

JSEALS Special Publication No. 8

**PAPERS FROM THE 30TH MEETING OF
THE SOUTHEAST ASIAN LINGUISTICS
SOCIETY (2021)**



Edited by

Mark J. Alves and Paul Sidwell



**UNIVERSITY of
HAWAII
PRESS**

© 2022 University of Hawai'i Press
All rights reserved
OPEN ACCESS – Semiannual with periodic special publications
E-ISSN: 1836-6821
<http://hdl.handle.net/10524/52498>



Creative Commons License

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

JSEALS publishes fully open access content, which means that all articles are available on the internet to all users immediately upon publication. Non-commercial use and distribution in any medium are permitted, provided the author and the journal are properly credited.

Cover photo adapted from the cover of the JSEALS journal provided by Paul Sidwell.

JSEALS

Journal of the Southeast Asian Linguistics Society

Editor-in-Chief

Mark Alves (Montgomery College, USA)

Managing Editors

Nathan Hill (Trinity College Dublin, UK)

Sigrid Lew (Dallas International University, USA)

Paul Sidwell (University of Sydney, Australia)

Editorial Advisory Committee

Edith ALDRIDGE (Academia Sinica, Taiwan)

Nathan BADENOCH (Villanova University, USA)

Luke BRADLEY (University of Freiburg, Germany)

Marc BRUNELLE (University of Ottawa, Canada)

Christopher BUTTON (Independent researcher)

Kamil DEEN (University of Hawaii, USA)

Gerard DIFFLOTH (Cambodia)

Rikker DOCKUM (Swarthmore College, USA)

Ryan GEHRMANN (Payap University)

San San HNIN TUN (INALCO, France)

Kitima INDRAMBARYA (Kasetsart University, Thailand)

Peter JENKS (UC Berkeley, USA)

Mathias JENNY (University of Zurich, Switzerland)

Daniel KAUFMAN (Queens College, City University of New York & Endangered Language Alliance, USA)

James KIRBY (University of Edinburgh, Scotland)

Hsiu-chuan LIAO (National Tsing Hua University, Taiwan)

Bradley MCDONNELL (University of Hawai‘i at Mānoa, USA)

Alexis MICHAUD (CNRS (Le Centre National de la Recherche Scientifique), France)

Marc MIYAKE (The British Museum)

David MORTENSEN (Carnegie Mellon University, USA)

Peter NORQUEST (University of Arizona, USA)

John D. PHAN (Columbia University, USA)

Pittayawat PITTAYAPORN (Chulalongkorn University, Thailand)

Amara PRASITHRATHSINT (Professor Emeritus of Linguistics, Chulalongkorn University, Thailand)

Alexander D. SMITH (University of North Texas)

Kenneth VAN BIK (California State University, Fullerton, CA)

Heather WINSKEL (Southern Cross University, Lismore, Australia)

Alice VITTRANT (Aix-Marseille Université / CNRS-DDL, France)

Heather WINSKEL (Southern Cross University, Lismore, Australia)

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

<u>Introduction from the Volume Editors</u>	v
<u>Statement from the JSEALS Editor-in-Chief</u>	vi
<u>List of SEALS XXX (2021) Conference Presentation Videos</u>	vii
Historical Linguistics	
<u>Vietic and Early Chinese Grammatical Vocabulary in Vietnamese: Native Vietic and Austroasiatic Etyma versus Early Chinese Loanwords</u>	1
	Alves, Mark
<u>The Split of Proto-Tai Voiced Stop Consonants in Nung of Trang Dinh District</u>	25
	Hirano, Ayaka
<u>The Tai Meuay of Bolikhamxay Province (Laos), their ethnonym, and their language</u>	33
	Pacquement, Jean
<u>Austroasiatic Dispersal: the AA “Water-World” Extended</u>	56
	Sidwell, Paul
<u>More Austro-Tai Comparisons and Observations on Vowel Correspondences</u>	73
	Smith, Alexander
Syntax and/or Morphology	
<u>Basic Clauses in Itawit</u>	96
	Ayunon, Chirbet and Shirley Dita
<u>Directional Terms in Idu, a Language of Northeast India</u>	111
	Blench, Roger
<u>Notes On Nominal Marking And Noun Phrase Elements In Kinaray-A</u>	123
	Casalan, Marvin and Shirley Dita
<u>Nominalization and Its Various Functions in Liangmai</u>	134
	Daimai, Kailadbou
<u>Syllable Structure and Morphemicity in Tone Patterns on Verbs in Kanise Khumi</u>	145
	Ikeda, Elissa
<u>Syntactic and Semantic Boundness of Noun Phrases in Burmese</u>	163
	Jenny, Mathias

<u>On the Usages and Functions of the Case Marker -niu in Colloquial Liangmai</u>	176
Mataina, Wichamdinbo and Kailadbou Daimai	
<u>Kapampangan Perfective Infix -i-</u>	187
Pelagio, Earvin Christian; Rexie Quizon; Johans Cruz; Jemima Atok; and Kevin Ray Abesamis	
<u>Nominal Anchoring Functions of Porohanon Common Noun Markers</u>	195
Santiago, Vincent Christopher	
<u>Noncanonical pronouns in Vietnamese and Chinese</u>	207
Song, Chenchen and Li Nguyen	
<u>The Grammatical Nature of Sino-Vietnamese “Verb-Object Compounds”</u>	233
Takahashi, Yasunori	
Phonetics and/or Phonology	
<u>Bagobo-Klata Phonology</u>	244
Estrera, Edward	
<u>Pseudoregister in Pacoh: Preliminary Acoustic Analysis and Implications for a General Model of Pseudoregister Formation in Austroasiatic</u>	285
Gehrmann, Ryan	
<u>Towards a prosodic typology of Kherwarian Munda languages: Santali of Assam</u>	298
Horo, Luke and Gregory Anderson	
<u>Effects of voiceless and preglottalized nasals on F0 in Eastern Khmu (Khmumu’ Am)</u>	318
Kirby, James	
<u>Preliminary Acoustic Analysis of Minor Syllables in Kanise Khumi</u>	334
Tan, Wen Hao	

INTRODUCTION FROM THE VOLUME EDITORS

The Southeast Asian Linguistics Society (SEALS) was founded in 1990 by Martha Ratliff and Eric Schiller in 1990. Until 2020, SEALS conferences had been convened annually for 29 years, but the Covid-19 pandemic resulted in sudden and dramatic restrictions on travel. The original 30th meeting was to be hosted by the Department of Linguistics of the University of Hawaii, but that conference was cancelled, like so many other academic activities. As the pandemic lingered, it was decided that rather than lose another year, conference video presentations would be posted, and conference proceedings would follow. In part due to the shift to the JSEALS journal publications, the last SEALS conference proceedings volume for SEALS XVI was published in 2006, making these papers the first such SEALS proceedings in 17 years.

The papers in this special publication of JSEALS were written based on the video presentations of the SEALS 2021 conference launched online on the 1st of June. While this lacked the interactive aspects associated with a normal international conference, participants were enthusiastic. All the video presentations associated with the articles in these proceedings are listed with live links on page [vii](#) of this volume. The complete listing of videos and supplemental materials are available here:

<https://sites.google.com/site/sealsjournal/seals-and-jseals-history/seals-xxx-online-2021/seals-2021-program>.

The papers were written after the talks, and authors necessarily made modifications to their works, but the videos, much like handouts of a conference, provide a point of reference of the core ideas of the authors. This is the first time that SEALS conference presentation videos have been made available with the subsequent conference proceedings.

That over 350 pages of 21 articles have been assembled and published during challenging circumstances around the world attests to the vigor of the field of Southeast Asian linguistics.

Editors

Mark Alves
Montgomery College
JSEALS Editor-in-Chief
SEALS Publishing Officer

Paul Sidwell
University of Sydney
SEALS President

FROM THE JSEALS EDITOR-IN-CHIEF

This is the eighth JSEALS Special Publication. The goal of JSEALS Special Publications is to share collections of linguistics articles, such as select papers from conferences or other special academic events, as well as to offer a way for linguistic researchers in the greater Southeast Asian region to publish monograph-length works.

The volume contains 21 papers in total: five papers on historical linguistics, eleven papers on syntax and/or morphology, and five papers on phonetics/phonology. The languages covered in this volume are spoken in throughout the greater Southeast Asian region: Mainland Southeast Asia, Insular Southeast Asia, Southern China, and the Indian Subcontinent. The papers range from detailed descriptions of linguistic aspects of understudied languages to probing questions related to multiple groups of languages in the region.

We are very pleased that JSEALS is able to contribute to the sharing of quality linguistic research in Southeast Asia, and we welcome and encourage proposals for issues going forward.

Mark J. Alves

January 15th, 2022

Montgomery College

Rockville, Maryland

List of SEALS XXX (2021) Conference Presentation Videos

The list below contains all the videos associated with the articles in the SEALS XXX conference that were posted online on 1 June 2021. The entire collection of videos are housed at the SEALS-JSEALS site on Google (<https://sites.google.com/site/sealsjournal/seals-and-jseals-history/seals-online-2021/seals-2021-program>) and can also be found on the SEALS 2021 Youtube page (https://www.youtube.com/channel/UCTT4-3HEvM_Z65tHQHXfvFw). Additionally, the downloadable videos in the links below have permanent DOI identifiers provided by Zenodo and have Creative Commons 4.0 licensing, meaning they can be freely downloaded and shared.

