bethwyn Evans, and Luisa Miceli, 251–65. Amsterdam: John Benjamins.

Sidwell, Paul. 2009. *Classifying the Austroasiatic languages: History and state of the art*. Munich: Lincom Europa.

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

Sidwell, Paul. 2015. Austroasiatic classification. In *The Handbook of Austroasiatic Languages*, ed. by Matthias Jenny and Paul Sidwell,144–220. Leiden: Brill.

Smith, Kenneth and Paul Sidwell. 2015. Sedang. In *The Handbook of Austroasiatic Languages*, ed. by Matthias Jenny and Paul Sidwell, 789–836. Leiden: Brill.

Smith, Ronald. 1970. *Ngeq rhyme dictionary* (Vietnam Data Microfiche Series no. VD45-75). SIL Unpublished database.

Smith, Ronald. 1973. Reduplication in Ngeq. *Mon-Khmer Studies* 4:85–111.

Solntseva, Nina. 1996. Case-marked pronouns in the Ta'oih language. *Mon-Khmer Studies* 26:33–36.

Srivises, Prasert. 1978. *Kui (Suai) – Thai – English dictionary*. Jerry W. Gainey and Theraphan L. Thongkum (eds.). Bangkok: Indigenous Languages of Thailand Research Project, Chulalongkorn University Language Institute.

Sulavan, Khamluan, Thongpheth Kingsada, and Nancy Costello. 1998. *Katu-Lao-English Dictionary*. The Ministry of Information and Culture and The Institute of Research on Lao Culture, Lao P.D.R.

Svantesson, Jan-Olof. 1983. *Kammu phonology and morphology*. Lund: Gleerup.

Thomas, David. 1992. On sesquisyllabic structure. *Mon-Khmer Studies* 21:206–10.

THE INTEGRATION OF FRENCH LOANWORDS INTO VIETNAMESE: A CORPUS-BASED ANALYSIS OF TONAL, SYLLABIC AND SEGMENTAL ASPECTS

Vera Scholvin and Judith Meinschaefer
Freie Universität Berlin
vera.scholvin@fu-berlin.de, judith.meinschaefer@fu-berlin.de

Abstract

Due to its history of language contact with French, modern Vietnamese contains numerous loanwords of French origin, many of which refer to a variety of culturally transmitted items (such as clothing, food, technology, tradeable objects more generally). The present study deals with the phonological aspects of such loans, considering tone, syllable structure and segmental structure. The analysis is based on a corpus of roughly 500 Vietnamese nouns of French origin that, according to native speakers' judgments, are still in use. As for tonal structure, generalizations about tone assignment made in previous research are modified. The systematic analysis of repair strategies applying to French consonant clusters in onsets and codas shows that Vietnamese generally prefers deletion over epenthesis, unlike many other languages, with two additional repair processes being attested in specific contexts, as well.

Keywords: loanwords, phonology, tones, syllables, language contact
ISO 639-3 codes: vie, fra

1 Introduction

The integration of loanwords is one of the classical research topics in linguistics, since the processes occurring in loanword integration potentially shed light on questions pertaining to a variety of linguistic subdisciplines, among which are sociolinguistics, historical linguistics as well as grammatical theory. The present study addresses the integration of French loanwords into Vietnamese, with a focus on the phonology of tone and syllable structure.

The phonological systems of French, an Indo-European language of the Romance branch, spoken in Western Europe, and Vietnamese, an Austroasiatic language of the Vietic branch, spoken in Vietnam, are structurally distinct. First, concerning the prosodic type (in the sense of Hyman 2006), French is (probably) a stress accent language (Pulgram 1965; Di Cristo 1999), while Vietnamese is a tone language (Nguyễn 1997; Pham 2003; Kirby 2011; Brunelle 2014; Brunelle and Kirby 2016). Second, French and Vietnamese have different phonotactics: While French allows complex syllable onsets and codas (Klausenburger 1970; Tranel 1987), in Vietnamese onsets and codas consisting of more than a single consonant are illicit (Nguyễn 1997; Kirby 2011). In addition, only a subset of the Vietnamese consonants can occur in coda position, but in French the inventories of onset and coda consonants are roughly identical.

Furthermore, the French lexicon contains many content words consisting of three or more syllables, while a relatively high proportion of Vietnamese content words are mono- or disyllabic (cf. Đ. H. Nguyễn 1997 and Trần 2011). One might expect these differences to be reflected in maximality and minimality constraints on the size of prosodic words in each language, but the prosodic structure of French and Vietnamese may not be so different, after all. For both languages, it is controversial whether they have prominence at the word level, cf. Brunelle (2017) for Southern Vietnamese and Bosworth (2017) and Özçelik (2017) for two recent - and conflicting - views on French. Therefore, it is not clear to what extent the level of the prosodic word is relevant to the description of the phonology of Vietnamese — and less so — of French; cf. in particular Schiering, Bickel and colleagues for Vietnamese (Schiering, Bickel and Hildebrandt 2010) and Pulgram (1965) and much subsequent work for French (e.g., Delais-Roussarie 1996; Jun and Fougeron 2002). Finally, the segmental inventories of French and Vietnamese overlap only partially, both with regard to consonants and to vowels (cf. section 4).

In the light of these structural differences, when adapting a French word into Vietnamese, speakers need to assign each syllable a tone, simplify consonant clusters, and map French segments without direct correspondents in Vietnamese onto word forms permitted in the target language. In the present study, French loanword integration in Vietnamese is analysed on the basis of a corpus of roughly 500 loanwords that are

still in use in contemporary Vietnamese, selected from a more comprehensive loanword corpus currently containing around 1000 French loans. The sub-corpus analysed here is accessible online; see section 4. In what follows, section 2 defines some basic concepts concerning the integration of loanwords and provides background information on the language contact situation between Vietnamese and French. A brief summary of previous research on the integration of French loanwords into Vietnamese is given in section 3. The corpus is described in section 4. Section 5 presents the result of the present study, starting with tone assignment in section 5.1. The mapping of French consonantal segments onto Vietnamese consonants is discussed in section 5.2, while section 5.3 deals with the integration of French consonant clusters. Conclusions are presented in section 6.

2 Processes of loan integration and the contact situation between Vietnamese and French

2.1 Lexical borrowing and loan integration

Language contact, however shallow it may be, often leads to the borrowing of words from one language (the 'source language') into the other (the 'target language'). Borrowing is thus an uncontroversial case of language change caused by contact (see also Thomason 2006). Language contact occurs whenever a given speaker makes use of, in addition to his or her first language ('L1'), linguistic material of another language (ranging from a few words to fluent production in that language). This language may have been acquired as a second language ('L2'), but it may also be a first language in the case of multilingual first language acquisition. For the sake of simplicity, we assume here a somewhat prototypical definition of the terms 'first' and 'second' language, primarily based on age of acquisition (i.e., roughly speaking, before or after the age of six years, cf. Saville-Troike 2006; Lenneberg 1967). Depending on the sociolinguistic characteristics of the contact situation, borrowing may be symmetric, i.e., both languages borrow from each other to a similar degree, or, as is more frequently attested, asymmetric, i.e., borrowing proceeds primarily from the language with more overt prestige in a given contact situation into the language with less overt prestige in that situation (Haspelmath 2009).

The present article focuses on situations of language contact between, on the one hand, speakers with Vietnamese as a first language and French as a second language and, on the other hand, French as L1. During the period of close contact between Vietnamese and French for almost a century of French rule from 1867 to 1954, linguistic borrowing — in the sense of 'language change' with somewhat stable effects on the lexicon, as conceived of by Thomason and colleagues (Thomason and Kaufman 1988; Thomason 2001) — occurred primarily from the language with more overt prestige in that specific situation, i.e. French, into the language with less overt prestige, Vietnamese.

Following Paradis & LaCharité (1997:391), who in turn base their definition on Poplack, Sankoff & Miller (1988), we consider a word of a target language L1 (here: Vietnamese) to be a 'loanword' from a source language L2 (here: French) if it 'is incorporated into the discourse of L1;... has a mental representation in L1; and... is made to conform with... the... phonological constraints of L1.' According to this definition, processes of loanword adaptation consist of the integration of a non-native lexeme, drawn from a source language L2, into the lexicon of a recipient language L1, modifying, among other things, the word's phonetic and phonological representation such as to adapt it to the phonetics and phonology of L1. It is precisely these processes of phonetic and phonological integration that are the focus of the present study.

Two aspects of this definition are worth further mention. First, a form is considered a loanword only if it is actually used ('incorporated into the discourse') by speakers of L1 (i.e., Vietnamese) and if it is considered part of the lexicon ('has a mental representation'). The present study has ensured that the data adhere to this condition by analysing only data which are still in use, checking potential loanwords against both native speaker judgments and a current Vietnamese dictionary; see section 4.

Second, and more importantly, the study of loanword integration provides a window into the productive phonetic and phonological constraints of the target language, which become visible in the form of changes that word forms of the source language undergo in the course of their integration into the target language. The native lexicon of a language contains words that have been living in the language for centuries and that often have accumulated a host of morphophonological irregularities that are no longer related to productive alternations. Loanwords, in contrast, are new words, and the integration of a loanword into the target language is a creative process in which native speakers draw on their knowledge of currently productive rules and patterns of the language. For this reason, productive processes and default properties of the target

language may be more readily visible in loanword adaptation than in the historically evolved native lexicon. Hence, we consider the study of loanword adaptation as a fruitful path to a better understanding of the productive patterns of Vietnamese phonology.

Finally, research of the last two decades has yielded a growing body of knowledge on universal principles of loanword adaptation that is too comprehensive to be summarized here; recent reviews are provided by, e.g., Uffmann (2015); Kang (2011); Haspelmath & Tadmor (2009) and Paradis & Lacharité (2011). The integration of loans in Vietnamese appears instructive in this respect, as it does not follow commonly accepted typological generalizations concerning the repair of consonant clusters. First, cross-linguistically there seems to be, at least in word initial (onset) position, a preference for epenthesis over deletion (cf. Kang's 2011 discussion of more than 30 languages, Shinohara's 2006 study on the five typologically distinct languages Cantonese, Marshallese, Fijian, Yoruba and Samoan). Second, strategies of segmental integration have been found to be more variable in word-final position as compared to word-initial position (Kang 2011). In 5.4, we will discuss the results of the present study in the light of these two generalizations.

2.2 Language contact between French and Vietnamese

According to Alves' (2009) study on a selection of about 1,200 loanwords, around 90 per cent of the loanwords in Vietnamese are of Chinese origin. Loanwords from French, in contrast, make up only around 4 per cent of Vietnamese loanwords, with the proportion of English loanwords being even smaller. During the Chinese domination from 111 B.C. to 938 A.D., i.e., for roughly a millennium, the Chinese administrators introduced, among other innovations, a Chinese-style educational system (Wright 2002). The French, in contrast, dominated Vietnam for less than a century. In 1867, the South of Vietnam became a French colony (Cochinchina), and the French rulers aimed at replacing the traditional Chinese-style education with a French school system, though with little success (Le 2008). Education according to the Chinese model was preferred by the Vietnamese elites even during the French presence (Le Failler 2015). Consequently, the teaching of French from elementary school onwards between 1876-1906 did not succeed in spreading knowledge of French and was abandoned in the 20th century (Nguyen and Nguyen 2008). Finally, in 1954, the French lost all political power in Vietnam.

Though there are ample general historical records of this period to date, we do not have a precise picture of the language contact situation between French and Vietnamese during the French domination. Given that at the end of the nineteenth century less than 10 per cent of the population of Vietnam was of French origin (Le 2008), and given the low number of native speakers of Vietnamese enrolled in French-style primary or secondary schools (with less than 2 per cent of the total population having completed elementary school according to Nguyen & Nguyen 2008), we consider it likely that most L1 speakers of Vietnamese had little to no knowledge of French. Uneducated speakers of Vietnamese communicated with French speakers in a French-Vietnamese pidgin language, but little is known of the structure of this pidgin, as serious attempts at its description were made only after it had already fallen out of use (Reinecke 1971; Phillips 1975).

We would like to speculate that in a situation with – supposedly – a low degree of bilingualism, where few speakers of the target language Vietnamese had knowledge of the source language French, it appears likely that loanword adaptation has been based on the phonetic surface structure of French, without interference from any knowledge of French phonology. The hypothesis, ultimately to be checked against much more data, is thus that adaptation of French words into Vietnamese is based on the French phonetic surface structure, as perceived by L1 speakers of Vietnamese with little knowledge of French and filtered through the phonological system of Vietnamese. The processes of adaptation of French loanwords into Vietnamese thus provide a window onto Vietnamese phonology, with minimal interference of French phonology.

3 Previous research on lexical borrowing from French into Vietnamese

The integration of French loans in Vietnamese has been the topic of a couple of previous studies, beginning with an article by Barker (1969), who formulates a number of generalizations about segment integration and tone assignment. Barker's study is based on a corpus of 136 loans, published in full length in his article. Most of his observations remain valid today. In the following three decades, the integration of French loans into Vietnamese received little interest in the research literature. A thesis by Vương (1992) and an article by Nguyễn (1997) focus on the phonology and orthography of French loans, dealing with truncation, tone

assignment, adaptation of consonants as well as consonant cluster repair. In a more recent monograph, Vương (2011), building on his thesis (1992), considers language contact in Vietnam in a broader setting, providing insights into dialectal variation found in processes of French loanword adaptation in the North as compared to the South. Nguyễn (2013), in another monograph on loanwords in Vietnamese, deals with orthographic differences between source lexeme and loanword. Huynh's (2008; 2010) work on French loans in Vietnamese is based on a corpus of approximately 600 words (including mostly nouns, but also adjectives and verbs), focussing on tone assignment in French loans. The corpus is published in full length in Huynh (2010), complemented with a thorough documentation and discussion of the data.