1. **Alves, Mark**
Vietic and Early Chinese Grammatical Vocabulary in Vietnamese: Native Vietic and Austroasiatic Etyma versus Early Chinese Loanwords
<https://doi.org/10.5281/zenodo.5778105>
2. **Ayunon, Chirbet and Shirley Dita**
Basic Clauses in Itawit
<https://doi.org/10.5281/zenodo.5780295>
3. **Blench, Roger**
Directional Terms in Idu, a Language of Northeast India
<https://doi.org/10.5281/zenodo.5778109>
4. **Casalan, Marvin**
The Nominal Marking System of Kinaray-a
<https://doi.org/10.5281/zenodo.5778128>
5. **Daimai, Kailadbou**
Nominalization and Its Various Functions in Liangmai
<https://doi.org/10.5281/zenodo.5781319>
6. **Estrera, Edward**
Bagobo-Klata Phonology
<https://doi.org/10.5281/zenodo.5780339>
7. **Gehrmann, Ryan**
A Preliminary Study on the Acoustic Correlates of Pseudoregister in Pacoh
<https://doi.org/10.5281/zenodo.5780409>
8. **Hirano, Ayaka**
The Split of Proto-Tai Voiced Stop Consonants in Nung of Trang Dinh District
<https://doi.org/10.5281/zenodo.5780418>
9. **Horo, Luke and Gregory Anderson**
Towards a prosodic typology of Kherwarian Munda languages: Santali of Assam
<https://doi.org/10.5281/zenodo.5780420>
10. **Ikeda, Elissa**
Syllable Structure and Morphemicity in Tone Patterns on Verbs in Kanise Khumi
<https://doi.org/10.5281/zenodo.5780428>

11. **Mathias Jenny**
Syntactic and Semantic Boundness of Noun Phrases in Burmese
<https://doi.org/10.5281/zenodo.5780432>
12. **Kirby, James**
Acoustic Properties of Voiceless and Preglottalized Continuants in Khmu
<https://doi.org/10.5281/zenodo.5780434>
13. **Mataina, Wichamdinbo and Kailadbou Daimai**
On the Usages and Functions of the Case Marker -niu in Colloquial Liangmai
<https://doi.org/10.5281/zenodo.5780439>
14. **Pacquement, Jean**
The Tai Meuy of Bolikhamxay Province (Laos) and their language
<https://doi.org/10.5281/zenodo.5781298>
15. **Pelagio, Earvin Christian; Rexie Quizon; Johans Cruz; Jemima Atok; and Kevin Ray Abesamis**
Kapampangan Perfective Infix -i-
<https://doi.org/10.5281/zenodo.5781300>
16. **Santiago, Vincent Christopher**
Nominal Anchoring Functions of Porohanon Common Noun Markers
<https://doi.org/10.5281/zenodo.5781304>
17. **Sidwell, Paul**
Austroasiatic Dispersal: the AA “Water-World” Extended
<https://doi.org/10.5281/zenodo.5773247>
18. **Smith, Alexander**
More Austro-Tai Comparisons and Observations on Vowel Correspondences
<https://doi.org/10.5281/zenodo.5781307>
19. **Song, Chenchen and Li Nguyen**
Noncanonical Pronominal Items in Vietnamese and Chinese: Imposters or True Pronouns?
<https://doi.org/10.5281/zenodo.5781312>
20. **Tan, Wen Hao**
Preliminary Acoustic Analysis of Minor Syllables in Kanise Khumi
<https://doi.org/10.5281/zenodo.5781315>
21. **Takahashi, Yasunori**
The Grammatical Nature of Sino-Vietnamese “Verb-Object Compounds”
<https://doi.org/10.5281/zenodo.5781317>

VIETIC ETYMA VERSUS EARLY CHINESE LOANWORDS IN THE DOMAIN OF GRAMMATICAL VOCABULARY

Mark J. ALVES
Montgomery College
mark.alves@montgomerycollege.edu

Abstract

In this study, over a hundred and twenty Vietnamese function words with Vietic reconstructions and/or status as early Chinese loanwords (i.e., those borrowed before the Sino-Vietnamese layer associated with Late Middle Chinese character readings) have been assembled, grouped by subcategory, and given historical linguistic context. In this lexical data, most basic function words (e.g., numerals, pronouns, locative terms, time words related to the natural world) are native etyma, while most words with other functions (e.g., aspect, modality, comparison, time/aspect, etc.) are early Chinese loanwords. Despite the early borrowing of Chinese function words, current evidence suggests structural/typological changes in Vietic happened primarily in the later Viet-Muong period, and before that, this northern Vietic group did not undergo substantial morphophonological restructuring for several centuries from the time of language contact with Sinitic. This scenario implies significant Sinitic-Vietic bilingualism in those early centuries, but it also shows Vietic maintained sufficient sociolinguistic status to retain core function words. However, this early language contact set the stage for typological convergence into the second millennium and the speciation of Viet-Muong.

Keywords: Vietic, Austroasiatic, Chinese loanwords, function words, historical syntax
ISO 639-3 codes: vie, zhx, och, ltc, mtq, tai

1 Grammatical features and words in Vietnamese

Vietnamese is a Vietic language which has undergone morphophonological restructuring largely due to language contact with and lexical borrowing from Sinitic (and presumably Tai but with much less lexical exchange). This is also the case for the dozens of closely related Muong lects, together constituting the Viet-Muong sub-branch of Vietic. Based on historical, archaeological, and comparative linguistic data, the earliest period of substantial language contact between the pre-Proto-Viet-Muong (i.e., a stage after Proto-Vietic but before the speciation of the Viet-Muong sub-branch) speakers and Sinitic-speaking groups is from the Han Dynasty (202 BCE–220 CE), with the most significant early contact from the first century CE.

What was the degree of typological similarity or difference between Sinitic and Vietic at that time? There is a good deal of information about Sinitic linguistic structures at the time of Sinitic-Vietic contact. Textual data shows that Archaic Chinese had SVO structure, and modifiers preceded nouns in noun phrases (Aldridge 2015, Peyraube 1996). Reconstructions of Old Chinese phonology include presyllabic material (Baxter and Sagart's (2014a and 2014b)) and final fricatives *-s and *-h and a glottal stop but no tones (Pulleyblank 1977-1978, Zhengzhang 2000, Schuessler 2007, Baxter and Sagart 2014a and 2014b). Old Chinese has also been reconstructed with derivational prefixes and suffixes (Pulleyblank 2000, Schuessler 2007:38-50, etc.), but without evidence of a high degree of productivity.

Regarding this northern pre-Proto-Viet-Muong speech community, as an Austroasiatic language group, it most likely had a typical Austroasiatic typology and core Austroasiatic vocabulary, in addition to any innovations this branch had developed since the Austroasiatic dispersal. It would have had reduced presyllables and no tones, but final fricatives *-s and *-h and a glottal stop, and possibly phonemic phonation (e.g., Diffloth 1986). As for morphosyntax, Vietic had derivational prefixes and infixes, though no data exists to help ascertain the degree of productivity. Vietic noun phrase structure

has been hypothesized to have been like that of Old Khmer (i.e., head-initial structure with all modifiers following the head noun and no development of classifiers), SVO clause structure with a topic-comment tendency, and the lack of explicit passive-voice marking (see Alves 2020).

Overall, with exception of headedness in noun phrases, Vietic shared many typological features with Sinitic. This was likely before (or early in the process of) major sub-branching in Vietic, and thus long before the Viet-Muong stage. We can also assume that this group's grammatical vocabulary was entirely native at that point, including Austroasiatic retentions and Vietic innovations.

However, from the early centuries of language contact, dozens of grammatical words were borrowed from Sinitic into this portion of the Vietic lexicon. While Muong lects have some early Chinese loanwords (hereafter, "ECLs"), Vietnamese has the largest quantity, as will be shown in this study. As a result, today, Vietnamese grammatical vocabulary consists of a mixture of native Vietic and Austroasiatic etyma and Chinese loanwords. The latter include (a) ECLs from Late Old Chinese (i.e., towards the end of the Han Dynasty) and Early and pre-Late Middle Chinese (i.e., before the second millennium CE), all of which have native Vietic phonological features, and (b) the later Sino-Vietnamese layer, which is associated with Chinese character pronunciations of Late Middle Chinese of the early second millennium.

With an admittedly wide scope of the meaning of "grammatical vocabulary", the focus of this study is on function words, or at least vocabulary with a functional nature in semantico-syntactic structure and with relatively abstract, functional semantic properties. This study considers numerals and quantity terms, pronouns and interrogative terms, measure words, and locational and temporal terms, among others. Such vocabulary overlaps with the borrowing of matter and of pattern (see Sakel 2007). Thus, while grammatical vocabulary does not on its own demonstrate structural influence and typological changes, such vocabulary can at the very least enhance semantico-syntactic features, as well as possibly impact relevant semantic systems (e.g., restructuring of the system of Vietnamese terms of address and reference (see Alves 2017)).

Moreover, the borrowing of functional words can indicate the degree of intensity of language contact and bilingualism. It tends to correspond to the degree of structural impact, as posited by Thomason and Kaufman (1988:74-75). Morphophonological restructuring and some change in the noun phrase is what eventually occurred in the northern part of Vietic, which became Viet-Muong and from which Vietnamese developed. However, this was a process involving several centuries to a millennium, not all of which can be directly attributed to language contact with Chinese. Nevertheless, mitigating the perceived linguistic impact of Chinese, the lack of influence in certain subdomains highlights how this pre-Viet-Muong northern Vietic maintained a distinct linguistic status and identity. We will return to the question of ethnolinguistic implications in the conclusion.

In previous studies (Alves 2001, 2005, 2007a, 2007b, 2009), I have explored Chinese loanwords in the domain of grammatical vocabulary in Vietnamese, as well as possible structural influence of language contact with Chinese. However, these studies incorporated limited historical phonological and historical syntactic data for support. In this study, regarding the historical phonological aspects, the data herein have been evaluated with respect to historical phonological studies of Ferlus (1982, 1992, 2014, etc.), Nguyễn T. C. (1995), and Nguyễn V. T. (2005), consideration of Middle and Old Chinese of Baxter and Sagart (2014a) and Schuessler (2007), as well as my own observations about the chronological development of tones in both Sinitic and Vietnamese (Alves 2018a). I have also increasingly applied ancient Chinese and Vietnamese textual data together with historical phonological matters (see Alves 2018b regarding the triplet *cũng* 'also', *cùng* 'together with', and *cộng* (a bound morph with the general sense of 'together'), all related to Chinese 供 *gòng* 'total' (OC *N-k(r)oŋʔ-s, MC *gjoŋŋH*) and Alves 2020 regarding Vietic noun phrase structure). Lastly, we now have a substantive study of the Austroasiatic grammatical lexicon, with Proto-Austroasiatic reconstructions (Shorto 2006, used for all Proto-Austroasiatic reconstructions except those marked #, which are my suggested reconstructions) and comparative data (Alves, Jenny, and Sidwell 2020), which serve as a further useful point of comparison.

I cannot demonstrate the same depth as the latter articles for all the words in this study, but I have made efforts to (a) note historical phonological patterns for all selected items and (b) check ancient Chinese textual resources. I have periodically checked two Chinese language dictionaries of Ancient

Chinese words (Guhanyu Cidian: Dongfang xuesheng gongjushu (1997), Guhanyu Changyongzi Zidian (2005)) and Vrong Loc’s (2002) dictionary of ancient words in Vietnamese to determine that certain meanings and syntactic structures are attested in the past. I have also checked ancient Chinese textual data at the ctext.org database, focusing on the early period in question, though finding examples with precise meanings and syntactic distribution is not always possible. For native Vietnamese words, there is no textual data outside of Vietnamese-language Nôm documents extending back to the 1200s. However, the types of native vocabulary tend to be less controversial and have higher certainty (e.g., numerals and basic time words) and often supported by comparative data in Austroasiatic (e.g., pronouns and locative words), further strengthening the certainty of chronological depth. Even stronger arguments can be made with more careful identification of textual evidence, but that is a long-term matter for after this study.

2 Grammatical vocabulary in Vietnamese

This study of grammatical vocabulary presents over sixty native etyma and about sixty ECLs. Thus, these words have a history extending from the second millennium BCE before the Han expansion into northern Vietnam into the first several centuries of Sinitic-Vietic contact in the first millennium CE.

Table 1 provides very general time periods based on historical and archaeologically attested periods. This lasts from the Austroasiatic dispersal hypothesized to be about 4000 BP into the mid-first millennium CE. Regarding contact with Sinitic, that period lasts from Late Old Chinese through Middle Chinese before the Late Middle Chinese stage. The key Chinese dynasties when historically documented early migrations of Chinese arrived in northern Vietnam include the Han Dynasty (206 BCE-220 CE), the Jin Dynasty (266-420 CE), and undoubtedly more arrived in following centuries, but as yet, I have seen no other supporting textual evidence with details. It is also presumably leading into the period of development of the hypothetical Annamese Chinese, as per Phan (2013). The descendants of speakers of this variety of Chinese ultimately shifted to Viet-Muong in the early second millennium. Chinese function words borrowed from the turn of the second millennium CE and the emergence of Viet-Muong (i.e., dictionary Chinese character readings) are not part of this study.

Regarding the Annamese Chinese hypothesis, whether a specific Chinese dialect emerged in this area—parallel to the emergence of Chinese dialect groups elsewhere in southern China—cannot be known. Nevertheless, the grammatical lexical data in this study could not have been borrowed without a large enough Chinese-speaking community embedded within this pre-Proto-Viet-Muong community.

Table 1: Stages from Austroasiatic to Vietic

Stages	Time Periods
1. Austroasiatic dispersal	c. 2000 BCE
2. Development of Vietic as a distinct group and early contact with Kra-dai (pre-Proto-Tai?) and early but limited contact with Sinitic	Between 2000 and 1 BCE
3. Initial substantial Sinitic-Vietic contact and development of local Annamese Chinese (Late Old Chinese to Middle Chinese)	First several centuries of first mill. CE
4. Development of Viet-Muong and shift of Annamese Chinese to Viet-Muong (Late Middle Chinese)	Early 2 nd mill. CE

Analysis of subcategories of ECLs allows for historical linguistic inferences from the pre-Proto-Viet-Muong period. Native proto-language etyma include native numerals (‘one’ to ‘ten’), core pronouns (first-, second-, and third-person singular, second-person plural, and determiners), question words (‘who’, ‘what’, ‘where’), several locational terms, and core time words. Thus, the grammatical terms which are reconstructable to an early stage of Vietic are genuinely basic vocabulary (e.g., pronouns and numbers) or relatively more basic (e.g., interrogative words and location words). Some of these have central roles in clause structure.

ECLs in Vietnamese, in contrast, are less central to clause structure but instead have other semantic-syntactic functions (e.g., modal verbs, conjunctions, unit terms) or even have a cultural nature (e.g., calendar terms). ECLs fall into several categories, including a handful of locative, measure, and numeric terms, but a notable number of clause-connective and preverbal auxiliary words, as well as

time words related to the Chinese zodiac calendar. Grammatical ECLs in Vietnamese are mostly not seen in Vietic languages other than Muong, but they nevertheless represent borrowing in a relatively early stage of Vietic history, at least the predecessor to Viet-Muong. Vietnamese and other Vietic languages have numerous additional Chinese loanwords from the Late Middle Chinese period, but these were borrowed after the emergence of Viet-Muong and a stage in which Vietnamese was distinct. Thus, those true Sino-Vietnamese loanwords are not relevant to the sociocultural impact in first part of the first millennium CE.

The matter of identification of ECLs requires explanation. Sino-Vietnamese vocabulary stemming to the Late Middle Chinese period can be readily identified since such words are listed in dictionaries as Chinese character readings. In contrast, ECLs must be identified through the comparative method, and there generally cannot be absolute certainty regarding such proposed loanwords. The ECLs in this study have been assessed for their phonological, semantico-syntactic, and sociocultural viability, and based on these factors, they have been evaluated with medium to high certainty of ECL status. Many other candidates have been excluded, and some items in this study may later be excluded as data is further evaluated. But in general, considering the consistency of the phonological patterns and quantity of words in semantic domains which provide reinforcing support for certainty, a large majority of these are strong candidates as ECLs. Thus, though some details herein remain tentative, the broader claims (e.g., sufficient Sinitic-Vietic bilingualism to facilitate substantial borrowing of grammatical words in this early period) must be considered viable hypotheses unless strong counterevidence shows otherwise.

Additional clarification of these words' histories is provided through Muong lexical data found in the substantive Muong Bi dictionary of Nguyễn V. K. et al. (2001). Comparable Muong words are provided in various tables in this paper to show the presence of probable ECLs in Viet-Muong, not only Vietnamese, thus supporting the dating of these items to that early period, but data from more Muong varieties is needed. While Vietnamese cannot be ruled out as a donor language in some of the Muong words, others have phonological features that mark them as predating the modern era. Phan (2013:320-321) has also presented some data of Sino-Vietnamese grammatical morphemes shared by a few varieties of Muong. He similarly noted phonological features of these words to suggest that these were indeed borrowed directly from Annamese Chinese, rather than through Vietnamese. He further proposes that these predate the speciation of Viet-Muong. While this is one reasonable hypothesis, another possibility is that they were borrowed variously before and after speciation. Regardless, at least some are genuine ECLs, which supports the presence of a Chinese-language community and widespread bilingualism with Viet-Muong in the first millennium.

However, as will be shown, many grammatical ECLs in Vietnamese do not appear in the Muong Bi data, suggesting a different type and degree of language contact with Chinese. Also, though Nguyễn V. T.'s (2005) 1,200-word list for some 30 Muong varieties does contain several grammatical ECLs listed in this study, in various instances, not all varieties of Muong are shown to have those ECLs. For example, while all the Muong varieties have cognates for the Vietnamese ECL *khá* 'rather/very' (see Table 20), Muong cognates for Vietnamese ECL *cũng* 'also' (see Table 19) are in only 10 of the 30 Muong lects, with another form /i/ in the other 20 (Nguyễn V. T. 2005: 230 and 203 respectively). Thus, we can hypothesize that different groups of Viet-Muong had different periods of contact with Chinese in different locations, perhaps during the time in which Vietnamese and varieties of Muong were differentiated. Nevertheless, the idea that Viet-Muong, as opposed to other Vietic groups, shared a more intense degree of contact with Chinese—thereby spurring the speciation of Viet-Muong as a Vietic sub-branch—is supported by the larger quantity of Chinese loanwords into Vietnamese and varieties of Muong than other sub-branches of Vietic.