On the basis of Barker's (1969) corpus and generalizations, Pham (2012) develops an optimality-theoretical analysis of tone assignment in Vietnamese loans. A detailed recent study by Kang, Phạm & Storme (2016) has been conducted on the basis of a very large, but so far unpublished corpus of more than 1,000 words, with a focus on the adaptation of vowels. The authors show that French phonotactic tendencies with respect to vowel quality (such as the *Loi de position*, regarding the differing distribution of lax and tense vowels in closed or open syllables, cf. Storme 2017 and Eychenne 2014) seem to be preserved in loan adaptions by Vietnamese speakers. A recent study by Nguyen & Dutta (2017) proposes an optimality-theoretical analysis of consonant cluster integration, based on Barker's (1969) & Huynh's (2010) data. Unfortunately, this study contains no information about the size of the corpus.

4 Methods

The analysis presented here is based on a selection of 533 Vietnamese nouns of French origin, drawn from a corpus of currently 1038 words, which was compiled on the basis of various published sources. Corpora from Barker (1969), Huynh (2010) and V. K. Nguyễn (2013) were taken as a starting point. Informal interviews with Vietnamese informants helped to expand the corpus. The informants are native speakers of Vietnamese living in Germany who have learned Vietnamese in Vietnam as a first language and acquired German in their adult life as a second language. Although they do not have any knowledge of French, they are aware of the French origin of the words they mentioned. For all 533 selected nouns, it has been checked that they are still in use, drawing on native informants' judgments as well as on word frequency and use in the World Wide Web and a Vietnamese dictionary (Bùi *et al.* 2003). Concerning the pronunciation of loanwords in the corpus, the phonetic transcriptions of the Vietnamese loanwords were first generated automatically on the basis of the orthographic representation (Kirby 2008) and then checked with reference to native informants' pronunciation. Phonetic transcriptions of the French source words are based on the standard hexagonal pronunciation as may be found in common dictionaries (Rey-Debove and Rey 2013). The corpus is accessible online at http://dx.doi.org/10.17169/refubium-1023.

5 Results and Discussion

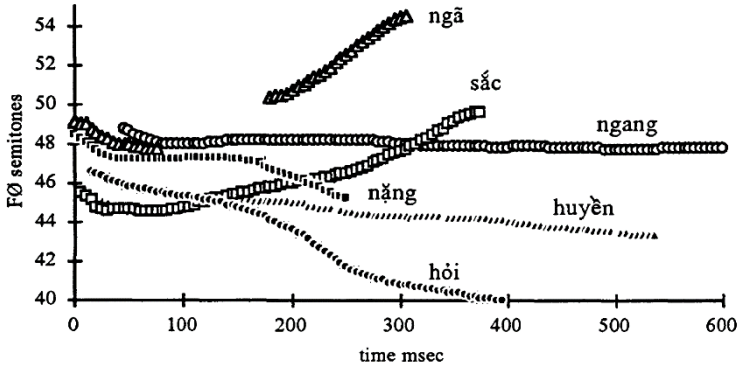
In this section, three aspects of Vietnamese loanwords from French are dealt with: first, we briefly discuss our results with respect to tone assignment, basically confirming and refining generalizations stated in previous research, then we consider the integration of French consonantal segments, and finally we deal with processes of repair of syllable structure.

5.1 Tone assignment in French loanwords in Vietnamese

French is a stress accent language (in the sense of Hyman's 2006 typology); yet, the main correlate of stress accent in French is pitch. Vietnamese, in contrast, is a tone language (again in the sense of Hyman 2006). Vietnamese tones are obligatory and culminative, with the tonal domain being the syllable (more precisely the rhyme), so that that every syllable is realized with a tone. Six tones (five in Southern varieties) can be discriminated in open syllables and in syllables ending in a sonorant, whereas only two tones occur in syllables ending in an oral stop (Kirby 2011). It is a topic of debate whether the two tonal categories to be distinguished in stop-final syllables are identical to two of the six tonal categories occurring in open and sonorant-final syllables or not, that is, whether the phonological system of Vietnamese differentiates six tones or eight tones. While the traditional stance is that Vietnamese has six tones, A. H. Pham (2001, 2003) as well as Michaud (2004) argue for the eight-tone view, based on evidence from tonal constraints in traditional poetry as well as in reduplication processes. Here, we follow the assumption that Vietnamese has a six tone system (Nguyễn 1997; Brunelle 2014; Brunelle and Kirby 2016). Phonetically, tonal distinctions

are implemented by pitch contour, intensity and duration, as well as laryngealization, glottalization and other aspects of voice quality (Brunelle 2009). To date, there is no consensus which of the various phonetic correlates of tone are phonologically relevant. While the traditional assumption is that pitch contour is phonemic (Vũ 1981), A.H. Pham's view is that 'instead of pitch height being contrastive as is generally assumed, it is phonation types of creakiness and breathiness which are distinctive as the register feature in North Vietnamese, and the differences in pitch heights are predictable' (A. H. Pham 2001, p. ii). The pitch differences of one speaker of Northern Vietnamese are shown in Figure 1.

Figure 1: *Vietnamese Tones (Northern standard variety), adapted from Nguyễn & Edmondson (1998)*



In Table 1, we list the Vietnamese names of the tones, their phonetic features (cf. Brunelle 2009), diacritics used in the standard orthography, as well as the numbers employed standardly in IPA transcriptions (cf. Kirby 2008). In following, we only refer to the numbers.

Table 1: *Vietnamese Tones*

Name	Phonetic Features			Orthographic representation	Number (IPA representation)
	Contour	Height/Register	Voice Quality		
Ngang	Level	High	Modal	a (no diacritic)	1
Huyền	Falling	Low	Modal/Breathy	à	2
Ngã	Falling-Rising	High	Creaky/Glottal	ã	3
Hỏi	Falling (-Rising)	Low	Creaky	ả	4
Sắc	Rising	High	Modal	á	5
Nặng	Falling	Low	Glottal[1]	ạ	6

Let us start with the two basic generalizations aboreut tone assignment of French loanwords in Vietnamese already described in the literature (cf. Barker 1969, Huynh 2008, M. Pham 2012), confirmed by our data. The majority of loanwords are assigned tone 1, as shown by the examples in (1) as well as in Tables 2 and 3.

(1)	<pénicilline>	[penisilin]	<pê ni ci lin>	[pe1 ni1 si1 lin1]	'penicillin'
	<relais>	[r(ə)lɛ]	<rơ le>	[zɤ1 lɛ1]	'relay'
	<ragoût>	[ʁagu]	<ra gu>	[za1 ɣu1]	'ragout'

If a syllable ends in a plosive, it is never assigned tone 1, but either tone 5 or 6, as shown in (2). While the most frequent pattern is assignment of tone 5 (191 syllables = 92 percent of 208 ending in a plosive), tone 6 is assigned in some cases (17 syllables = 8 percent),[1] in line with the distribution in Huynh (2010).

[1] We did not count cases with potential intra- and inter-speaker-variability between tone 5 and 6 when there is no tone specification in the orthography.

(2)	<biciclette>	[siklɛt]	<xích>	[sik5]	‘bike’
	<atlas>	[atlas]	<át-lát>	[at5 lat5]	‘atlas’
	<cartable>	[kaʁtabl]	<cặp táp>	[kặp6 tap5]	‘briefcase’
	<gaze>	[gaz]	<gạc>	[ɣak6]	‘gauze’

Table 2: *Tone assignment in monosyllabic loanwords*

	Absolute number of words	**Per Cent**
Tone 1	94	55 %
Tone 2	10	6 %
Tone 3	0	0 %
Tone 4	0	0 %
Tone 5	62	36 %
Tone 6	6	3 %
Total	172	

Further generalizations can be drawn with respect to disyllabic loanwords, as shown in Table 3.

Table 3: *Tone assignment in disyllabic loanwords*

Tonal sequence	**Absolute number of words**	**Per Cent**
1 1	163	53,4 %
1 2	3	1,0 %
1 5	33	10,8 %
1 6	7	2,3 %
2 1	17	5,6 %
2 2	4	1,3 %
2 5	6	2,0 %
2 6	5	1,6 %
4 1	2	0,7 %
4 5	2	0,7 %
4 6	1	0,3 %
5 1	41	13,4 %
5 5	16	5,2 %
5 6	1	0,3 %
6 1	2	0,7 %
6 2	1	0,3 %
6 5	1	0,3 %
Total	305	

In disyllabic loanwords, the syllable structure of the initial syllable appears to be relevant to tone assignment of this syllable. First, if the initial syllable is closed, tone 2 is hardly ever assigned. As shown in (3a-b), either tone 5 or tone 6 is assigned to word-initial syllables ending in a plosive (59 items = 58 per cent of all 103 disyllabic loanwords with a closed word-initial syllable). Tone 1 is assigned to most closed word-initial

syllables ending in a sonorant (41 items = 40 per cent of 103), as shown in (3c-d), but in three of the relevant words (3 per cent of 103) the first syllable is assigned tone 2; see (3e-g).

(3)	a. <antenne>	[ãtɛn]	<ăng ten>	[ăŋ1 tɛn1]	'antenna'
	b. <balcon>	[balkɔ̃]	<ban-công>	[ban1 koŋ͡m1]	'balcony'
	c. <taxi>	[taksi]	<tắc xi>	[tăk5 si1]	'taxi'
	d. <tabiler>	[tablije]	<tạp-dề>	[tap6 ze2]	'apron'
	e. <bidon>	[bidɔ̃]	<bình toong>	[biŋ2 tɔŋ͡m1]	'water bottle'(milit.)
	f. <paletot>	[palto]	<bành tô>	[bɛŋ2 to1]	'long coat'
	g <mouchoir>	[muʃwaʁ]	<mùi soa>	[muj2 swa1]	'handkerchief'

If the initial syllable is open, it is likewise sometimes assigned tone 2 (26 words, i.e. 14 per cent of 187 disyllabic loans with an open word-initial syllable); see (3b-e). Of the 27 word-initial open syllables bearing tone 2, the nucleus is a central vowel – [a] or [ɤ] – in 21 words (84 per cent), as shown in (4a-c), as compared to 5 items with other vowels and tone 2, as shown in (4d).

(4)	a.	<chemise>	[ʃ(ə)miz]	<sơ mi>	[sɤ1 mi1]	'shirt'
	b.	<carotte>	[kaʁɔt]	<cà rốt>	[ka2 zot1]	'carrot'
	c.	<blouse>	[bluz]	<bờ lu>	[bɤ2 lu1]	'blouse'
	d.	<ressort>	[ʁəsɔʁ]	<lò xo>	[lɔ2 sɔ1]	'spring'(tec.)

As shown above, the generalizations about tone assignment to the first syllable of disyllabic loans are gradient. They complement Barker's (1969) less specific claim that the first syllable of disyllabic borrowed words *often* takes tone 2, as well as M. Pham's (2012) statement that in disyllabic words with an open initial syllable and a final closed syllable, the initial syllable mostly receives tone 2. According to our data, whether the second syllable is closed or open is irrelevant. In sum, while tone 1 may be considered the default in tone assignment to French loanwords, segmental quality plays a role, as well. On the one hand, it is relevant whether a syllable ends in a plosive or a sonorant; on the other hand, whether the vowel is a central vowel or a front/back vowel.

5.2 Adaptation of segmental structure

Let us start with two basic generalizations concerning the adaptation of segmental structure. Subsequently, a more detailed view of onset and coda retention and replacement will be provided. On the one hand, segments found in the inventories of both languages are retained; on the other hand, French segments which are not part of the Vietnamese inventory are replaced. Given that in Vietnamese, differently from French, only a subset of consonants is licit in the coda of a syllable, consonants that are illicit in the coda are likewise replaced. As a consequence, repairs occur more frequently in coda positions than in onset position, an observation about loan integration that holds for other language pairs, as well (Shinohara 2006; Kang 2011). In general, segments illicit in the target language are replaced by segments that are similar to the source segment.

5.2.1. Onset consonants

Before providing a more detailed view of the integration of onset consonants, we start with an overview of the segment inventories of French (based on standard descriptive works such as Tranel 1987; Walker 2001), as shown in Table 4, and of Vietnamese consonants that are licit in onset position, illustrated in Table 5 (cf. Kirby 2011; Thompson 1965; Nguyễn 1997; Brunelle 2014).

Table 4: *French onset consonants (Ile-de-France-variety)*

	Labial	**Dental**	**Alveolar**	**Palatal**	**Dorsal**	**Glottal**
Plosive	b	t d			k g	
Nasal	m	n		ɲ		
Frikative	f v		sz	ʃ ʒ	ʁ	
Lateral		l				
Approximant	w			j ɥ		

Table 5: *Vietnamese onset consonants (Hà-Nội-/Northern standard variety)*

	Labial	**Dental**	**Alveolar**	**Palatal**	**Dorsal**	**Glottal**
Plosive	(p[2]) ɓ	t tʰ	ɗ	tɕ	k	ʔ
Nasal	m	n		ɲ	ŋ	
Flap			ɾ[3]			
Frikative	f v		s z		x ɣ	h
Lateral		l				
Approximant	w					

In onset position, twelve of the French consonants have direct correspondents in Northern standard Vietnamese, i.e., [ɓ, t, ɗ, k; m, n, ɲ; f, v, s, z; w]. French onset consonants without a corresponding segment in the Vietnamese inventory are replaced systematically by similar segments; as shown in Table 6.

Table 6: *Replacement of onset consonants*

Replacement	**What changes?**		**Example**
ʃ → s	Place	Loss of feature [high]	<choc> [ʃɔk] → <sốc> [sok͡p5] 'choc'
ʒ → z	Place	Loss of feature [high]	<gène> [ʒɛn] → <gien> [zɛn1] 'gene'
g → ɣ	Manner (Constriction)	[plosive] → [continuant]	<golf> [gɔlf] → <gôn> [ɣon1] 'golf'

Three onset consonants, i.e., [j], [ʁ] and [p], show variable integration. The integration of [p] has been discussed in previous studies (cf. Nguyễn 1997, Đoàn, Nguyễn & Phạm 2009, Kirby 2011) and shall not be dealt with here.