Finally, a more thorough study of these proposed ancient loanwords would require careful identification of relevant developed senses, functions, and distributional patterns of grammatical words in ancient Chinese texts. However, in most cases, the semantics are straightforward and either have been retained in Chinese into the modern era or, in a few cases, are older senses that can be readily located in references on ancient literary Chinese. For example, Vietnamese *khá* 'rather/very' has a *sắc* rather than *hỏi* tone marking it as an ECL, and the Chinese source 可 *kě* 'able' has, among other functions, precisely this preverbal intensifier position and semantic function and did so also at least as early as the Tang Dynasty (618-907 CE). In some instances, some ECLs were not originally

grammaticalized throughout Sinitic but became grammaticalized in Vietnamese, and sometimes Muong. Such cases are noted and of course cannot be considered direct influence of Chinese, but they are, nonetheless, ECLs with grammatical meanings. They could also potentially represent developments of the Annamese Chinese speech community in the region in the first millennium CE, in which case, some of these could be borrowed grammatical words, but then the timing and source of the grammaticalization would become harder or even impossible to determine.

3 Categories of function words Vietic and Early Chinese loanwords

This section covers Vietnamese function words and their etymological origins. Table 2 presents an overview of the eight categories of grammatical words and the numbers of words in the two categories of Vietic etyma and ECLs. As is shown, pronominal, quantity, and locational words are predominantly native etyma. ECLs are dominant in the categories of measure words and words with conjunctive, comparative, and modal functions. Thus, these are words with potential significance to or even impact on phrase structure in northern Vietic. The number of time words is similar in both columns, but native words are core time words (e.g., ‘day’, ‘month’, ‘year’, etc.). ECLs in this domain are secondary (e.g., ‘hour’, ‘now’, ‘turns’, ‘instances’, etc.) or are related to the sociocultural system of the Chinese zodiac calendar. The total number of grammatical ECLs in Table 2 is 62. The number is 73 when Chinese zodiac calendar terms, which are cultural content words rather than function words, though with semantic overlap of these with temporal function words, so they are noted herein.

Table 2: Categories and numbers of Vietic and ECL items in Vietnamese

Category	Vietic Etyma	Early Chinese Loans
1. Pronouns and interrogative terms	11	0
2. Numbers and quantity expressions	18	9 (mostly quantity terms)
3. Locational words	12	5
4. Time words	11	12 (and 11 calendar terms)
5. Measure words and units	8	17
6. Conjunctive terms	1	5
7. Comparative terms	1	8
8. Modals and preverbs	1	6
Total	63	62 (or 73)

A few ECLs have spread widely throughout Vietic, enough to be reconstructable, as shown in Table 3. The timing of the borrowing is obviously not to the proto-Vietic level but rather sometime in the first millennium CE. Vietic languages have borrowed grammatical words from Vietnamese recently, as shown by their phonological traits (e.g., the Sino-Vietnamese passive marker *bị* in Chut languages), and thus cannot be valid reconstructions. However, the phonological features of the words in other Vietic languages listed in Table 3 mark them as earlier loans, though whether directly from Chinese or another Vietic language cannot be known. Regardless, this small number of items highlights the spread of ECLs in Vietic many centuries ago but also the limited impact on function words outside of Viet-Muong.

Table 3: Grammatical ECLs which are reconstructable in Vietic

Gloss	Viet.	SV	Proto-Vietic	Muong	Chinese	OC	MC
(classifier)	<i>cái</i>	<i>cá</i>	#kajʔ	<i>cải</i>	個 <i>gè</i> ‘piece, item’	*kʰa[r]-s	kaH
many/much	<i>nhiều</i>	<i>nhiêu</i>	#new	(<i>từ</i>)	饒 <i>ráo</i> ‘abundant’	*[ŋ]ew	nyew
pair	<i>đôi</i>	<i>đôi</i>	*to:j	<i>tôi</i>	對 <i>duì</i> ‘pair’	*[tʰ][u]p-s	twojH
side	<i>bên</i>	<i>biên</i>	#pe:n	<i>pên</i>	邊 <i>biān</i> ‘side’	*pʰe[n]	pen

Another aspect considered in this study is borrowability rates. WOLD (the World Loanword Database) lists borrowability rates of classes of words, providing criteria for a select list of basic vocabulary, the Leipzig-Jakarta 100-word list (Tadmor et al. 2010), which is relatively resistant to borrowing. For this study, the statistical range of borrowability rates are interpreted in a relative manner: one can posit very low (e.g., below 0.10, typical for items in the Leipzig-Jakarta list), somewhat low (from 0.10 to 0.25),

and medium or higher rates. These are impressionistic ranges, but they are based on ranges of rates found in WOLD statistics. Where necessary in subsequent subsections, I provide interpretation of the borrowability rates seen in subgroups of words.

This study secondarily explores the matter of relatively more basic vocabulary within the domain of grammatical vocabulary, as demonstrated precisely in the kinds of native retentions in Vietnamese versus ECLs. In this way, the data from WOLD gives some perspective regarding the etymological sources of the words, and while WOLD statistics cannot be considered strong support for claims of loanword status, the statistical tendencies in WOLD provide another angle to consider and should be used as a point of reference in establishing word origins. The following subsections of types of grammatical words are organized as they are in Table 2.

3.1 Pronouns and Interrogative Terms

Vietnamese pronouns and interrogative terms highlight the native Vietic and Austroasiatic origins of the Vietnamese lexicon. In the data of WOLD, among the forty-one languages, these terms have very low borrowability rates, 0.07 or lower, as in Table 4. Such rates match rates of other basic vocabulary in the Leipzig-Jakarta wordlist, such as ‘you (2s)’ at 0.04, ‘this’ at 0.00, and ‘who’ at 0.03.

Table 4: Borrowability rates of pronouns and referential terms

Borrowability Rates	Categories
0.00 - 0.07	Personal pronouns
0.00 - 0.06	Interrogative, information-question words
0.00 - 0.03	Demonstratives/Deixis

The modern Vietnamese system of pronouns has been somewhat restructured from the original Vietic system (which is admittedly only partially reconstructable). There is notable impact on the overall system and related pragmatics, resulting directly and indirectly from language contact with Chinese (see Alves 2017). Despite this impact stretching out over several centuries from the first to second millennium CE, all pronouns and interrogative terms in this data are almost exclusively native retentions. The native pronouns in Table 5 include first, second, and third-person singular pronouns and the proximal demonstrative.

Table 5: Proto-Vietic pronouns

Gloss	Proto-Vietic	Proto-Austroasiatic	Vietnamese
1s	*so:	NA	<i>tao</i>
2s	*mi:	*mi[i]ʔ; *miih	<i>mày</i>
2p	#baj	*pej	<i>bay</i>
3s	*hanʔ	*[ʔ]anʔ	<i>hấn</i>
3s	#na:ʔ	*nVʔ	<i>nó</i>
this	*-ni:	*niʔ; *nih	<i>nay, này</i>

The data in Table 5 deserves comments. While the second and third-person Proto-Vietic forms are Austroasiatic etyma (again, from Shorto 2006 in this table and all others), the first-person form is an apparent Proto-Vietic innovation of unknown origin. Not listed in Table 5 is the only pronominal ECL, Vietnamese *họ*, which is a third-person plural pronoun connected to Chinese 戶 *hù* ‘household’ (OC *m-qʰaʔ, MC huX). However, this grammaticalization represents an innovation in Vietnamese, so this ECL might not have been borrowed as a pronoun. Only further exploration in Nôm texts can determine whether any evidence can elucidate the chronology of grammaticalization of this word. It is not used this way in Muong Bi, which instead uses *pâu* ‘they’ (nor is this item seen in descriptions of other Vietic languages), but data for this word in other Muong lects is not readily available.

The form for the proximal ‘this’, Vietnamese *này*, is comparable to a form with complex regional distribution. This form is seen in Austroasiatic (*niʔ; *nih ‘this’), Austronesian (*i-ni ‘this/here’), Kra-dai (*naj^C), and Cantonese *nei^l/ni^l* 呢 ‘this’ (but not widespread among other varieties of Chinese

or even Yue). Regardless of any possible deeper historical origins, it must be treated first as an Austroasiatic etymon and a lexical retention, not a loanword. Other determiners in Vietnamese constitute an phonaestheme-like analogical pattern of initials and rhymes (see Thompson 1985:142), making etymological identification challenging. For example, words with initial /d/ (*đó* ‘that’, *đấy* ‘there’, *đâu* ‘where’) may be related, but there is currently no means of determining their etymological origins or derivational sources.

All Vietnamese interrogative words in Table 6 are native etyma. However, the terms *mô* ‘where’ and *chi* ‘what’ are central Vietnamese dialect words that have different counterparts of unknown etymological origin in northern and southern dialects (*đâu* ‘where’, as just noted, and *gi* ‘what’). Central Vietnamese has retained some archaic traits in its grammatical vocabulary, including some of those words’ phonological features, while northern Vietnamese innovated in other ways (Alves 2012). Some Vietnamese interrogative terms are bisyllabic compounds, which cannot be reconstructed to the distant past. The compounds sometimes contain a mixture of etymological sources (e.g., ‘why’ *tại sao* (at.how) with a Sino-Vietnamese syllable (在 *zài* ‘at’) and a native etymon; ‘when’ *bao giờ* (how much.time) with a native etymon *bao* and an ECL *giờ* (see in Table 14)).