The dorsal fricative [ʁ] is integrated into Vietnamese in most cases by the coronal fricative [z]. It seems possible that this sound has been integrated into Southern Vietnamese as [ɾ], due to the perceptive similarity between [ʁ] and [ɾ]. Subsequently, it may have been replaced in the North by its allophonic counterpart [z]; it could also be a reading adaptation. Still, some speakers of the Northern standard variety use the sound [ɾ] when they are aware of the word's status as a loanword. If they have knowledge of English, they sometimes use the approximant [ɹ]. Below, we refer to the pronunciation of one speaker, using [z] for some words (5a-b), [ɾ] or [ɹ] for others (5c-d).

[2] A voiceless bilabial as an allophonic variant of [ɓ] occurs in only a few loanwords and is not realized by all speakers.

[3] The same holds for the alveolar flap [ɾ]. In many other varieties of Vietnamese, [ɾ] is an allophone of [z]; therefore, speakers of all varieties are familiar with that sound.

(5)	a.<rail>	[ʁɑj]	<ray>	[zăj1]	‘rail’
	b.<relais>	[r(ə)lɛ]	<rơ le>	[zɤ1 lɛ1]	‘relay’
	c.<radio>	[ʁadjo]	<ra đi ô>	[ɾa1 đi1 o1, ɹa1 đi1 o1]	‘radio’
	d.<rideau>	[ʁido]	<riđô, ri-đô>	[ɾi1 đo1]	‘curtain’

The palatal glide [j] may be replaced by [ŋ], [z] or [i]. At first sight, these sounds have little phonetic similarity to each other. Under a phonological perspective, however, the adaptation of [j] as [ŋ], [z] or [i] appears systematic. As to its replacement by [z], let us briefly mention that for socio-historical reasons it seems plausible that the contact variety for many words has been Southern Vietnamese (cf. Huynh 2008). In Southern Vietnamese varieties, the sound [j] is, in fact, a possible onset consonant. Crucially, its allophonic counterpart in the Northern standard variety is [z]. Hence, the French consonant [j], which may originally have been integrated as [j] into Southern Vietnamese, is replaced by [z], as shown in (6a-c). There is only one item replacing the glide [j] with the corresponding vowel; see (6d).

(6)	a. <yaourt>	[ja.uʁt]	<da ua>	[za1 ʔuə1]	‘yogurt’
	b. <billiard>	[bijaʁ]	<bi-da>	[ɓi1 za1]	‘billiard’
	c. <tablier>	[tablije]	<tạp-dề>	[tap6 ze2]	‘apron’
	d. <iode>	[jɔd]	<i-ốt>	[ʔi1 ot5]	‘iodine’

What has been said in the previous paragraph holds for [j] in simple onset position not preceded by a vowel. If, in contrast, the sound [j] stands in word-internal simple onset position and is preceded by a vowel, it is syllabified as a coda consonant and therefore preserved as [j]; see (7). This is possible only because Vietnamese (cf. Nguyễn 1997), unlike French, is apparently not subject to the principle of onset maximization (Vennemann 1988).

(7)	<glaïeul>	[glajœl]	<lay-ơn>	[lăj1 ɤn1]	‘gladiolus’
	<maillot>	[majo]	<may-o>	[măj1 o1]	‘vest’
	<maillechort>	[majʃɔʁ]	<may-so>	[măj1 sɔ1]	‘nickel silver’
	<moyeu>	[mwa.jø]	<moay-ơ>	[mwăj1 ʔɤ1]	‘hub’

Finally, as shown in (8), if the [high] segment [j] stands in complex onset position and is preceded by a nasal consonant [m] or [n], it is either replaced by the [high] nasal consonant [ɲ] or by the vowel [i]. Phonologically, the former process may be conceived as a progressive (or perseverative) spreading of the feature [nasal] to the following glide, with the result of changing the illicit onset [j] into the licit one [ɲ], as shown in (8a-b). Where the glide [j] is replaced by the vowel [i], all features are preserved, but the segment is syllabified as a syllable nucleus rather than as a syllable margin; see (8c-d).

(8)	a. <camion>	[ka.mjɔ̃]	<cam-nhông>	[kam1 ɲo͡ŋm1]	‘truck’
	b. <aluminium>	[alyminjɔm]	<nhôm>	[ɲom1]	‘aluminium’
	c. <amiante>	[amjãt]	<a-mi-ăng>	[a1 mi1 ăŋ1]	‘asbestos’
	d. <ammoniac>	[amɔnjak]	<a-mô-ni-ác>	[a1 mo1 ni1 ak5]	‘ammonia’

The data presented in this paragraph show that the integration of onset consonants is systematic and may be accounted for by phonological as well as by socio-historical factors. Furthermore, orthography may have played an important role. It seems possible that certain words are reading adaptations (cf. Vendelin & Peperkamp 2006).

5.2.2. Coda consonants

In Vietnamese, only ten consonants are licit in coda position: the three voiceless obstruents [p, t, k], three (non-palatal) nasal consonants [m, n, ŋ], the glides [j,w] as well as the double-articulated sounds [ŋ͡m, k͡p], standing in complementary distribution with [ŋ, k] after back rounded vowels (cf. Kirby 2011). Fricative, palatal (with the exception of the palatal glide [j]), glottal and lateral segments as well as voiced obstruents

are illicit in coda position. In French, in contrast to Vietnamese, basically all consonants are licit codas. French coda consonants that are not licit codas in Vietnamese are thus replaced by similar segments, delinking or replacing as few features as possible; an overview of selected replacement processes is given in Table 7. Note that one and the same segment may be replaced by different segments, depending on whether it occurs in coda or in onset position. To give an example, French [ʁ] is replaced by [z] in onset position and by [k] in coda position.

Table 7: *Replacement of selected coda consonants*

Replace-ment	What changes ?		Example (Fr.|Viet.|Glosse)		
l → n	Manner of articulation	[Lateral] → [Nasal]	<caramel> ⌈ka ʁa mɛl⌉	<caramen> ⌈ca ra men⌉	'caramel'
d → t	Voicing	Delinking of [Voiced]	<acide> ⌈asid⌉	<a-cít> [a1 sit5]	'acid'
f → p	Manner & Place	[Continuant] → [Plosive] ⌈Coronal⌉ → ⌈Dorsal⌉	<bifteck> ⌈bif tɛk⌉	<bíp tết> ⌈bip5 tek5⌉	'beef-steak'
s → t	Manner (Constriction)	[Continuant] → [Plosive]	<caisse> ⌈kɛs⌉	<két> ⌈kɛt5⌉	'cash desk'
ʃ → t, k	Manner & Place	[Continuant] → [Plosive] Delinking of feature ⌈high⌉	<bâche> ⌈baʃ⌉	<bát> ⌈bat6⌉	'tarpau-lin'
		[Continuant] → [Plosive] ⌈Coronal⌉ → ⌈Dorsal⌉	<fiche> ⌈fiʃ⌉	<phích> ⌈fik5⌉	'plug'
ʁ → k	Manner (Constriction) & Voicing	[Continuant] → [Plosive] Delinking of [Voiced]	<garde> ⌈gaʁd⌉	<gác> ⌈ɣak5⌉	'guard'

As shown in Table 7, a few cases of consonant replacement are variable, while others are categorical. In other cases, French coda consonants that are not licit in Vietnamese are deleted, and in a few cases, they are replaced by one of the vowels [i, o, u]. When considering the whole picture, the integration of coda consonants appears to be based on a complex interaction of constraints that for reasons of space are not considered here.

5.3 Adaptation of consonant clusters by deletion and epenthesis

In what follows, we briefly summarize the most important generalizations concerning French and Vietnamese syllable structure, followed by an analysis of the two major repair processes applying to consonant clusters: vowel epenthesis and consonant deletion. A third, and minor, strategy consists in the syllabification of the first consonant in an onset cluster as a coda of the preceding syllable. Table 8 presents an overview of the frequency of different repair processes in onset and coda clusters.

Table 8: *Adaptation of consonant clusters by deletion and epenthesis*

	Onset clusters		Coda clusters		Total	
	Absolute number of words	Per Cent	Absolute number of words	Per Cent	Absolute number of words	Per Cent
Deletion	34	60 %	33	100 %	67	74 %
Epenthesis	15	26 %	0	0 %	15	17 %
Resyllabification	8	14 %	0	0 %	8	9 %
Total	57		33		90	

For reasons of space, we disregard the rather complex processes of adaptation observed in French consonant clusters preceded or followed by a schwa-vowel (25 words).

In Vietnamese, the onset is an obligatory constituent of the syllable. A syllable may have a coda, but only a subset of the consonant inventory is licit in coda position; see 5.2.2. Complex onsets and codas are disallowed, with the exception of the sequence C[w]V (Nguyễn 1997; Kirby 2011). It is, however, unclear whether the glide [w] should be analysed as part of the onset. As this structure occurs in both languages, no repair is needed for loans. In contrast to Vietnamese, French does allow complex onsets and codas (Klausenburger 1970; Tranel 1987). Here, we consider only French onset and coda clusters consisting of two consonants; more complex clusters are possible in French, but are not attested in the corpus analysed here.

French onset and coda clusters, illicit in Vietnamese, thus need to be repaired in loanword adaptation. Speakers generally use two possible repair strategies, i.e., vowel epenthesis (**CC**VC → **C**V.**C**VC) and consonant deletion (**CC**VC → **C**VC). As shown in Table 8, deletion is much more frequent than epenthesis (cf. also Nguyen and Dutta 2017). While deletion (5.3.1) is found in onset and coda clusters, epenthesis (5.3.2) is restricted to onset clusters. Resyllabification, i.e., the syllabification of the first consonant in an onset cluster as a coda of the preceding syllable, is by definition only possible in onset clusters. In onset clusters containing the glide [j], the glide is often replaced by the corresponding vowel [i]; see 5.3.3.

5.3.1. Deletion

Where deletion applies, the most common strategy is to maintain the consonant in the first position and to delete the second one. This is valid for both onset and coda clusters, with few exceptions (cf. Table 9).

Table 9: *Deletion of the first vs. the second consonant in a cluster*

	First consonant deleted		**Second consonant deleted**		**Total**
	Absolute number of words	**Per cent**	**Absolute number of words**	**Per cent**	**Absolute number of words**
Onset	11	32 %	23	68 %	34
Coda	4	12 %	29	88 %	33
Total	15	22 %	52	78 %	67

French consonants are replaced whenever they are either not part of the Vietnamese inventory or illicit in coda position. This also holds for consonant clusters, and the replacement patterns are the same as for single consonants; see Tables 6 and 7. An illicit consonant in the first position of a cluster is thus typically replaced rather than deleted.

In the corpus analysed here, many cases of deletion in onset clusters are sequences of C+[ʁ] (20 words) and C+[l] (6 words), exemplified in (9) and (10). The pattern exemplified in (10) constitutes an exception: In onset clusters with a lateral consonant in second position, it is the first consonant that is deleted, while the second is maintained. These findings fall in line with Vương (1992).

(9) Deletion in onset clusters: C+[ʁ] → C (Deletion of second consonant)

<brancard>	[bʁɑ̃kaʁ]	<băng ca>	[ɓăŋ1 ka1]	‘stretcher’
<cravatte>	[kʁavat]	<cà vạt>	[ka2 vat6]	‘tie’
<fromage>	[fʁɔmaʒ]	<pho mát>	[fɔ1 mat5]	‘cheese’

(10) Deletion in onset clusters: C+[l] → [l] (Deletion of first consonant)

<complet>	[kɔ̃plɛ]	<com lê >	[kɔm1 le1]	‘suit’
<glaïeul>	[glajœl]	<lay-ơn>	[lăj1 ʔɤn1]	‘gladiolous’
<chou-fleur>	[ʃuflœʁ]	<su lơ>	[su1 lɤ1]	‘cauliflower’

As to coda clusters, it is generally the second consonant which is deleted; the first is replaced if illicit in coda position; see (11). Examples in which the first consonant is preserved and the second deleted are given in (12); the first consonant is replaced and the second deleted in (13).

(11)

[ʁ]+C → [k]	14 items (and three exceptions, see 13 a,c,d)
[l]+C → [n]	6 items (and one exception, see 13b)
[s]+C → [t]	6 items
[k]+C → [k] or [k͡p]	2 items
[m]+C → [m]	1 item

(12) Deletion in coda clusters: First consonant preserved, second deleted

<contact>	[kɔ̃takt]	<công-tắc>	[koŋ͡m1 tak5]	‘switch’
<inox>	[inɔks]	<i-nốc>	[ʔi1 nok͡p5]	‘stainless steel’
<pompe>	[põp]	<bơm>	[ɓɤm1]	‘pump’

(13) Deletion in coda clusters: First consonant replaced, second deleted

<harpe>	[aʁp]	<hạc>	[hak6]	‘harp’
<citerne>	[sitɛʁn]	<xitéc>	[si1 tɛk5]	‘tank’
<talc>	[talk]	<tan>	[tan1]	‘talc’

The integration of the consonant [ʁ] in coda position has been studied by Vương (1992) and in detail by Kang et al. (2016), who claim that the neutralization of the French phonemes /ʁ/ and /k/ is due to Vietnamese phonological restrictions, ‘but the Vietnamese adaptation systematically retains the contrast in the quality and length difference in the preceding vowel’ (Kang et al. 2016, p. 11). The same holds for clusters with [ʁ]+C in the following examples given in their article: French <cirque> [siʁk] and <course> [kuʁs] are adapted as Vietnamese <xiếc> [siək5] ‘mustard’ and <cuốc> [kuək5] ‘ride’.

Let us now briefly turn to the three exceptions for [ʁ]+C-clusters and one exception for [l]+C-clusters, where the output is not, as expected, [k] or [n], as shown in (14). In the first two cases (14a, b) [ʁ/l+m] → [m], the first consonant is deleted, but the second preserved. This may be due to the saliency of the second consonant of the cluster, the nasal [m]. In the third case (14c) V+[ʁ]+C → VV, the consonant [ʁ] is replaced by a vowel, possibly due to perceptual similarity between [ʁ] and low vowels. Finally, (14d), is an irregular variant to the regular integration of French *moutarde*. Corpus deletion patterns are given in Table 10.

(14) Exceptional cases for the deletion in coda clusters

a. <forme>	[fɔʁm]	<phom>	[fɔm1]	‘form’
b. <film>	[film]	<phim>	[fim1]	‘film’
c. <yaourt>	[jauʁt]	<da ua>	[za1 ʔuə1]	‘jogurt’
d. <moutarde>	[mutaʁd]	<mù tạt>	[mu2 tat6]	‘mustard’

***Table 10*:** Patterns of deletion in the adaptation of consonant clusters

Integration of the cluster	**Consonant**		**Position Plays a role**	
	1st	**2nd**		
C+[ʁ] → [C]	preserved	deleted	yes	Onset
C+[l] → [l]	deleted	preserved	--	
[l]+C → [n]	replaced	deleted	yes	
[s]+C → [t]	replaced	deleted	yes	
[k]+C → [k]/[k͡p]	replaced	deleted	yes	
[m]+C → [m]	preserved	deleted	yes	Coda
[ʁ]+C → [k]	replaced	deleted	yes	
[ʁ]+C → V	replaced (vowel)	deleted	yes	
[ʁ,l]+[m] → [m]	deleted	preserved	--	

5.3.2. Epenthesis

In the adaptation of French consonant clusters into Vietnamese, epenthesis applies far less frequently than deletion, attested only in onset clusters; see Table 8. A few words are adapted alternatively with deletion or epenthesis (4 items).[4] Some examples for epenthesis are given in (15).

(15) Epenthesis in CC sequences

<blouse>	[bluz]	<bờ lu>	[ɓɤ5 lu1]	'blouse'
<clef>	[kle]	<cơ lê, cờ lê>	[kɤ1 le1], [kɤ2 le1]	'spanner'
<crème>	[kʁɛm]	<kem, cà rem>	[kɛm1], [ka2 zɛm1]	'ice-cream'
<scandal>	[skɑ̃dal]	<xì căng đan>	[si2 kăŋ1 ɗan1]	'scandal'

Three epenthetic vowels are attested in the corpus, [a, i, ɤ]; of these, [ɤ] has the highest frequency. It seems possible that the place of articulation of the preceding consonant is one of the factors that determine the choice of the low, high, or mid vowel (cf. Uffmann 2006); additional data is needed to confirm this hypothesis.

5.3.3. Adaptation of the glide [j] in onset clusters

The corpus analysed here contains a total of 23 clusters of the structure C+[j] in onset position. In these clusters, the glide [j] is mapped onto the vowel [i] in 19 instances (87 per cent), as shown in (16a-f) and onto the vowel [ɯ] in one instance; see (16g).

(16) Adaptation of C+[j] sequences

a. <barrière>	[baʁjɛʁ]	<barie>	[ɓa1 zi1]	'fence, gate'
b. <magnesium>	[maɲezjɔm]	<magie>	[ma1 zi1]	'magnesium'
c. <radium	[ʁadjɔm]	<ra-đi>	[za1 ɗi1], [ɾa1 ɗi1]	'radium'
d. <diode>	[djɔd]	<đi-ốt>	[ɗi1 ot5]	'diode'
e. <piano>	[pjano]	<piano>	[pi1 a1 no1]	'piano'
f. <violette>	[vjolɛt]	<vi-ô-lét>	[vi1 o1 lɛt5]	'pancy'
g. <légion>	[leʒjɔ̃]	<lê dương>	[le1 zɯəŋ1]	'Fr. Foreign Legion'

The same pattern of replacement of [j] by [i] is found where the glide [j] occurs in simple onset position; see (8c-d) above, i.e., all features of [j] are preserved, but the segment is syllabified as syllable nucleus rather than as syllable margin.

5.4. The adaptation of consonant clusters in a cross-linguistic perspective

When compared to generalizations about cluster integration in the scholarly literature, Vietnamese appears to be cross-linguistically unusual. According to Paradis & Lacharité (1997), it appears that epenthesis is generally preferred over deletion. In fact, typological generalizations about deletion and epenthesis in loanword adaptation made in previous studies state that deletion is generally infrequent in word-initial position, though some languages use both strategies, or even use deletion only (cf. Kang 2011 for an overview). In many other languages, however, such as Sesotho (Rose and Demuth 2006), Shona (Uffmann 2006) or Akan (Adomako 2008), epenthesis is the only repair strategy available in word-initial position. In Vietnamese, in contrast, the preferred strategy in onset position is deletion. Furthermore, it seems that the segmental context is not relevant in the choice between epenthesis and deletion, differently to what has been shown for, e.g., Hawaiian (Adler 2006), Thai, and a number of other languages discussed in Fleischhacker (2005). Concerning repair strategies in word-final clusters, 'it is not clear whether epenthesis is cross-linguistically the preferred strategy over deletion in this position' (Kang 2011: 14). A number of other languages are like Vietnamese in that epenthesis is unattested in word-final position, or in coda position

[4] In the sample of 77 illicit consonant clusters, these four items were counted twice.

more generally. In Thai, for instance, 'loans with a final cluster never employ epenthesis' (Kenstowicz & Suchato 2006 : 932).

Another aspect in which Vietnamese may be unusual relates to the factors that determine which of the two consonants in a cluster undergoes deletion. In some other languages, such as Cantonese, Marshallese, Yoruba, Fijian, patterns of deletion have been found to depend on the segmental identity of the consonants (cf. Shinohara 2006). Deletion patterns in Vietnamese, in contrast, depend on the position of a segment in the cluster rather than on the segmental content (with the exception of sequences consisting of an obstruent followed by a nasal or lateral). In this respect, however, Vietnamese is similar to Thai: In Vietnamese, it is mostly and in Thai it is always the second consonant that deletes (cf. Kenstowicz & Suchato 2006).

6 Conclusion

From an empirical perspective, the present study has contributed a couple of new generalizations, both with respect to the question of how tones are assigned as well as to how consonant clusters are adapted in Vietnamese loanwords from French. From a theoretical perspective, it has become clear that the phonological structure of Vietnamese is a crucial factor in the adaptation of French single consonants and consonant clusters. The data analysed here do not suggest that French phonological structure (as opposed to phonetic form) plays a role in loanword integration into Vietnamese.

In future research, we will both extend the methods employed and the amount of data analysed. Concerning the methodological perspective, it may be fruitful to compare experimentally elicited native speakers' pronunciations for nonce formations having specific phonological properties to loanword patterns and to lexico-statistical patterns extracted from a large electronic corpus of Vietnamese. Empirically, the loanword corpus is being enlarged in order to be able to describe patterns of syllable truncations and augmentations (via vowel epenthesis) and to better understand the role of minimality and maximality requirements on word length that may be relevant in loanword adaptation.

Acknowledgment

We would like to thank the audiences at the 7th International Conference on Austro-Asiatic Linguistics (ICAAL) in Kiel and at the Linguistics Colloquium of the Free University Berlin as well as two anonymous reviewers for helpful comments on an earlier version of this article. We are grateful to Bruce Mayo for checking our English. Needless to say, all remaining errors are our own.

References

Adler, Allison N. 2006. Faithfulness and Perception in Loanword Adaptation: A Case Study from Hawaiian. *Lingua* 116(7):1024–45.

Adomako, Kwasi. 2008. Vowel Epenthesis and Consonant Deletion in Loanwords: A Study of Akan. Unpublished MA thesis. University of Tromsø.

Alves, Mark J. 2009. Loanwords in Vietnamese. In *Loanwords in the world's language: A Comparative Handbook*, ed. by Martin Haspelmath and Uri Tadmor, 617–637. Mouton de Gruyter.

Barker, Milton E. 1969. The Phonological Adaption of French Loanwords in Vietnamese. *Mon-Khmer Studies Journal* 3:138–47.

Bosworth, Yulia. 2017. High Vowel Distribution and Trochaic Markedness in Québécois. *The Linguistic Review* 34(1):39–82.

Brunelle, Marc. 2017. Stress and Phrasal Prominence in Tone Languages: The Case of Southern Vietnamese. *Journal of the International Phonetic Association* 47(03):1–38.

Brunelle, Marc. 2009. "Tone Perception in Northern and Southern Vietnamese." *Journal of Phonetics* 37(1):79–96.

Brunelle, Marc. 2014. Vietnamese (Tiếng Việt). *The Handbook of Austroasiatic Languages. Volume 2*, edited by Mathias Jenny and Paul Sidwell, 907–53. Leiden: Brill.

Brunelle, Marc and James Kirby. 2016. Tone and Phonation in Southeast Asian Languages. *Linguistics and Language Compass* 10(4):191–207.

Bùi, Khắc Việt et al. 2003. *Từ điển tiếng Việt. Viện Ngôn Ngữ Học*. Hà Nội - Đà Nẵng: Nhà Xuất Bản Đà Nẵng.

Di Cristo, Albert. 1999. Vers Une Modélisation de l'accentuation Du Français: Première Partie. *Journal of French Language Studies* 9(02):143–79.

Delais-Roussarie, Élisabeth. 1996. Phonological Phrasing and Accentuation in French. *Dam Phonology : HIL Phonology Papers II*, vol. 3, ed. by Marina Nespor and Norval Smith. The Hague: Holland Academic Graphics.

Đoàn, Thiện Thuật, Khánh Hà Nguyễn, and Như Quỳnh Phạm. 2009. *A Concise Vietnamese Grammar*. Hanoi: Thế Giới Publishers.

Eychenne, Julien. 2014. Schwa and the Loi de Position in Southern French. *Journal of French Language Studies* 24(2):223–53.

Le Failler, Philippe. 2015. L'intégration des exonymes à la langue vietnamienne ou quand l'usage d'internet force la normalisation. *Moussons* 25:79–97.

Fleischhacker, Heidi Anne. 2005. *Similarity in Phonology: Evidence from Reduplication and Loan Adaptation*. PhD dissertation. Los Angeles: University of California

Haspelmath, Martin. 2009. Lexical Borrowing: Concepts and Issues. In *Loanwords in the World's Languages*, ed. by Martin Haspelmath and Uri Tadmor, 35–54. Berlin: De Gruyter.

Haspelmath, Martin and Uri Tadmor (eds.) 2009. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: de Gruyter Mouton.

Huynh, Sabine. 2008. L'assimilation Des mots d'emprunts français à la langue vietnamienne: La question des tons [The Assimilation of French Loanwords into the Vietnamese Language: The Question of Tones]. *Cahiers de Linguistique – Asie Orientale* 37(2):223–40.

Huynh, Sabine. 2010. *Les mécanismes d'intégration des mots d'emprunt français en vietnamien*. Paris: Harmattan.

Hyman, Larry M. 2006. Word-Prosodic Typology. *Phonology* 23(02):225–57.

Jun, Sun-Ah and Cécile Fougeron. 2002. Realizations of Accentual Phrase in French Intonation. *Probus* 14(1):147–72.

Kang, Yoonjung. 2011. Loanword Phonology. In *The Blackwell companion to phonology*, ed. by Marc van Oostendorp, Colin Ewen, Elizabeth Hume, and Karen Rice, 2258–82. Malden, MA: Wiley-Blackwell.

Kang, Yoonjung, Andrea Hoa Pham, and Benjamin Storme. 2016. French Loanwords in Vietnamese: The Role of Input Language Phonotactics and Contrast in Loanword Adaptation. *Proceedings of the Annual Meetings on Phonology 2*. https://doi.org/10.3765/amp.v2i0.3749

Kenstowicz, Michael and Atiwong Suchato. 2006. Issues in Loanword Adaptation: A Case Study from Thai. *Lingua* 116(7):921–49.

Kirby, James. 2011. Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41(03):381–92.

Kirby, James. 2008. VPhon: A Vietnamese Phonetizer.

Klausenburger, Jürgen. 1970. *French Prosodics and Phonotactics*. Berlin: de Gruyter.

Le, Thu Hang. 2008. Le Viêt Nam, un pays francophone atypique: Regard sur l'emprise française sur l'évolution littéraire et journalistique au Viêt Nam depuis la première moitié du XXe Siècle. *Documents pour l'histoire du Français langue étrangère ou seconde* 40/41:341–50.

Lenneberg, Eric Heinz. 1967. *Biological Foundations of Language*. New York: Taylor and Francis.

Michaud, Alexis. 2004. Final Consonants and Glottalization : New Perspectives from Hanoi Vietnamese. 61(2–3):119–46.

Nguyễn, Đình Hoà. 1997. *Vietnamese: Tiếng Việt không son phấn*. Amsterdam: John Benjamins.

Nguyễn, Đức Dân. 1997. Về âm và chính tả các từ tiếng Việt gốc Pháp. *Ngôn Ngữ* 3:40–44.

Nguyen, Huynh Trang and Hemanga Dutta. 2017. The Adaptation of French Consonant Clusters in Vietnamese Phonology: An OT account. *Journal of Universal Language* 18(1):69–103.

Nguyen, Quang Kinh and Quoc Chi Nguyen. 2008. Education in Vietnam: Development History, Challenges and Solutions. In *An African exploration of the East Asian education experience*, ed. by Birger Fredriksen and Jee-Peng Tan, 109–54. Washington D.C.: World Bank.