Table 6: Proto-Vietic interrogative words

Gloss	Proto-Vietic	Proto-Austroasiatic	Vietnamese
how many	*bəl?	NA	<i>máy</i>
what	#ci:	*(?)ci?	<i>chi</i> (dialectal)
where	#mo:	*m(o)?	<i>mô</i> (dialectal) ‘where/what’
where/which	*-mo:	NA	<i>mô</i> (dialectal)
who	*?e:	NA	<i>ai</i>

3.2 Numbers and quantity expressions

Compared to the low borrowability rates of pronouns, rates among number words are noticeably higher, with numerals two to ten having rates of 0.21 to 0.29. Numeral term systems are sometimes shared among languages, and in this region, language contact with Chinese has significantly affected the numeral systems of Tai (and Japanese and Korean). Nevertheless, numerals must be considered in studies of affiliation in language families. In WOLD, other quantity expressions vary in borrowability rates, but in general, they range from 0.09 to 0.23, only somewhat higher than truly basic vocabulary.

Like pronouns and interrogative terms in Vietnamese, core numeral words are native etyma, either stemming to Vietic or Austroasiatic. The overall situation of numerals in the Austroasiatic language family suggests developments of numeral terms above four after the Austroasiatic dispersal (Sidwell nd), but still, all Vietic numerals have cognates in multiple branches of Austroasiatic, as noted for the terms for five to nine in Table 7. While Chinese was clearly a donor language for many cultural elements, including measure words, as described in Section 2.5, the retention of native numerals over a thousand years of language contact and a bilingual Sinitic-Vietic community is evidence of Vietic’s sociocultural status in the first millennium CE. Otherwise, we could expect to see a situation like that of Tai, in which core numerals have been replaced with the Chinese system. These include the Proto-Tai numbers for two to ten, and then through lexical compounding, to ninety-nine. In addition, in a list of lexical data for sixteen Kradai lects (Liang and Zhang 1996:44-45), most have forms clearly stemming to Chinese numeral terms for 100, 1,000, and 10,000 (with Thai as a notable exception, having other etymological sources for 100 and 1,000). This is very different from the circumstance in Vietnamese and Vietic broadly.

An unanswerable question is how much—if at all—the Annamese Chinese community adopted Vietic numerals in their daily speech in a bilingual community. Regardless, other than some very limited circumstances noted below, Vietic speakers did not predominantly use Chinese numerals.

Table 7: Proto-Vietic numbers

Gloss	Proto-Vietic	Proto-Austroasiatic	Vietnamese
one	*mo:c	*muuj	<i>một</i>
two	*ha:r	*baar (unlikely related to the Vietic form)	<i>hai</i> (cf. <i>vài</i> ‘a few’)
three	*pa:	*pi? (#pe:)	<i>ba</i>
four	*po:n?	*pun?	<i>bốn</i>
five	*dām	NA (cf. Bahnaric, Monic, Aslian)	<i>năm</i>
six	*p-ru:? > p ^h ru:? / k ^h lu:?	NA (cf. Bahnaric, Monic, Aslian)	<i>sáu</i>
seven	*pəs	NA (cf. Bahnaric, Monic, Aslian)	<i>bảy</i>
eight	*sa:m?	NA (cf. Bahnaric, Monic, Aslian)	<i>tám</i>
nine	*ci:n?	*dciin[?]	<i>chín</i>
ten	*ju:k	NA	<i>chục</i>
ten	*ma:l	NA	<i>mười</i>
hundred	*k-lam	NA	<i>trăm</i>
few	*ʔi:t	NA	<i>ít</i>
more	#də:h	NA	<i>nữa</i>
half	#CVdə:h	NA	<i>nửa</i>

There are only a few Vietic general quantity terms in Table 8, but these express significant quantitative senses (‘few’ and ‘more’). None are Austroasiatic etyma, but the fact that these are all reconstructable in Vietic suggests they could have been used before Sinitic-speaking groups arrived.

Table 8: Proto-Vietic quantity words

Gloss	Proto-Vietic	Proto-Austroasiatic	Vietnamese
few	*ʔi:t	NA	<i>ít</i>
more/additional	#də:h	NA	<i>nữa</i>
half	#CVdə:h	NA	<i>nửa</i>

As for ECL quantity terms, there are several quantity expressions and a few numerals with very restricted usage. The impact of Chinese quantity expressions on pre-Viet-Muong and Vietic broadly is small but still notable, including ‘every’ and ‘much/many’. The Chinese words ‘pair/twins’ and ‘many/much’ have spread widely throughout Vietic, as noted in Table 2. The numbers that have been borrowed have little to no usage in modern Vietnamese, and it is unclear exactly how these scattered items were used in the pre-Proto-Viet-Muong period. The number 10,000 was not likely to have been useful in daily interaction, except with a metaphorical meaning, as in the sense ‘crowded/numerous’ in Table 9 (*muôn* has phonological features showing it to be an older loanword than *vàn*). Also, the ECLs for ‘two’ (with different tones suggesting multiple periods of borrowing) and ‘four’ have no widespread usages in modern Vietnamese, and there are no other apparent ECL numerals.

The entire set of Chinese numerals is available in the later Sino-Vietnamese pronunciations, but these are used primarily in literary writing or in limited compounds, not as free morphemes. Again, as noted above, Vietnamese has retained all core native number terms, so while the system of quantity terms was modified somewhat in the first millennium, the Vietic numeral system has received no apparent influence.

Table 9: Early Chinese loanwords for quantity terms and numbers

Category	Gloss	ECL	SV	Muong	Chinese	OC	MC
general	enough	<i>đủ</i>	<i>túc; tú</i>	<i>tú</i>	足 <i>zú</i>	*[ts]ok-s	tsjowk
	every	<i>mọi</i>	<i>mỗi</i>	(<i>tháy</i>)	每 <i>měi</i>	*m ^ə ʔ	mwojX
	much, many	<i>nhiều</i>	<i>nhiều</i>	(<i>từ</i>)	饒 <i>ráo</i>	*[ŋ]ew	nyew
	lacking/little	<i>thiếu</i>	<i>thiếu</i>	<i>thiếu</i>	少 <i>shǎo</i>	*[s.t]ewʔ	syewX
	pair	<i>đôi</i>	<i>đôi</i>	<i>tôi</i>	對 <i>duì</i>	*[t]ʰ[u]p-s	twojH
number	two	<i>nhi, nhì, nhĩ</i>	<i>nhị</i>	NA	二 <i>èr</i>	*ni[j]-s	nyijH
	four	<i>tư</i>	<i>tứ</i>	NA	四 <i>sì</i>	*s.li[j]-s	sijH
	ten thousand	<i>muôn,</i>	<i>vạn</i>	<i>muôn</i>	萬 <i>wàn</i>	*C.ma[n]-s	mjonH
	crowded, numerous	<i>vàn</i>	<i>vạn</i>	<i>muôn</i>	萬 <i>wàn</i>	*C.ma[n]-s	mjonH

3.3 Locational words

The borrowability rates of the two dozen locative words in WOLD can be divided into two main categories: those with rates below 0.10 and those with rates above 0.20. Rates of basic locational and directional terms with core cognitive linguistic properties are at the lower end (e.g., ‘up’ 0.0, ‘above’ 0.03, ‘down’ 0.05, ‘far’ is 0.06, etc.). In contrast, cardinal direction terms have notably higher borrowability rates of 0.20 or higher (i.e., ‘north’ 0.20, ‘south’ 0.26, ‘east’ 0.23, ‘west’ 0.24). This situation largely matches that of Vietnamese locational terms, with some exceptions. Most locational and directional terms in Vietnamese are native etyma, while cardinal terms are all Chinese loanwords from the later Sino-Vietnamese layer, not ECLs. However, a few locational terms were borrowed from Chinese in the early, pre-Sino-Vietnamese period.

Table 10: Proto-Vietic location words

Gloss	Proto-Vietic	Austroasiatic	Vietnamese	Muong
after	#k ^h saw	*krawʔ	<i>sau</i>	<i>khau</i>
before/forehead	*k-la:k	NA	<i>trước</i>	NA
end/extremity > nipple	*gɔ:jʔ / kɔ:jʔ	NA	<i>cuối</i>	<i>cuối</i>
far	*s-ŋa:jʔ	*[c]ŋajʔ	<i>ngái</i> (dialectal)	(<i>xa</i>)
inside	*k-lɔ:ŋ	*kluuŋ ‘belly/middle’	<i>trong</i>	<i>tlong</i>
left side	*k-la:jʔ	NA	<i>trái</i>	<i>tlái</i>
middle	#Csah	NA	<i>giữa</i>	<i>khũa</i>
right side	*dam / tam	*st ₁ am; *st ₁ uum	<i>đăm</i> (archaic)	<i>tăm</i>
stay/be at	*ʔəh	NA	<i>ở</i>	<i>ớ</i>
top (upstream?)	*k-le:ŋ	NA	<i>trên</i>	<i>tliênh</i>
top/crown	*ŋɔ:nʔ	NA	<i>ngon</i>	<i>ngon</i>
under	*-ta:lʔ	*kt ₁ aal	<i>dưới</i>	<i>tín</i>

Several of the Vietic reconstructions in Table 10 are Austroasiatic etyma (e.g., ‘after’, ‘under’, ‘inside’, ‘right side’, and ‘far’), giving these items truly deep history. Both ‘far’ and ‘right side’ are not used in modern mainstream Vietnamese, but they still demonstrate native etymological connections. As for ‘inside’, it is a possible grammaticalized form of ‘belly’, a noted grammaticalization cline (i.e., BELLY > IN (Heine and Kuteva 2002:53)), but it may not be possible to determine the timing of grammaticalization in Vietic. There is a comparable form in Chinese 中 *zhōng* ‘middle’ (OC *trujŋ, MC trjuwŋ) considering its phonetic shape, including a source *u which often leads to Vietnamese ‘o’ /ɔ/, and reasonable semantic overlap. However, I propose we consider native sources first as a default unless stronger evidence can show an external source is more likely.