Nguyễn, Văn Khang. 2013. *Từ ngoại lai trong tiếng Việt*. Thành Phố Hồ Chí Minh: Nhà xuất bản Tổng hợp Thành phố Hồ Chí Minh.

Nguyễn, Văn Lợi and Jerold A. Edmondson. 1998. Tones and Voice Quality in Modern Northern Vietnamese. *Mon-Khmer Studies* 28:1–18.

Özçelik, Öner. 2017. The foot is not an obligatory constituent of the prosodic hierarchy: 'Stress' in Turkish, French and Child English. *The Linguistic Review* 34(1):157–213.

Paradis, Carole and Darlene Lacharité. 2011. Loanword Adaptation: From Lessons Learned to Findings. In *The Handbook of Phonological Theory*, ed. by John Goldsmith, Jason Riggle, and Alan Yu, 751–78. Oxford: Wiley-Blackwell.

Paradis, Carole and Darlene Lacharité. 1997. Preservation and Minimality in Loanword Adaptation. *Journal of Linguistics* 33(2):379–430.

Pham, Andrea Hoa. 2001. *Vietnamese tone, tone is not pitch*. PhD dissertation. Toronto: University of Toronto.

Pham, Andrea Hoa. 2003. *Vietnamese Tone. A New Analysis*. New York: Routledge.

Pham, Mike. 2012. *Tone Assignment of French Loanwords in Vietnamese*. Unpublished manuscript.

Phillips, John Seward. 1975. *Vietnamese Contact French: Acquisitional Variation in a Language Contact Situation*. PhD dissertation. Bloomington: Indiana University.

Poplack, Shana, David Sankoff, and Christopher Miller. 1988. The Social Correlates and Linguistic Processes of Lexical Borrowing and Assimilation. *Linguistics* 26(1):47–104.

Pulgram, Ernst. 1965. Prosodic Systems: French. *Lingua* 13:125–44.

Reinecke, John. 1971. Tây Bôi: Notes on the Pidgin French of Vietnam. In *Pidginization and creolization of languages*, ed. by Dell Hymes. Cambridge: Cambridge University Press.

Rey-Debove, Josette and Alain Rey, eds. 2013. Le Nouveau Petit Robert. Dictionnaire Alphabétique et Analogique de La Langue Française. Version Électronique.

Rose, Yvan and Katherine Demuth. 2006. Vowel Epenthesis in Loanword Adaptation: Representational and Phonetic Considerations. *Lingua* 116(7):1112–39.

Saville-Troike, Muriel. 2006. *Introducing Second Language Acquisition*. Cambridge: Cambridge University Press.

Schiering, René, Balthasar Bickel, and Kristine A. Hildebrandt. 2010. The Prosodic Word Is Not Universal, but Emergent. *Journal of Linguistics* 46:657–709.

Shinohara, Shigeko. 2006. Perceptual Effects in Final Cluster Reduction Patterns. *Lingua* 116(7):1046–78.

Storme, Benjamin. 2017. The Loi de Position and the Acoustics of French Mid Vowels. *Glossa* 2(1):1–25.

Thomason, Sarah G. 2006. Language Change and Language Contact. In *Encyclopedia of Language and Linguistics*, ed. by Keith Brown, 339–46. Amsterdam: Elsevier.

Thomason, Sarah Grey. 2001. *Language Contact. An Introduction*. Edinburgh: Edinburgh University Press.

Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Thompson, Laurence C. 1965. *A Vietnamese Grammar*. Seattle: University of Washington Press.

Trần, Thị Thúy Hiền. 2011. *Processus d'acquisition des clusters et autres séquences de consonnes en langue seconde: De l'analyse acoustico-perceptive des séquences consonantiques du vietnamien à l'analyse de la perception et production des clusters du français par des apprenants vietnamiens du FLE*. PhD dissertation. Grenoble: Université de Grenoble.

Trần, Thi Thuy Hien, Nathalie Vallée, and Silvain Gerber. 2016. Syllabe CVC et cycle mandibulaire : Une étude articulatoire des asymétries. Le Cas Du Vietnamien. *Journées d'Etudes sur la Parole*. Paris.

Tranel, Bernard. 1987. *The Sounds of French: An Introduction*. Cambridge: Cambridge University Press.

Uffmann, Christian. 2006. Epenthetic Vowel Quality in Loanwords: Empirical and Formal Issues. *Lingua* 116(7):1079–111.

Uffmann, Christian. 2015. Loanword Adaptation. In *The Oxford Handbook of Historical Phonology*, ed. by Patrick Honeybone and Joe Salmons. Oxford: Oxford University Press.

Vendelin, Inga and Sharon Peperkamp. 2006. The Influence of Orthography on Loanword Adaptations. *Lingua* 116(7):996–1007.

Vennemann, Theo. 1988. *Preference Laws for Syllable Structure and the Explanation of Sound Change: With Special Reference to German, Germanic, Italian, and Latin*. Berlin: Mouton de Gruyter.

Vũ, Thanh Phương. 1981. *The Acoustic and Perceptual Nature of Tone in Vietnamese*. Phd dissertation. Canberra: Australian National University.

Vương, Toàn. 2011. *Tiếng việt trong tiếp xúc ngôn ngữ từ giữa thế kỷ XX. Le vietnamien en contact linguistique depuis la deuxième moitié du XXe Siècle*. Hà Nội: Nhà Xuất Bản Dân Trí.

Vương, Toàn. 1992. *Từ gốc pháp trong tiếng Việt*. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

Walker, Douglas C. 2001. *French Sound Structure*. Calgary: University of Calgary Press.

Wright, Sue. 2002. Language Education and Foreign Relations in Vietnam. *Language Policies in Education: Critical Issues*, ed. by James W. Tollefson, 225–44. Mahwah: Erlbaum.

Yip, Moira. 1993. Cantonese Loanword Phonology and Optimality Theory. *Journal of East Asian Linguistics* 2(3):261–91.

WATERWORLD: LEXICAL EVIDENCE FOR AQUATIC SUBSISTENCE STRATEGIES IN AUSTROASIATIC

Roger Blench
McDonald Institute for Archaeological Research, University of Cambridge
rogerblench@yahoo.co.uk

Abstract
The Austroasiatic language phylum has long been established, but limited progress has been made towards a consolidated reconstruction of its proto-lexicon. Hence its homeland and routes of dispersal, as well as the potential subsistence systems of early speakers remain disputed. Sidwell & Blench (2011) put forward an aquatic dispersal model, hence the lexicon shouldreflect water and aquatic exploitation of resources. Indeed, it turns out that many items associated with these can indeed be reconstructed, including waterways, boats and water transport, fish and other river fauna and fish capture techniques. Recent redating of the SE Asian Neolithic suggests that agriculture only begins in the region between northern Vietnam and Thailand around 4000 BP. This correlates well with an aquatic dispersal based on access to both livestock and crops, as well as new types of watercraft. Speakers spread rapidly in all directions, following the main river arteries and even crossing the sea to the Nicobar Islands.

Keywords: Austroasiatic; reconstruction; homeland; dispersal

Acronyms: MKED (Mon-Khmer Etymological Dictionary), PB (proto-Bahnaric), PK (proto-Katuic), PKha (Proto-Khasic), PP (proto-Pearic), PPa (proto-Palaungic), PV (proto-Vietic)

1 Introduction
Although the Austroasiatic phylum has been long identified, limited progress has been made in the reconstruction of its proto-lexicon. For a summary of the current situation see Sidwell & Rau (2015). Individual branches have been reconstructed, and there are many scattered proposals for common lexemes shared between branches, but this is not reconstruction. In a number of instances the putative proto-forms in Shorto (2006 and online) are supported by citations from as few as two branches of Austroasiatic. These lacunae make it problematic to draw conclusions about the origin and routes of dispersal, as well as the potential subsistence systems of early speakers, a classical goal of historical linguistics. This in turn has implications for dating, since the SE Asian Neolithic is now very well known.

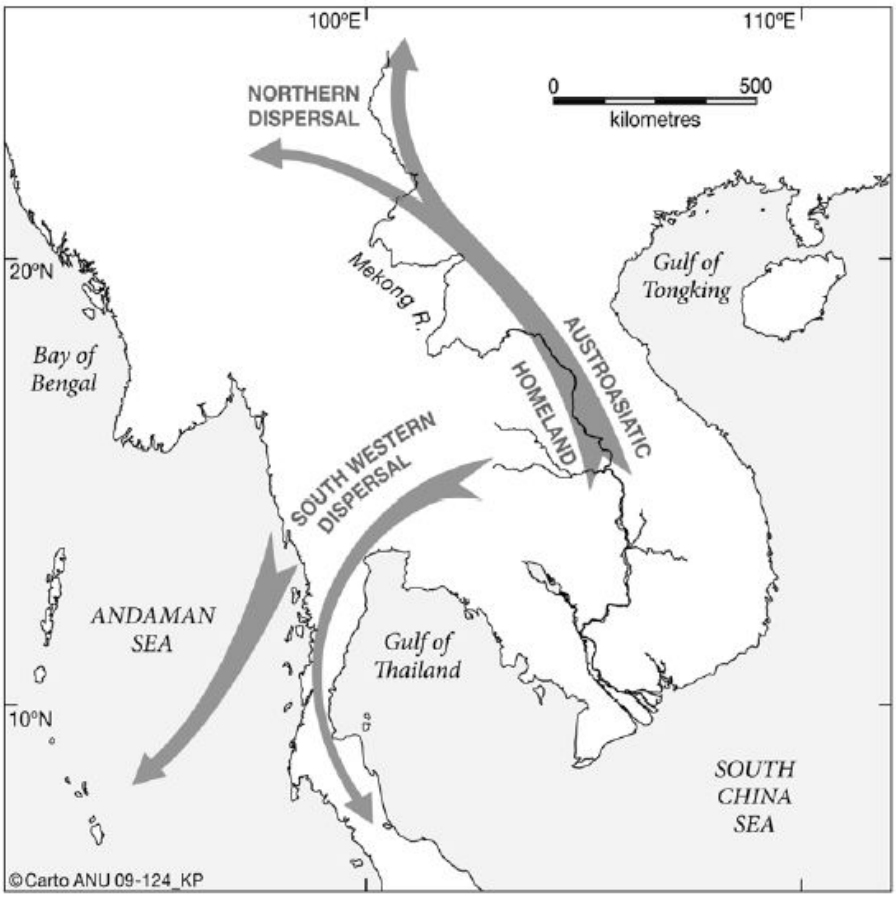
There may be a problem connected with the internal structure of Austroasiatic. Historical linguistics works best with apical structures where proto-forms can be attributed to different nodes following the identification of sound-shifts. But it seems likely Austroasiatic has a flat structure, its thirteen[1] branches developing from the diversification of a dialect chain rather than a series of hierarchical splits. This would make it 'innovation-linked' rather like Western Malayo-Polynesian; lexemes common to all branches might be rather rare and instead many terms would be shared by a series of near-contiguous branches.

Diffloth (2005) argued that the geographical dispersal characteristic of Austroasiatic reflects a quest for river valleys. Map 2 shows how the scattering of the branches of Austroasiatic indeed follows this pattern to a large extent, although Nicobaric, Aslian and Munda are exceptions. If the argument in Sidwell & Blench (2011) is correct, the flat array arises from an initial phase of aquatic dispersal, driven by improved boats, crops suitable for cultivation in humid soils (Blench 2011b). Blench (2011a) has also proposed that Austroasiatic speakers reached Island SE Asia, specifically Borneo, before being assimilated by expanding Austronesian speech communities. This in turn reflects the early spread of the SE Asian Neolithic, which can tracked through sites exhibiting a characteristic artefact cluster, including 'incised and impressed' pottery (Rispoli 2008; Higham et al. 2011). In this model, the original homeland of Austroasiatic would have been in

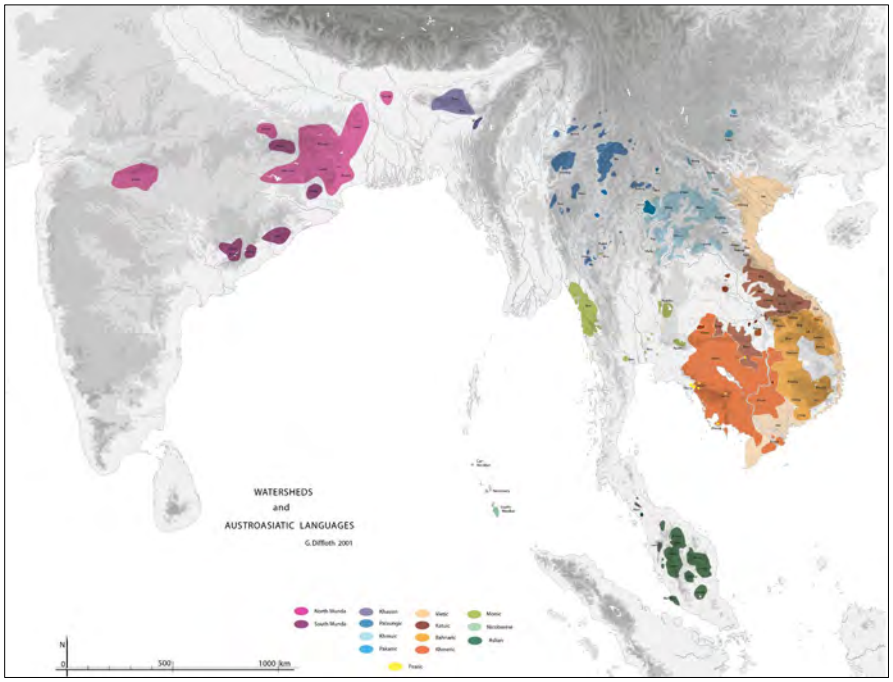
[1] Or fourteen (see Blench & Sidwell 2011).

the middle Mekong and speakers of the gradually differentiating dialects would have dispersed both north and southwest, as shown in Map 1.

Map 1. *Proposed dispersal pattern of Austroasiatic (Sidwell & Blench 2011)*



Map 2. *Austroasiatic languages (Diffloth 2001)*



However, it is not necessary to subscribe to this model, nor even to a middle Mekong homeland, to accept the importance of rivers in stimulating the early dispersal of Austroasiatic. The Mekong is the most biodiverse river in the world, surpassing even the Amazon, with over 1200 species of fish (Rainboth 1996) and many Austroasiatic subgroups are situated within its basin. If aquatic subsistence was indeed important at the period of dispersal, then this should be reflected in the lexicon. A preliminary attempt to draw attention to some possible common forms is given in Sidwell & Blench (2011: Table 5). However, this was still framed in the discredited Mon-Khmer model. This paper[2] is an attempt to draw together the lexical evidence for Austroasiatic, making no presumptions about subgroupings. Table 5 shows the lexical categories for common roots relevant to aquatic subsistence.

Table 1: *Lexical categories relevant to aquatic subsistence*

Category	**Class**	**Examples**
Rivers		
Water transport	boat	
Fauna	fish	
	crustaceans	
	chelonians	
	others	otter, crocodile, heron
Capture techniques	fish traps	
	fish poison	

2. Data

The core of this paper consists of tables of lexemes which are either cognate or are borrowings. The focus is on Austroasiatic languages, but where I consider there are cognates in other language phyla, I have also included these, together with my hypotheses concerning the direction of borrowing. Some regional words have a broader distribution, such as the main word for 'river', whose cognates seem to encompass almost every type of water-body from the Mekong to a puddle.

For Austroasiatic, the main source for citations is the online Mon-Khmer Etymological Dictionary (MKED)[3], which provides access to many of the important lexical sources, retranscribed to IPA where this is relevant, for example in Nicobarese. Where no source is given for the data, the reference can be found in the MKED. Many of these etymologies were first pointed out in Shorto (2006) and where this is the case, I have cited the number of Shorto's reconstructions beneath those tables (e.g. Shorto #115).[4] I have usually cited reconstructions for a subgroup where these are available. Occasionally, when a single form is attested across many languages, I have given a 'common' form, such 'Common Pearic' where the data seems to warrant it. Two groups of Austroasiatic, Munda and Aslian, have undergone extensive relexification, such that older roots which may have shown cognacy have been replaced. Typical Munda dictionaries show widespread borrowing from Hindi or other Indic languages, while Aslian (more surprisingly) borrows extensively from Malay, even in the area of fishing and foraging. As for other language phyla, Hmong-Mien material is cited from Ratliff (2010). For Sino-Tibetan languages I have used the online STEDT database[5], occasionally supplemented by my own field materials. Austronesian is largely drawn either from Wolff (2010) on Robert Blust's online Austronesian Comparative Dictionary[6]. There is no convenient online source for Daic languages, so I have referenced individual online publications.

---

[2] This paper is a revised version of one given at the VII Austroasiatic Meeting, held in Kiel, September 2017. Thanks to the audience and the reviewers for comments. Thanks to Nicole Kruspe for additional comments on the Aslian material.

[3] http://www.sealang.net/monkhmer/dictionary/

[4] It is also worth noting that Shorto did not have access to online search tools.

[5] http://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl

[6] http://www.trussel2.com/acd/

A significant problem is the extent to which these terms can be regarded as reconstructions. Sidwell & Rau (2015) have put forward Proto-Austroasiatic reconstructions. The initial section of Shorto (2006) deals with the reconstruction of Proto-Mon-Khmer phonology, omitting Munda and Nicobarese. However, the actual lexical evidence for individual phonemes is highly variable and in the case of words tabulated in this paper, attestations are usually in a small number of branches. Shorto did not have the advantage of published reconstructions at branch level, such as are now available for Bahnaric (Sidwell 2000; Sidwell & Jacq 2003), Katuic (Sidwell 2005), Khmuic (Sidwell 2014), Khasian (Sidwell 2014) and Palaungic (Sidwell 2015). Nonetheless, this does not yet constitute enough evidence for a starred form for the items discussed here. The existing comparative evidence suggests that these items are potentially reconstructible.
Frankly, the literature is marred by imprecise definitions and a lack of interest in ethnoscientific terminology. One assumes that fishing peoples such as the Nicobarese must have hundreds of terms for marine and possibly freshwater fish species, but if so, this is not recorded in the literature. The situation is similar for other aquatic species on the Mekong and Salween systems. The quality of recorded fish names evidenced in Ross *et al*. (2010) for Oceanic makes possible a fine level of detail not possible for Austroasiatic.

3. Rivers

Mainland SEA has a widespread stem applied to watercourses, or by extension valleys, *#ro[o]ŋ*, which can take a variety of prefixes. The simplest form of the root generally seems to mean channel, gully or ditch, as shown in Table 2.

Table 2: *A SEA regional term for -ro(o)ŋ for 'ditch, canal'*

Phylum	**Branch**	**Language**	**Attestation**	**Gloss**
Austroasiatic	Bahnaric	Rengao	roːŋ	drainage channel, side-channel of river
	Katuic	P-Katuic	*rɔɔŋ	stream, river
	Monic	Mon	pəròŋ	gully
	Palaungic	P-Palaungic	*rɔŋ	river valley
	Vietic	Vietnamese	giòng	current, flow, stream
Sino-Tibetan	Lolo-Burmese	Burmese	mroŋ	gully
Sino-Tibetan	Lolo-Burmese	Burmese	mroŋː	canal
Daic	Tai	Thai[7]	rôŋ	channel, ditch
Daic	Tai	Shan[8]	hɔ̀ŋ3	gully, river

(Shorto #668)

However, his stem seems to have acquired a widespread prefix, k~kh very early, which acted to increase the size of rivers to which it applied. It must have subsequently spread independently from *#ro[o]ŋ*, as it is attested in many subgroups where the bare root is unknown. In this form it is often applied to the Mekong, whose name is incorporated in it, and elsewhere the Salween. Table 3 shows that it is attested in all the major phyla of MSEA except Hmong-Mien.

[7] Thai citations from http://sealang.net/thai/dictionary.htm.
[8] Shan citations from http://sealang.net/shan/dictionary.htm.

Table 3: *A SE Asian regional term for 'river', 'valley'*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Bahnaric	PB	*krɔːŋ	river
	Khmuic	Khmu Yuan	krɔ́ːŋ	Mekong
	Mangic	Bolyu	huːŋ[13]	river, ditch
	Monic	P-Monic	*krooŋ	stream, creek, river
	Munda	Kharia	khirom	large river
	Palaungic	proto Waic	*klɔŋ[9]	river
	Palaungic	Palaung	klɔŋ	quantifier for watercourses
	Pearic	Pear [Kompong Thom]	kraŋ	large river
	Vietic	P-Vietic	*k-rɔːŋ	river
Austronesian	Chamic	Proto-Chamic[10]	*krɔːŋ	river
	Chamic	Acehnese	kruəŋ	river
Daic	Tai	Thai	khoŋ	Salween
Daic	Tai	Shan	khōŋ4	Salween
Sino-Tibetan	Kachinic	Kachin	kruŋ	valley
	Lepcha	Lepcha	kyoŋ	valley
	Sinitic	Old Chinese	*k-hlun	river
	Tibetic	Written Tibetan	kluŋ	river
	Lolo-Burmese	Old Burmese	kʰloŋ	river

(Shorto #733)

A distribution like this makes it difficult to establish where the extended root originated. However, for Austroasiatic it is lacking only in the southern languages, Aslian and Nicobaric, whereas it is highly restricted in Sino-Tibetan, having been picked up by Sinitic and Tibetic, but not attested at all in western languages. This suggests a borrowing into Sino-Tibetan, Daic and Austronesian. There is another, apparently unrelated root in Austroasiatic which is applied only to large rivers and by extension the sea (Table 4). This is attested in Nicobaric, apparently replacing the *#loŋ* root.

Table 4: *Evidence for reconstructing 'large river, sea' in Austroasiatic*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Bahnaric	Chrau	[daːʔ] nleː	large river
	Katuic	Kuy	thlèː	sea
	Khmer	Khmer	tùənlè	(large) river
	Munda	Kharia	dhara	stream, river
	Nicobaric	Nancowry	kamalɛ́ʔ	sea
Austronesian	Chamic	Cham	tathiʔ	sea

(Shorto #210)

[9] One reviewer requested I note the possibility that forms with medial -l- are an unrelated etymon.
[10] Chamic data and reconstructions are from Thurgood (1999).

Ratliff (2010) reconstructs **ɢlaew*^{A} for 'river' but one of only two exemplified languages, the West Hmongic Luopohe, has *ʁlei*^{A}, which may be related to this root. Finally, Austroasiatic may have a number of local roots which refer to water currents. Table 5 puts these forward as suggestions only. They may prove to be more widespread or possibly just coincidence. Shorto (entry 1686) merges the first two together with roots meaning 'pour', 'dribble' etc. but these are provisionally kept apart.

Table 5: *Possible Austroasiatic roots relating to river currents*

Branch	**Language**	**Citation**	**Original Gloss**
Khmeric	Surin	wuaʔ	to be strong, swift, rapid (current)
Nicobaric	Nancowry	wua	current (of water)

(Shorto #1686)

Bahnaric	Sre [Koho]	cɔ:	to lead (by a current)
Palaungic	PPa	*cɔɔr	current

(Shorto #1686)

Monic[11]	Mon	həmɔ̀	flow, current, flood
Nicobaric	Car	ha-nɛ:-mə	current of water

4. Water transport

The rivers and seas of MSEA throng with a wide variety of vessels, and in Vietnam, some early river transport has been excavated, preserved in silt, so we can get a sense of the construction of these early river-craft. One of these, oddly, turns out to exhibit a constructional technique otherwise only reported from the Mediterranean (Bellwood et al. 2007). Recent research in the region of the South China Seas only serves to underline the intensity of maritime traffic from the early Neolithic (Bellwood 2017).

Austroasiatic has two widespread roots for 'boat' which appear to be indigenous. The root *#duuk* is discussed in Diffloth (2011) and is confined to core families in the Central Mekong area, and was presumably lost as Austroasiatic spread west and south. Table 6 shows the reflexes of this root.

Table 6: *The #duuk root for 'boat' in Austroasiatic*

Branch	**Subgroup, language**	**Citation**
Bahnaric	Proto-Bahnaric	*duuk
Katuic	Proto-Katuic	*duuk
Khmeric	Khmer	tuuk
Monic	Nyah Kur	thù:k
Nicobaric	Nancowry	düe
Pearic	Common	#tɔ̀k
Vietic	PV	*ɗu:k

(Shorto #336)

Pearic may well be borrowed from Khmer. The implosive initial in Vietic is probably not original. If Malay *bidok* 'canoe' is connected this must be a recent borrowing into Malay. The other root for 'boat' is *#C.lɔɔŋ*, which has a more scattered distribution and is found only sporadically in some branches (Table 7). However,

[11] Shorto (B94) reconstructs a verb **t.huum*, based on Palaung *thom*, to flood' and Lawa *thuam* 'to be flooded', which is provisionally treated as distinct.

it is clearly attested in Munda, which makes it more secure for proto-Austroasiatic than *#duuk*. The three different attestations in Mon show the optionality of the prefix over time.

Table 7: *Another Austroasiatic root for 'boat'*

Phylum	Branch	Subgroup, language	Citation
Austroasiatic	Bahnaric	PB	*pluŋ
	Katuic	Ngeq	roŋ
	Khasic	P-Kha	*lɛɛŋ
	Khmuic	Khmu	cləːŋ
	Monic	Old Mon	dluŋ
	Monic	Middle Mon	gluŋ
	Monic	Mon	klɜ̀ŋ
	Munda	Kharia	ɖoloŋ
	Palaungic	P-Palaungic	*ɟnlɔɔŋ
Sino-Tibetan	Kuki-Chin	Lushai	loŋ
	Kuki-Chin	Kyo Chin	mlauŋ
	Naga	Chang	loŋ
	Lolo-Burmese	Written Burmese	lâuŋ
	Lolo-Burmese	Akha	lɔ̀

(Shorto #747)

Matisoff (2003) reconstructs **m.loŋ* for proto-Tibeto-Burman, although the distribution shows clearly this is a regional loanword, borrowed from Austroasiatic, only found in some Lolo-Burmese languages and the Naga-Kuki-Chin complex. The similarity with the reconstructions for 'river', 'valley' (Tables 2 and 3). suggests the possibility of a nominalisation although there is no direct evidence for this. One term for boat is attested in both Austronesian and Austroasiatic, whose reflexes are laid out in Table 8.

Table 8: *A SEA regional term for 'boat'*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Aslian	Jahai	kupon	boat
	Bahnaric	Biat	baŋ	coffin
	Aslian	Semai, Temiar	kapal[12]	boat
	Monic	Old Mon	kḅaŋ	ship
	Mangic	Mang	ɓaaŋ	ferry, boat
	Nicobaric	Central	kopòk	boat, ship
Austronesian	PAN		*qabaŋ	boat, canoe
	Taiwan	Siraya	avaŋ	canoe
	Taiwan	Favorlang	abaŋɯ	boat
	Philippines	Magindanao	kaban	boat
	Philippines	Tagalog	baŋkaʔ	canoe
	Philippines	Sulu	guban	boat
	Ibanic	Iban	boŋ, buuŋ	long, shallow boat,

[12] ? < Malay or Tamil

Phylum	Branch	Language	Attestation	Gloss
	Chamic	PC	*bɔɔŋ	coffin
	Malayic	Moken	kabaŋ	boat
	Malayic	Malay	kəbaŋ	vessel
	Malayic	Sekah	gobaŋ	boat
	Barrier	Nias	owo	boat
	Barrier	Sichule	ofo	boat
	Bima-Sumba	Sawu	kowa	boat

(Shorto #633)

The lack of Muṇḍā and Khasi cognates makes it difficult to assign this term to proto-Austroasiatic; and it does not reconstruct to the proto-language in any Austroasiatic branch. Nonetheless the Nicobarese and Aslian forms are clearly not just Malay borrowings, and the stem must be assigned to an early period in Austroasiatic expansion. Clearly these common forms are a consequence of early interactions with Austronesian maritime culture. Mahdi (1999) has identified the links, both cultural and lexical, between coffins and boats, attested in Bahnaric. The widespread Austronesian *#baŋka* for 'canoe' (e.g. Wolff 2010) is surely a reversal of the elements of *#kabaŋ*.