The number of locative ECLs in Vietnamese is smaller than the number of native terms, but these ECLs have significant semantico-syntactic functions. A general locative term for ‘side’ (in contrast with

‘left’ or ‘right’, which are native terms in Vietnamese) was borrowed. There is an odd pattern, for which I see no explanation: the borrowing of only one item of semantic pairs, including ‘near’ but not ‘far’, ‘from’ but not ‘to’, ‘outside’ but not ‘inside’, and ‘bottom’ but not ‘top’. Overall, the locative ECLs unquestionably impacted parts of this lexical system in that early period, but a core native system is still mostly in place.

Table 11: Early Chinese loanwords for locational terms

Gloss	ECL	SV	Muong	Proto-Vietic	Chinese	OC	MC
bottom	<i>đáy</i>	<i>đẽ</i>	<i>đáy</i>	NA	底 <i>dǐ</i>	*tʰijʔ	tejX
from	<i>từ</i>	<i>tư</i>	(<i>pó</i>)	NA	自 <i>zì</i>	*s.[b]i[t]-s	dzijH
near	<i>gần</i>	<i>cận</i>	<i>khênh</i>	*t-kəŋ	近 <i>jìn</i>	*N-kərʔ	gj+nH
northern (of wind)	<i>bắc</i>	<i>bắ</i>	NA	NA	北 <i>běi</i>	pok	*pʰək
outside	<i>ngoài</i>	<i>ngoại</i>	<i>wài</i>	NA	外 <i>wài</i>	*[ŋ] ^{ws} a[t]-s	ngwajH
side	<i>bên</i>	<i>biên</i>	<i>pên</i>	#pe:n	邊 <i>biān</i>	*pʰe[n]	pen

As for possible syntactic influence, it seems such ECLs did not impact Vietic prepositional phrase structure. While the use of prepositions, not postpositions, in Vietic has not been shown conclusively, all evidence suggests that Vietic had right-branching structures in noun phrase and clauses. Because both Vietic and Sinitic were prepositional languages, loanwords for ‘from’ and ‘near’ could have been readily fit into Vietic structures. However, for Sinitic locational nouns, such as ‘outside’ and ‘side/the location of’, the words take preceding modifiers, such as adjectives, possessives, and determiners (e.g., Chinese *wū wài* 屋外 (house-**outside**) ‘outside the house’ (found in both early and modern Chinese), in which ‘house’ precedes the locative noun). This is unlike Vietnamese and other Vietic languages with post-noun modifiers (e.g., Vietnamese *ngoài nhà* (**outside**-house) ‘outside the house’, in which ‘house’ follows). Thus, not only did Sinitic not impact Vietic phrase structure related to locational concepts; some of these locative words were also adapted to the typologically distinct Vietic syntactic patterns.

3.4 Time words and calendar terms

Table 12 presents three levels of borrowability rates of time words, focusing on the terms in this study: low (0.01-0.19), medium (0.22-0.45), and high (0.49-0.76). The lowest level consists of terms that are identifiable in the natural world (e.g., ‘day’ and ‘year’), while those at the upper end of the range are arbitrarily defined and culturally specific (e.g., ‘hour’ and ‘week’). As might be expected, those in the lower range are native in Vietnamese, as seen in Table 13, while those at the upper end are ECLs, as in Table 14. Those in the middle category include both native etyma and ECLs.

Table 12: WOLD borrowability rates of time words

Rates	Subcategories
0.01 to 0.19	<ul style="list-style-type: none"> • Terms related to ‘days’ (day, today, tomorrow, etc.) • Parts of the day (morning, midday, evening, night) • Month, year • Seasons (spring, summer, autumn, winter)
0.22 to 0.45	<ul style="list-style-type: none"> • Various (the beginning, immediately, always, ready, the season)
0.49 to 0.76	<ul style="list-style-type: none"> • Hour, week, days of the week

A few of the Vietic time words in Table 13 are also Austroasiatic etyma. The words ‘day’ and ‘year’ are truly basic as they are observable natural phenomena. In contrast, ‘about to’ is a probable grammaticalized word derived from ‘ready/prepare’, and while the word form is reconstructable in Vietic, it is not possible to determine when the grammaticalization occurred. However, the etymon with the aspectual function is seen in both Muong and the conservative Vietic Ruc language *k^hrap*³ (Nguyễn V. L. 1993:100), both with retained early onset, suggests some time depth of this development). All other time words in Table 13 have Proto-Vietic reconstructions, including the parts of the day and two

additional aspectual terms. The latter two, like ‘about to’, are likely grammaticalized from verbs, though ideally, there will be textual data in early Vietnamese Nôm writing to clarify the timing of such developments.

Table 13: Proto-Vietic time words

Category	Gloss	Vietnamese	Proto-Vietic	Proto-Austroasiatic
Units	day	<i>ngày</i>	*-ŋi:	*tŋiiʔ
	month	<i>tháng</i>	*k-ra:ŋʔ	NA
	year	<i>năm</i>	*c-n-əm < cəm	*cn ₁ am
	day after tomorrow	<i>mốt</i>	#CVmu:t	NA
Parts of days	morning	<i>sớm</i>	*k-rə:mʔ	NA
	midday/noon	<i>trưa</i>	*k-la:	NA
	dark/evening	<i>tối</i>	*su:lʔ	NA
	night	<i>đêm</i>	*te:m	NA
Adverbials	right/immediate/just	<i>ngay</i>	*t-ŋar	NA
	finished/complete	<i>xong</i>	#cɔ:ŋ	NA
	about to/prepared to	<i>sắp</i>	#srap < k ^h rap (?)	*srap ‘ready/prepare’

The ECLs in Table 14 include a range of adverbials and units of time which do not have a natural world point of reference, as the native etyma do. Instead, these adverbials primarily function as aspectual/temporal markers for past, present, or future time frames. Other than ‘hour’, the unit terms have a somewhat more abstract sense of an instance of time. The grammaticalization paths of all of these words will require more careful sifting of both early Chinese and Vietnamese textual data. Nevertheless, the phonological features of these mark them as loanwords in that early period.

Of note is the Chinese word 時 *shí*, which is listed twice in Table 14. Thus, the loanwords constitute a triplet with two ECLs and a corresponding Sino-Vietnamese form. In Chinese, this morph has a wide range of meanings, including ‘time (in the general sense)’, ‘currently’, ‘hour’, among others. Correspondingly, this triplet in Vietnamese expresses multiple meanings: (a) *giờ* ‘time (in the general sense)’, ‘now’, and ‘hour’, (b) *chờ* ‘now’, and (c) *thì/thời* ‘time (in the general sense)’. The Sino-Vietnamese syllable *thì* is a later Sino-Vietnamese borrowing, while *giờ* is probably the oldest, considering the affricate onset. The timing of the borrowing of the ECLs is less certain, but their onsets and vowels show they are both ECLs. The Sino-Vietnamese morph *thì* is also a possible source of the homophonous high-frequency topic marker *thì* ‘then’, becoming more dominant over the ECL *bèn* ‘then’, as discussed in Section 3.6.

Also of note is Vietnamese *đang* ‘during’, of which some aspects weaken my claim of its ECL status. The phonological form is reasonable, including the retention of *a from the early period of borrowing, in contrast with the diphthongization in the Sino-Vietnamese form. However, the use of this word in modern Chinese is as a kind of adverbial clause marker, meaning ‘when’, not as a preverbal form in Vietnamese (and I cannot find syntactic descriptions of this in early Chinese texts). Meinsternerst (2011) shows how, in Pre-Tang Chinese, 當 *dāng* had multiple functions, including as a temporal and locative preposition, effectively “at” (in addition to a verb meaning ‘to match/correspond’ and modal expressing necessity). Thus, if this is indeed an ECL, there was a reinterpretation with impact on its semantico-syntactic features. Another issue is that some Tai languages have a comparable form. The progressive sense for a form [daaŋ] is seen in several central and northern Tai languages (Gedney 2008:115), though this is homophonous with ‘body’, a common source in grammatical clines. There is a Malay/Indonesian form *sedang*, which has a similar semantico-syntactic function and distribution, but considering the presyllable, this seems more likely to be a chance partial similarity. As for deeper historical origins, one word in isolation contributes nothing of value, and without wider distribution in Austronesian (perhaps instead grammaticalization from the homophonous word meaning ‘average’?), it does not even have relevance to Sagart’s Sino-Austroasian hypothesis. Regardless, the word would

be a likely ECL into both Viet-Muong and Tai. Hopefully, additional ancient Chinese and/or Vietnamese textual data can clarify the matter.

Table 14: Early Chinese loanwords for time terms

Category	Gloss	ECL	SV	Muong	Chinese	OC	MC
Adverbs	always	<i>xàng</i> (redup.)	<i>thường</i>	NA	常 <i>cháng</i>	*[d]aŋ	dzyang
	during	<i>đang</i>	<i>đương</i>	<i>tang</i>	當 <i>dāng</i>	*t'aŋ	tang
	ever	<i>từng</i>	<i>tằng, tãng</i>	NA	曾 <i>céng</i>	*[dz]ʰəŋ	dzong
	just about to	<i>chực</i>	<i>trực</i>	(<i>khấp</i>)	直 <i>zhí</i>	*N-t<r>ək	drik
	now	<i>chừ</i>	<i>thì</i>	(<i>cạ nì, chí nì</i>)	時 <i>shí</i>	*[d]ə	dzyi
	previous/old	<i>xưa</i>	<i>sơ</i>	<i>hơ</i>	初 <i>chū</i>	*[ts]ʰra	tsrhjo
	then	<i>bèn</i>	<i>tiện</i>	(<i>đẹ ẹ</i>)	便 <i>biàn</i>	*ben-s	bjienH
Units	a turn	<i>lần</i>	<i>luân</i>	<i>lần</i>	輪 <i>lún</i>	*[r]u[n]	lwin
	batch; time/turn	<i>lượt</i>	<i>liệt</i>	<i>lướt</i>	列 <i>liè</i>	*[r]e]t	ljet
	hour (and 'now' and 'time')	<i>giờ</i>	<i>thì</i>	<i>dờ</i>	時 <i>shí</i>	*[d]ə	dzyi
	period of time; level	<i>chừng</i>	<i>trình</i>	NA	程 <i>chéng</i>	*l<r>eŋ	drjeng
	time/instance	<i>phen</i>	<i>phiên, phan</i>	(<i>cạ</i>)	番 <i>fān</i>	*pʰar	phjon

There are other ECLs related to time, but not fully grammaticalized ones. Vương (2002) notes a number of Chinese loanwords (what he sometimes labels “nativized” words, likely following Wang Li (1948, 1958), though no phonological traits distinguish them from ECLs), such as *thốt* ‘measure word’, *vàn* ‘numerous’, *bướp* ‘lack’, *tua* ‘must’. A few of these in Table 15 are ECLs which refer to punctuality. Overall, among ECLs, there is a domain of terms with reference to both grammatical issues of time (e.g., aspect) and timeliness (i.e., being late or on time), altogether supporting the idea of a bilingual language contact situation, one with a broader sociocultural impact.

Table 15: ECLs with general reference to timeliness

Gloss	ECL	SV	Muong	Chinese	OC	MC
gradual	<i>dần</i>	<i>tuần</i>	<i>đần</i>	馴 <i>xùn</i>	*zwin	*sə.lu[n]
in time	<i>lập</i>	<i>cập</i>	<i>lập</i>	及 <i>jí</i>	*[m-k]rəp	gip
in time	<i>kíp</i>	<i>cập</i>	(<i>lập</i>)	及 <i>jí</i>	*[m-k]rəp	gip
late	<i>chầy</i>	<i>trì</i>	(<i>muộn</i>)	遲 <i>chí</i>	*l<r>ə[j]	drij
urgent	<i>kíp</i>	<i>cấp</i>	NA	急 <i>jí</i>	*[k](r)əp	kip

ECL Calendar Words

Calendar terms are abstract cultural words rather than function words. However, considering that (a) they overlap in function with other time words, (b) they constitute a semantic system, and (c) there are a dozen ECLs in this category, they deserve attention. The core concept of the Chinese New Year is an ECL, *tết*, as noted in Table 16. As for seasons, Vietnamese has borrowed all four terms for seasons, but only ‘summer’ belongs to the category of ECL, while the other three are of the later Sino-Vietnamese stratum. The only ECL time word for a part of the day is for ‘noon’ (though the native *trưa* ‘noon’ is the dominant one in modern usage). This word is listed as the standard Sino-Vietnamese pronunciation, suggesting a Late Middle Chinese borrowing, but its vowel (i.e., ‘o’ which is statistically mostly associated with native words, instead of the expected ‘ô’) and tone (i.e., *nặng* instead of the expected *ngã*) suggest it was borrowed in an earlier period and retained as a formal Chinese character pronunciation.

The Chinese calendar system, along with its many names of years and months, is more complicated than what can be described here. Readers can readily locate sources listing these forms in various recipient languages in the region. Lexical evidence demonstrates that the Vietic spoken in northern Vietnam after Chinese groups settled borrowed this calendar system early in the first millennium CE, as demonstrated by the phonological features as well as the spread of the system to other languages in the region. Based on the Chinese system of twelve earthly branches with animal terms, Vietic speakers developed their own version with several native etyma.¹ This system subsequently spread to Old Khmer (Ferlus 2013). It is this nativized Viet-Muong system that is in both Vietnamese and Muong (see Ferlus 2013), while I can find no source of these original ECLs in available Muong sources. Regardless, we can hypothesize that the original Chinese system, and thus the ECLs in Table 16, were borrowed before that second native system was developed.

What is notable in Table 16 is that there are only six ECLs of the twelve Chinese year terms, and of these, the first four are formally listed in dictionaries of Chinese character readings (an interesting detail as others were replaced by later-stage pronunciations). However, all four have clear ECLs phonological features, notably tones, which place these borrowings in the first half of the first millennium. The reason for these seeming scattered replacements of earlier forms is unclear.

Table 16: Early Chinese loanwords for Chinese calendar terms

Gloss	ECL	SV	Muong	Chinese	OC	MC
new year	<i>tết</i>	<i>tiết</i>	<i>thết</i>	節 jiē	*ts ^h ik	tset
noon	<i>ngọ (SV)</i>	<i>ngọ</i>	(<i>cạ nửa ngày</i>)	午 wǔ	*[m].q ^h aʔ	nguX
summer	<i>hè</i>	<i>hạ</i>	(<i>nóng</i>)	夏 xià	*[g] ^h raʔ	haeH
1 st month	<i>giêng</i>	<i>chinh, chính, chánh</i>	<i>chiêng</i>	正 zhēng	*C.teŋ	tsyeng
12 th month	<i>chạp</i>	<i>lạp</i>	<i>chạp</i>	臘 là	*C.r ^h ap	lap
1 st earthly branch	<i>tí (SV)</i>	<i>tử and tí</i>	NA	子 zǐ	*tsoʔ	tsiX
2 nd earthly branch	<i>xú (SV)</i>	<i>sửu and xú</i>	NA	丑 chǒu	*ŋruʔ	trhjuwX
3 rd earthly branch	<i>dần (SV)</i>	<i>dần</i>	NA	寅 yín	*[g](r)ər	yin
4 th earthly branch	<i>mẹo (SV)</i>	<i>mão and mẹo</i>	NA	卯 mǎo	*m ^h ruʔ	maewX
5 th earthly branch	<i>thìn</i>	<i>thần</i>	NA	辰 chén	*[d]ər	dzyin
8 th earthly branch	<i>mùi</i>	<i>ị</i>	NA	未 wèi	*m[ə]t-s	mj+jH
10 th year in the cycle	<i>dậu</i>	<i>dậu</i>	NA	酉 yǒu	*N-ruʔ	yuwX

3.5 Measure words

WOLD does not include a category for measure words. The container word ‘bottle’ has a high borrowability rate of 0.60, and it overlaps with the class of measure terms, but this data does not allow generalizations. Still, words for containers, and thus related to trade, are in a sociocultural situation that can lead to lexical borrowing. The focus here is on general measure words, while classifiers cannot be considered in the earliest period of Sinitic-Vietic language contact. Fully grammaticalized classifiers in Chinese are not clearly attested until the later first millennium CE (e.g., Peyraube 1996, Behr 2009).

¹ I take a different position from that of Norman (1985) and Ferlus (2013), who followed Coedès’s (1935) reasonable claim of a Viet-Muong origin of the dozen-animal calendar in Khmer and Tai languages. Norman proposed an Austroasiatic origin of the system in Chinese, and Ferlus attempted to provide further phonological support. This calendar system is not practiced among hilltribe Austroasiatic groups, not even (to my knowledge) conservative Vietic groups, so it appears limited to Viet-Muong. Also, I do not find their phonological evidence of proposed loanwords into Chinese to be strong enough to support these claims. Also, it is problematic to claim the borrowing of the word for ‘horse’ since horse-raising could not have been part of Vietic culture before the Oracle bones, when the Chinese calendar is attested. Regardless of the origins of the system in Chinese, the phonological data in Table 16 leans strongly in the direction of borrowing the words and overall system from Chinese in the early centuries of the 1st millennium CE.

The category of measure words in Vietnamese shows clear lexical influence of Sinitic from the early period. Only several Vietic items have been identified in Vietnamese, as in Table 17, whereas Table 18 contains seventeen ECL measure words. The later Sino-Vietnamese layer contains many more measure words, though these are not in the scope of the current study, but overall, this highlights the long-term tendency for such words being borrowed in Viet-Muong languages. The classifier *cái* (generic classifier) has also been borrowed in Tai (see Alves 2015a), a kind of indirect evidence of the shared period of borrowing of these terms. However, Proto-Tai and later stage Tai vocabulary includes a significant number of grammatical ECLs not seen in Vietnamese, highlighting the distinct situations of Vietic and Tai with Sinitic.