5. River and sea fauna

5.1 Fish

Reconstructing individual fish species in Austroasiatic is problematic since the lexical sources are weak on scientific names. However, Table 9 shows a generic term for 'fish', **kaʔ*, attested in nearly every branch.

Table 9: *A general Austroasiatic term for 'fish'*

Branch	Language	Attestation
Aslian	P-Aslian	*kaːʔ
Bahnaric	Sre	ka
Katuic	Kuy	kaː
Khasic	PK	*kʰa
Khmeric	Khmer	kaː-[moŋ &c.] (in compounds)
Khmuic	Kammu-Yuan	káʔ
Monic	Old Mon	kaʔ
Munda	Kharia	ka-[13]
Nicobaric	Nancowry	ká
Palaungic	Lawa	kaʔ
Vietic	Vietnamese	cá

(Shorto #16)

This root is widespread in the region, turning up in Austronesian as *ikan* and possibly even in Japanese *sakana* (possibly related to PAN *Sikan). Two species of catfish are attested in a more restricted set of Austroasiatic branches, as in Tables 10 and 11. The second root is more doubtful, as the semantic shift to

[13] Pinnow (1959:64).

'sawfish' in Khmer is a bit unlikely. Another species described as a 'serpent headed fish' and is most likely to be a snakehead (*Channa* spp.)[14] (Table 12).

Table 10: *Catfish sp. in Austroasiatic*

Branch	Language	Attestation	Gloss
Bahnaric	Sedang	b.lɔŋ	
Bahnaric	Tarieng	loːn	
Katuic	Ngeq	k.loː	
Khmeric	Khmer	c.laŋ	prob. *Macrones* sp.
Palaungic	Lamet [Lampang]	ləːn	

Table 11: *Catfish sp. in Austroasiatic*

Branch	Language	Attestation	Gloss
South Bahnaric	Chrau	[kaː] kɔː	catfish
Monic	Mon	[kaʔ] həkɔʔ	catfish sp., *Clarias magur*
Khmeric	Khmer	thkɔː	sawfish

(Short #22)

Table 12: *Fish sp. in Austroasiatic*

Branch	Language	Attestation	Gloss
Bahnaric	Sedang	rə.lɔ̃n	fish sp.
Katuic	Ngeq	k.luan	fish sp.
Nicobaric	Nancowry	lúan	salt-water eel

Eel is widely attested in Austroasiatic and the root appears to be borrowed into Sino-Tibetan and Austronesian (Table 13). The cognacy of the Sino-Tibetan forms is uncertain. This word is poorly attested in many Sino-Tibetan languages. Austronesian cognates are clearly not PAN, which is something like **tuła* (Wolff 2010). Shorto (2006 No. 461) proposes *phook ~ *pʔook a form for 'fish-paste', the fermented paste common as a food flavourer in SE Asia. He notes the similarities to the word for 'fish-bone' (*prʔook) suggesting possible problems with the reconstruction.

[14] Diffloth (1979).

Table 13: *'Eel' in SE Asian language phyla*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Bahnaric	PB	*-duŋ	
	Katuic	PK	*ʔnduŋ	
	Khmer	Surin Khmer	ntuaŋ	
	Khmuic	Khmu	ʔɔntùəŋ	???
	Monic	Nyah Kur	nthòoŋ	swamp eel
	Monic	Mon	daluŋ	eel
	Palaungic	Lamet [Nkris]	təlaːɲ	eel
	Munda	Mundari	ɖuŋ.ɖuŋ	long, very slender fish
	Munda	Kharia	ɖuŋɖuŋ	eel
	Pearic	PP	*ml(ɔː)ŋ	eel
Sino-Tibetan	Sakish	Kadu	patùn	eel
	Isolate	Kman	p.lun	eel
Austronesian	Philippines	Cebuano	induŋ	moray eel sp.
	Borneo	Iban	lundoŋ	eel
	Sumatra	Karo Batak	duŋduŋ	eel
	Malayic	Acehnese	ndoŋ	eel
	Malayic	Acehnese	linɔŋ	eel sp.
	Malayic	Cham	lanuŋ	eel
	Malayic	Malay	[ular] londoŋ	sea-snake

(Shorto #579)

5.2 Crustaceans

In many ways, crustaceans seem to be more salient in Austroasiatic than fish. Table 14 shows a probable Austroasiatic root for 'prawn, shrimp'.

Table 14: *An Austroasiatic root for 'prawn'*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Bahnaric	Nyaheun	cɔŋ	prawn, shrimp
	Katuic	PK	*ʔɲcɔŋ	shrimp
	Khmer	Surin	trej-kɔːŋ	shrimp, prawn
	Khmuic	Phong	paː kuːŋ	shrimp
	Munda	Santal	icaʔ	
	Nicobaric	Nancowry	ʃoaŋ	marine shrimp
	Palaungic	Danaw	mɑiʔ³ tɔŋ⁴ kɔŋ¹	prawn
	Pearic	Chong [Kompong Som]	pkɔːŋ	prawn
	Vietic	Thavung	kɔːŋ	prawn
Daic	Tai	Proto-Zhuang-Tai	*kuŋ.C	shrimp

Phylum	Branch	Language	Attestation	Gloss
	Kra	Lakkia[15]	tsoŋ.[3]	shrimp
	Kra	Biao	kuŋ.[3]	shrimp
Sino-Tibetan	Kuki-Chin	proto-Kuki-Chin	ŋaay kuang	shrimp/prawn
	Naga	Ao	[a]-kuŋ	prawn
	Bodo-Garo	Deuri	cicô	shrimp/prawn

The restricted distribution in both Sino-Tibetan and Daic clearly argues for borrowing into these two phyla. Tables 15 and 16 show more restricted roots for 'shrimp'.

Table 15: *A central Austroasiatic root for 'shrimp'*

Branch	Language	Attestation	Gloss
Bahnaric	Chrau	kəmvih	
Khmu		kɔmpɯ̀h	
Khmeric	Khmer	kɑmpɨh	
Pearic	Chong [of Samray]	kəmpɨːs	small river shrimp

(Shorto #1919)

Table 16: *Minor Austroasiatic roots for 'prawn, shrimp'*

Branch	Language	Attestation
Khasic	Pnar [Rymbai]	cʰɨŋktat
Khmuic	Khmu	cntah
Palaungic	PP	*kntaas

(Shorto #1901)

Katuic	Kuy	kaː sum
Vietic	PV	*soːm

(Shorto #1419a)

Table 17: *A reconstruction for 'crab' in Austroasiatic*

Phylum	Language	Subgroup, language	Citation
Austroasiatic	Aslian	CA	#kantam
	Bahnaric	PB	*ktaːm
	Katuic	PK	*ktaam, *ʔataam,
	Khasic	PKha	*tʰaam
	Khmeric	Khmer	kdaam
	Khmuic	PKhm	*ktaːm
	Mangic	Mang	taːm[6]
	Munda	PNM	*kaʈkɔm
	Monic	PM	*kntaam

[15] Kra-Dai citations are from Ostapirat (2000).

Phylum	Language	Subgroup, language	Citation
	Nicobaric	Nancowry	katəŋ-cafa[16]
	Palaungic	PP	*ktaam
	Pearic	Pear [Kompong Thom]	kta:m
	Vietic	PV	kta:m
Austronesian	Malayic	Malay	kətam
	Malayic	Moken	kətam
	Chamic	Acehnese	gɯtɯəm
Daic	Kra	Laha	khlaat

(Shorto #1348)

Table 17 shows a comparative set for 'crab' in Austroasiatic. Blust (ACD) reconstructs PAN **kətəm* for 'crab' which is evidently related. Table 18 shows a minor root for 'crab' in Austroasiatic.

Table 18: *A minor root for 'crab' in Austroasiatic*

Branch	Language	Attestation	Gloss
Bahnaric	Jru'	trʌp	crab sp.
Palaungic	Proto-Pramic	*hra:p	crab
Vietic	Proto-Vietic	*ra:p	crab

Photo 1. *Terrapins and fish in water plants on the Bayon (Author photo)*

5.3 Chelonians

Turtles and tortoises are found throughout the region and constitute an important source of food, but also play a significant role in mythology and oral traditions. They are regularly represented in the historical iconography, notably at Angkor Wat (***Photo 1***). Although the lexicographic literature is extremely vague on species, it is likely that if these were better identified, the different roots might apply to different species. When the Nicobarese migrated to the islands, they must have re-applied the names to marine species. In Table 19 **kaap* represents a widely attested root in Austroasiatic, present in both the Nicobars and Aslian, but lost in western subgroups such as Munda and Khasic.

[16] Non-edible land crab.

Table 19: *A reconstruction for 'tortoise, turtle' in Austroasiatic*

Branch	Language	Attestation	Gloss
Aslian	Jahai	kɔh	tortoise sp.
Bahnaric	P-Bahnaric	*kɔːp	tortoise
Katuic	P-Katuic	*ʔakɔɔp	turtle
Khmuic	Tai Hat	kuːp	turtle
Palaungic	Palaung	kəpkəp	tortoise
Nicobaric	Car	kap	tortoise
Nicobaric	Nancowry	kap-ka	green turtle (*Chelonia virgata*)
Vietic	Chứt [Arem]	kòːp	shell (crab, tortoise)

(Shorto #1235)

Table 20 shows a more uncertain root, which was given by Shorto (2006) as proto-Mon-Khmer. The vowels in Monic are irregular, unless this is a different root. The ku- prefix, added in Munda is striking, because the root then resembles both the Malayic forms and also, more strikingly, those found all over Sub-Saharan Africa (Blench 2008). Table 21 shows a root, **tɪpaʔ*, which seems restricted to freshwater turtle species.

Table 20: *A common form for 'turtle' in Austroasiatic*

Phylum	Branch	Language	Citation	Gloss
Austroasiatic	Bahnaric	Stieng	blɔːu	tortoise shell
	Khasic	PK	*-ruʔ	turtle
	Khmeric	Surin	nʌːʔ	turtle
	Munda	Sora	'ku(ː)luː-n	turtle
	Munda	Kharia	'kulu	turtle
	Monic	Mon	naoh	turtle
	Palaungic	Riang [Sak]	ruːs²	tortoise, turtle
	Vietic	PV	ʔa-rɔː	tortoise
Austronesian	Malayic	Malay	kura-kura	tortoise

(Shorto #B118)

Table 21: *A reconstruction for 'freshwater turtle' in Austroasiatic*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Aslian	Jahai	pjðŋ	turtle
	Bahnaric	PB	*tɒa:	turtle
	Katuic	PK	*tɒaa	soft shelled turtle
	Khmuic	PKhm	*tmɒaʔ	snapping turtle
	Mangic	Mang	ma:[1] ɒa:[2]	turtle, tortoise
	Nicobaric	Car	təkurəɒə	land turtle
	Pearic	Chong [of Chantaburi]	kaɒʰa:	turtle soft-shelled
	Pearic	Chong [Kasong]	lɒʰa:	turtle soft-shelled
Sino-Tibetan	Mruish	Hkongso	pʰáˆ	soft shelled turtle

(Shorto #104)

Hkongso must be a borrowing from Austroasiatic. Possibly compare proto-Hlaic **tʰɯ:p* 'point-nosed turtle' 鱉 (Norquest2007). Table 22 shows several low-frequency roots for 'tortoise/turtle'. The first set, linking Bahnaric/Monic/Khasian is highly uncertain.

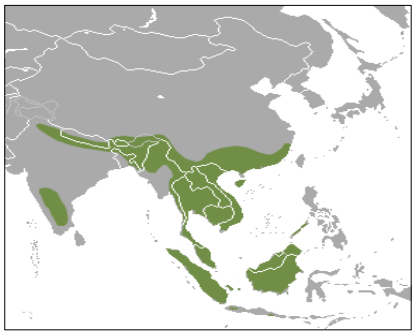
Table 22: *Low-frequency roots for 'tortoise/turtle' in Austroasiatic*

Branch	Language	Attestation	Gloss
Bahnaric	Mnong [Rölöm]	kra:	large turtle
Khasis	Khasi	dka:r	tortoise
Monic	Mon	klao	targe tortoise sp.
Bahnaric	Sapuan	ntə:k	tortoise, turtle
Nicobaric	Nancowry	ʔok-teka	tortoise
Bahnaric	Jruq	tmom	turtle (land)
Katuic	PK	*tmoom	turtle

5.4 Others

A few species characteristic of riverine habitats have significant reconstructible roots in Austroasiatic. These are the otter, the crocodile, and the heron. There are two species of otter found throughout the MSEA region, the oriental small-clawed otter, *Aonyx cinerea*, and smooth-coated otter, *Lutrogale perspicillata*. Map 3 and Map 4 show the range of these species (from IUCN Red List of Threatened Species 2010).

Map 3: *Range of the Oriental small-clawed otter, Aonyx cinerea*



Map 4: *Range of t:he smooth-coated otter, Lutrogale perspicillata*

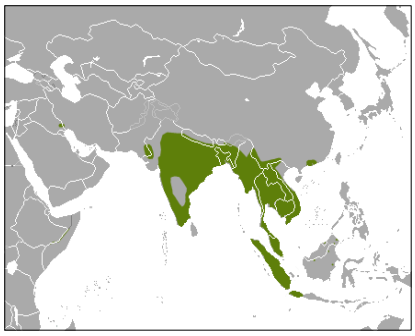


Table 23 shows a widespread Austroasiatic root for 'otter' borrowed into Chamic. It is likely that the original form was closest to Vietic **p-se:ʔ* which accounts for the long vowel and final glottal in other reflexes. The fricative /s/ would have weakened to /h/ in some branches, while Khasi was subject to prefix replacement.

Table 23: *A SE Asian root for 'otter'*

Phylum	Language	Subgroup, language	Citation
Austroasiatic	Aslian	Semelai	bəhɛʔ
	Bahnaric	Nvaheun	ɒhie
	Bahnaric	Mnong [Rölöm]	bhi:ŋ
	Katuic	PK	*ɒhav
	Katuic	Bru	ɒhɛ
	Khasic	Khasi	kəsiʔ
	Khmeric	Khmer	ɒhè:
	Monic	PM	*ɒhɛɛʔ
	Pearic	PP	#ɒʰe:
	Vietic	PV	*ɒ-se:ʔ
Austronesian	Chamic	PC	*buhay

(Shorto #104, #A50)

Another member of the regional riverine fauna is the crocodile. Crocodiles are regularly represented in historical sources, such as on the Bayon (***Photo 2***). Table 24 shows a widespread root for 'crocodile' which is missing in western branches.

Table 24: *An Austroasiatic root for 'crocodile'*

Branch	Language	Attestation	Gloss
Bahnaric	PWB	*krbɨw	crocodile
Katuic	PK	*krɓəə	crocodile
Khmeric	Khmer	krɑpəə	crocodile
Khmuic	Khmu [Cuang]	ckʰrɛ: (<Tai)	crocodile
Nicobaric	Car	rew[17]	crocodile
Pearic	Pear [Kompong Thom]	krəpə: tiek	crocodile

(Shorto #115)

[17] Not necessarily cognate.

Photo 2: *Crocodile catching fish on the Bayon (Author photo)*

Table 25 shows two local roots for 'crocodile' in Austroasiatic. They are conceivably related, although reflexes with front and back vowels in Vietic make this doubtful.

Table 25: *Local roots for 'crocodile'*

Branch	**Language**	**Attestation**
Pearic	Chong [of Kompong Som]	lkọ:
Pearic	Chong	rəkɔ̀ɔ
Vietic	Mương [Son La]	kʰuː³
Khmuic	Khmu [Cuang]	ckʰrɛː
Palaungic	Lamet [Lampang]	səkheːʔ
Vietic	Thavung	khɛ̂̃ː (?<Tai)

Tables 26 and 27 show two widespread roots for fishing birds. *#kok* seems to mean 'heron' underlyingly, but it has shifted to hornbill in both Aslian and Khasic and to cormorant in Vietnamese. Shorto (280) reconstructs the meaning as 'egret' but the evidence from additional cognates points towards a waterbird. Table 27 shows what is clearly a local root for 'pelican' in some central branches of Austroasiatic. The table also includes proposed cognates in non-Austroasiatic languages, but I have not been able to confirm these.

Table 26: *An Austroasiatic root #kok for 'heron', 'fishing bird'*

Phylum	**Branch**	**Language**	**Attestation**	**Gloss**
Austroasiatic	Aslian	Semai	dkuuk	helmeted hornbill, *Rhinoplax vigil*
	Bahnaric	PB	*kɔːk	egret, heron
	Katuic	Pacoh	ka.laːŋ kuːk	pelican
	Khasi	Khasi	koh-[karang]	hornbill
	Khmeric	Khmer	kok	heron, egret
	Munda	Kharia	kɔlɛʔ	heron
	Palaungic	PPa	*kVk	heron
	Vietic	Vietnamese	cốc	cormorant
	Malayic	Javanese	blekok	k.o. heron
Austronesian	Chamic	Acehnese	blökɔʔ	heron

(Shorto #278, #280)

Table 27: *An Austroasiatic root for 'pelican'*

Phylum	Branch	Language	Attestation	Gloss
Austroasiatic	Khmeric	Khmer	tuŋ	pelican (*Pelecanus* sp.)
Austroasiatic	Monic	Mon	tàŋ	bird including stork and pelican
Austroasiatic	Pearic	Chong [of Samray]	tuŋ	grey pelican (*Pelecanus philippensis*)
Sino-Tibetan	Lolo-Burmese	Burmese	duṁ:	not in dictionary
Daic	Tai	Thai	nók grà tung	pelican
Austronesian	Chamic	Cham	kaḍuṅ (!).	pelican

(Shorto #572)

6. Capture techniques

Any ethnographic museum in the region usually displays an abundance of fish traps, storage baskets and other devices. These are extraordinarily diverse and few dictionaries capture their specificity. shows some non-return traps made by the Khasi; the fish swims along the funnel and then is unable to turn back and escape. Traps of this type are made throughout the region, but we are not yet in a position to reconstruct individual types. Table 28 shows an Austroasiatic root for 'fish trap' (type unspecified).

Table 28: *An Austroasiatic root for 'fish-trap'*

Language	Subgroup, language	Citation	Original Gloss
Bahnaric	Sedang	trɔ̃	fish trap
Khmeric	Surin	trù:	bamboo fish trap
Katuic	Kui	thrụ:	cylindrical fish trap made of bamboo strips
Monic	Nyah Kur	thru	bamboo fish trap with a narrow neck
Munda	Kharia	londṛa	fish trap sp.
Pearic	Chong [Samre]	tûəɹ	fish trap
Vietic	Thavung	to:ŋ	fish trap

(Shorto #178 [under *druʔ])

There are no confirmed external cognates but Karo Batak has *tuar* 'small fish-trap placed with opening stream-upwards' which could be coincidence. Matisoff (2003: 285) reconstructs **tuŋ* for proto-Lolo-Burmese 'set a trap'. Given that no words for actual fish-trap in Sino-Tibetan seem to be shared with Austroasiatic, this may be just coincidence.

to 3. *Khasi bamboo fish-traps (Don Bosco Museum, Shillong)*

Another widespread fishing technique is the scoop-net or landing net, a large loose cord net for capturing fish that shoal. Table 29 shows a regional term for 'scoop net' recorded in three Austroasiatic branches.

Table 29: *A restricted Austroasiatic root for 'scoop net'*

Language	Subgroup, language	Citation	Original Gloss
Khmeric	Khmer	cʰnìːəŋ	scoop-net
Khmeric	Khmer	tnɑɑŋ	fishing net, landing net, scoop net
Monic	Mon	càiŋ; ~ (*ɟrɟaaŋ >)	net
Monic	Mon	hnàiŋ	net
Palaungic	Lawa Bo Luang	ʔacɯaŋ	to net [fish]

(Shorto #536)

Photo 4 shows a scoop-net depicted with considerable verisimilitude on the Bayon at Angkor in the 12th century.

Photo 4: *Khmer scoop-net on the Bayon (Author photo)*

Table 30: *A restricted Austroasiatic root for 'fish with line'*

Language	Subgroup, language	Citation	Original Gloss
Monic	Mon	dɔn (k)ḍan,	fish with a line
Monic	Proto-Nyah Kur	*cərndɛ̱ŋ	fishhook
Nicobarese	Central	koron-[hətə]	to fish with a line

(Shorto 1161 *kɗɔn)

Finally, a common method of catching fish in MSEA is the use of vegetable poisons. Thrown into a pond or pool, they stun the fish, which rise to the surface, without making them toxic. Table 35 shows a root which is spread across much of the range of Austroasiatic, although only attested in four families.

***Table 31:** An Austroasiatic root for 'to poison fish'*

Language	Subgroup, language	Citation	Original Gloss
Bahnaric	Proto-Bahnaric	*kraw	to poison (fish with plant)
Katuic	ProtoKatuic	*kraw	poison (fish)
Khasic	Khasi	*kʰəriaw	fish poison
Nicobaric	Car	ka-jaw	to poison fish (with the grated seeds of the *kin-yav*)

(Shorto #1846, also perhaps #1461)

7. Conclusions

A combination of linguistic geography and historical linguistics, suggests the possibility that Austroasiatic represents a 'flat array' of languages, and that this is due to an early riverine dispersal (Sidwell and Blench 2011). Using a 'centre of gravity' argument, the Middle Mekong is proposed as the original nucleus of dispersal. The period of dispersal is identified with the SE Asian Neolithic, currently dated to ca. 4000 BP. Although early Austroasiatic speakers were clearly crop producers, growing both taro and rice, if they were largely following river basins, aquatic technology and subsistence must have been highly salient in their vocabulary. The paper shows that a number of lexical items can be shown to be common to many of the branches of Austroasiatic, suggesting them as reasonable candidates for the proto-language. Other roots have more restricted distributions and apply only to local areas. Lexical data for Austroasiatic remains highly schematic and imprecise, as well as significantly defective for some branches. This suggests that with greater attention to biological and technical detail, it will be possible to refine some of the reconstructed items proposed here.

References

Bellwood, Peter 2017. *First islanders: prehistory and human migration in island Southeast Asia*. Hoboken: John Wiley & Sons.

Bellwood, Peter, Judith Cameron, Nguyen Viet and Bui Liem. 2007. Ancient boats, boat timbers, and locked mortise and tennon joints from Bronze-Iron Age Northern Vietnam. *International Journal of Maritime Archaeology* 36:1:2–20.

Blench, Roger M. 2008. The problem of pan-African roots. In: *In Hot Pursuit of Language in Prehistory*. John Bengtson ed. Amsterdam: John Benjamins.

Blench, Roger M. 2011a. Was there an Austroasiatic presence in island SE Asia prior to the Austronesian expansion? *Bulletin of the Indo-Pacific Prehistory Association* 30:133–44.

Blench, Roger M. 2011b. The role of agriculture in the evolution of Southeast Asian language phyla. In *Dynamics of Human Diversity: the case of mainland Southeast Asia*, ed. by Nicholas J. Enfield, 125–52. Canberra: Pacific Linguistics.

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.

Diffloth, Gerard. 2011. Austroasiatic word histories: boat, husked rice and taro. In: *Dynamics of Human Diversity in Mainland SE Asia*, ed. by N.J. Enfield, 295–313. Canberra: Pacific Linguistics.

Higham, Charles, Thomas Higham & & Amphan Kijngam. 2011. Cutting a Gordian Knot: the Bronze Age of Southeast Asia: origins, timing and impact. *Antiquity*, 85: 583–98.

Mahdi, Waruno. 1999. The dispersal of Austronesian boat forms in the Indian Ocean. In *Archaeology and Language III*, ed. by Roger M. Blench and Matthew Spriggs, 144–79. London: Routledge.

Matisoff, James A. 2003. *Handbook of proto-Tibeto-Burman*. Berkeley: University of California Press.

Norquest, Peter K. 2007. *A Phonological Reconstruction of Proto-Hlai*. PhD thesis. Department Of Anthropology, University of Arizona.

Ostapirat, Weera. 2000. Proto-Kra. *Linguistics of the Tibeto-Burman Area*, 23(1):1–215.

Rainboth, Walter J. 1996. *Fishes of the Cambodian Mekong*. Rome: Food & Agriculture Organisation.

Ratliff, Martha. 2010. *Hmong-Mien language history*. Canberra: Pacific Linguistics.

Rispoli, Fiorella. 2008. The incised and impressed pottery of Mainland Southeast Asia: following the paths of Neolithization. *East and West*, 57:235–304.

Ross, Malcom, Pawley, Andrew. & Meredith Osmond. eds. 2011. *The lexicon of proto-Oceanic: the culture and society of ancestral Oceanic society. 4: Animals*. Pacific Linguistics. Canberra: ANU.

Shorto, Harry L. 2006. *A Mon-Khmer comparative dictionary*. Pacific Linguistics 579. Canberra: ANU.

Sidwell, Paul. 2005. *The Katuic Languages: classification, reconstruction and comparative lexicon*. Munich: Lincom.

Sidwell, Paul. 2014. Proto Khasian: an emerging reconstruction. In *North East Indian Linguistics Volume 6*, ed. by Gwendolyn Hyslop, Linda Konnerth, Stephen Morey, and Priyankoo Sarmah,149–63.

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

Sidwell, Paul. 2000. *Proto South Bahnaric: a reconstruction of a Mon-Khmer language of Indo-China*. Canberra: Pacific Linguistics.

Sidwell, Paul. 2014. Khmuic classification and homeland. *Mon-Khmer Studies* 43(1):47–56.

Sidwell, Paul and Jacq, Pascale. 2003. *A Handbook of Comparative Bahnaric, volume 1-West Bahnaric*. Canberra: Pacific Linguistics.

Sidwell, Paul. & R.M. Blench. 2011. The Austroasiatic *Urheimat* : the Southeastern Riverine Hypothesis. In *Dynamics of Human Diversity*, ed. by : N. J. Enfield, 317–45. Canberra: Pacific Linguistics.

Sidwell, Paul and Felix Rau. 2015. Austroasiatic Comparative-Historical Reconstruction: An Overview. In *The Handbook of Austroasiatic Languages*, ed. by Mathias Jenny and Paul Sidwell, 221–63. Leiden: Brill.

Thurgood, Graham. 1999. *From ancient Cham to modern dialects: two thousand years of language contact and change*. Manoa: University of Hawaii press.

Wolff, John U. 2010. *Proto-Austronesian phonology*. 2 vols. Ithaca, NY: Cornell Southeast Asia Program Publications.