Table 17: Proto-Vietic measure words

Gloss	Proto-Vietic	Vietnamese	Muong
bunch (of bananas)	*bɔːŋ (AA *buuŋ)	<i>buồng</i>	(<i>tlóc</i>)
bunch/bouquet	*pɔːʔ	NA	<i>bó</i>
fathom	*p-laːs	<i>sải</i>	NA
handful of bananas	*c-naːs (< c-rn-aːs ?)	NA	<i>nải</i>
handful/contents of two cupped hands	*pɔːk	<i>vốc</i>	<i>pốc</i>
lump	#kok	<i>cục</i>	<i>cóc</i>
mouthful/piece of	*-mɛːŋʔ	<i>miếng / mảnh</i>	<i>miếng</i>
span	*c-kaːŋ	<i>gang</i>	(<i>nặng</i>)

A key question is how Vietic speakers incorporated these measure words into noun phrases in that early period. In modern Vietnamese noun phrases, the order is quantity, measure or classifier, and the head noun, followed by other modifiers (e.g., *một quả cam tươi* (one.CLSF.orange.fresh) ‘one fresh orange’). I have suggested that borrowing these kinds of words from Chinese impacted Vietnamese noun phrase structure as their pre-noun position goes against the typology in the region (Alves 2001). No textual evidence for Viet-Muong precedes the second millennium CE, but in the earlier part of the second millennium, texts show classifiers as optional and their position in noun phrases variable, both before and after nouns (Vũ 2014). As for the early Vietic noun phrase, I have hypothesized that quantity expressions were originally in post-nominal position, this order in Old Khmer texts from the first millennium CE and the implied change in progress considering the variable position in early Vietnamese Nôm texts (Alves 2020). If so, we can speculate that there was stimulus to use such ECLs in this post-nominal position at some point after the borrowing of Chinese measure terms but before numerals and quantity terms could be moved in front of head nouns. This change in Vietic noun-phrase word order may have been in progress for centuries, but no specific time can be offered based on existing data.

However, as noted in Section 3.2, the Vietic numeral system was not impacted by Sinitic. Native numerals moved to the front position along with Sinitic general quantity expressions and measure words. Perhaps this combined movement of native numeral terms with measure words is related to the loss of Annamese Chinese. We can assume that bilingual Annamese Chinese used Chinese numerals with Chinese measure words or classifiers in the pre-noun position, but this speech community eventually shifted to Viet-Muong. That bilingual situation could have contributed to the variability in word order of quantified noun phrases Viet-Muong noun-phrase, even as the Chinese numerals stopped being used. Again, there is no concrete evidence of this, making it speculation, but the speculation is based on observable data and possible language contact scenarios.

Table 18: Early Chinese loanwords for unit terms

Category	Gloss	ECL	SV	Muong	Chinese	OC	MC
Architecture	level/floor	<i>tàng</i>	<i>tàng</i>	<i>thòng</i>	層 céng	*N-s-tʰəŋ	dzong
	story, floor, building	<i>lầu</i>	<i>lâu</i>	<i>lầu</i>	樓 lóu	NONE	NONE
	unit for buildings	<i>căn</i>	<i>gian</i>	NA	間 jiān	*kʰre[n]	kean
Classifiers	classifier for vehicles	<i>chiếc</i>	<i>chích</i>	<i>chiếc</i>	隻 zhī	*tek	tsyek
	generic classifier	<i>cái</i>	<i>cá</i>	<i>cái</i>	個 gè	*kʰa[r]-s	kaH
	measure for pictures	<i>tranh</i>	<i>tránh</i>	NA	幀 zhēn	NONE	NONE
	measure word for groups of animals, classifier for elephant, garden, raft	<i>thót</i>	<i>thát</i>	NA	匹 pǐ	*pʰi[t]	phjit
	unit for flat things	<i>bức</i>	<i>phúc</i>	NA	幅 fù	*pək	pjuwk
General	situation/classifier for activities	<i>cuộc</i>	<i>cục</i>	<i>cuộc</i>	局 jú	*[g](r)ok	gjowk
	type	<i>giống</i>	<i>chủng, chúng</i>	<i>chống</i>	種 zhǒng	*k.təŋʔ	tsyowngX
	type/species	<i>loài</i>	<i>loại</i>	<i>thứ</i>	類 lèi	*[r]u[t]-s	lwijH
	part	<i>phần</i>	<i>phận</i>	<i>phần</i>	分 fēn	*[m]-pə[n]-s	pjun
	measure for divinations	<i>quẻ</i>	<i>quái</i>	NA	卦 guà	*[k]wʰre-s	kweaH
	ten thuoc in length (archaic)	<i>dương</i>	<i>trượng</i>	NA	丈 zhàng	*[d]raŋʔ	drjangX
Trade	unit of measurement (for cereals)	<i>thung</i>	<i>thặng</i>	NA	升 shēng	*s-təŋ	syng
	peck of	<i>đầu</i>	<i>đầu; đầu</i>	<i>tầu</i>	斗 dǒu, đầu	*tʰoʔ	tuwX
	tael	<i>lạng</i>	<i>hạng</i>	<i>lạng</i>	量 liàng	*[r]aŋ-s	ljang

3.6 Conjunctive words

The term “conjunctive” is here used loosely as the words in Table 19 include adverbial conjunctive words as well as conjunctions. WOLD contains only a few conjunctive words. In the database, the rate for ‘because’ is somewhat high at 0.35, while ‘and’ has borrowability rate of 0.19, towards the lower end of the range. The sense of ‘with’ is very low at 0.09, but overlap of usage of words ‘with’ and ‘and’ is common in languages of Asia, which complicates the situation.

The number of ECLs in this domain is small, but the items are functionally significant. As shown in Table 19, it is precisely these words ‘with’, ‘and’, and ‘because’ that have been borrowed in the early period. Details of multiple borrowings of Chinese *gòng* 共 ‘altogether’, a triplet of two ECLs and a Sino-Vietnamese morph, have been described elsewhere (Alves 2018b).

In contrast to several conjunctive ECLs, only one native conjunction, *hay* ‘or’, Proto-Vietic #hi:, is in Vietnamese. Again, additional conjunctive words are seen in the later Sino-Vietnamese layer, but the ECLs already show impact on multi-clausal constructions. One speculation to make is that the use of parataxis for conjunctive, cause-effect, and conditional (e.g., Ruc in Vietic (Nguyễn V. L. 1993:125, Pacoh in Katuic (Alves 2015b:892-893)) left gaps to fill with a lexeme. Another possibility is Chinese had a somewhat more formal status and involved higher frequency usage of connective words. This is the case in modern Southeast Asian languages: parataxis is more associated with minimal lexical marking, while more formal registers tend to include lexical clause linkers (Jenny 2021:608-611).

Table 19: ECL conjunctions

Gloss	ECL	SV	Muong	Chinese	OC	MC
and	và	hoà	(pợi)	和 hé	*[c]ʰoj	hwa
and, with	cùng	cộng	công	共 gòng	*N-k(r)onʔ-s	gjowngH
also	cũng	cộng	(i)	共 gòng	*N-k(r)onʔ-s	gjowngH
because	vì	vị	(tại)	為 wèi	*cʷ(r)aj-s	hjweH
if/supposing	giá	giả	đả	假 jiǎ	*Cə.kʰraʔ	kaeX

3.7 Comparative terms

The data in WOLD includes only two comparative senses: ‘similar’ with a borrowability rate of 0.17 and ‘more’ at 0.23, both medium-low rates. The former sense is seen in corresponding ECLs in Table 20, while the latter is manifested as an instance of native grammaticalization and likely a part of a regional typological tendency.

Vietnamese has only one comparative term that stems from an early reconstructable etymon: Vietnamese *hơn* ‘more than’, Proto-Vietic #hə:n, a possible grammaticalization of Proto-Austroasiatic *hən, *hə:n ‘to grow/to increase’ (see Alves, Jenny, Sidwell 2020:325). This grammaticalization path fits into the ‘surpass’ type construction seen among languages in the region, such as Yue Chinese, Tai, Lao, and others (see Ansaldo 2010). However, there is currently insufficient data to posit the structure of comparative constructions in early Vietic preceding this grammaticalization event. It is used with the comparative function in other Vietic languages, but this etymon has not been carefully studied in them, nor have I found studies of it in early Nôm texts. Lastly, the common Vietnamese intensifier *lắm* ‘very much’ is restricted to Viet-Muong languages, so ultimately, there is little reconstructable lexical data in this domain.

Table 20: Early Chinese loanwords for comparison terms

Gloss	ECL	SV	Muong	Chinese	OC	MC
<i>as much as</i>	tày	tề	NA	齊 qí	*[dz]ʰəj	dzej
<i>contrary</i>	ngược	ngịch	ngươc	逆 nì	*ŋrak	ngjaek
<i>equal; flat</i>	bằng	bình	pằng	平 píng	*m-breŋ	bjaeng
<i>more so</i>	càng	cánh; canh	cặng	更 gèng, gēng	*kʰraŋ-s & *kʰraŋ	kaeng
<i>rather</i>	khá	khả	(khí hơi)	可 kě	*[k]ʰəa[j]ʔ	khaX
<i>similar</i>	tựa	tự	NA	似 sì	*sə.ləʔ	ziX
<i>to resemble</i>	tợ	tự	NA	似 sì	*sə.ləʔ	ziX
<i>to compare</i>	ví	tỷ	NA	比 bǐ	*C.pijʔ	pjiX

As for ECLs, there are several terms that express comparison and similarity in Table 20. Chinese 似 *sì* ‘similar’ was borrowed twice in the early pre-Sino-Vietnamese period, resulting in a triplet. Of the two ECLs, distinguished only by vowel type, I am not sure which is the older borrowing, only that they have vowels corresponding to Old Chinese reconstructions (also note Schuessler’s Late Han Chinese *ziə^B with a diphthong). Vietnamese *khá* ‘rather’ is from Chinese 可 *kě* ‘able’, but it also has the sense of ‘very’ as an intensifier for adjectives, though this is less common in Chinese and somewhat more archaic. Vietnamese *bằng* ‘equal to’ is a probable grammaticalized form of Chinese ‘level’. As it is also in all thirty varieties of Muong as well (Nguyễn V.T. 2005:176), it most likely stems to a pre-Viet-Muong period, and we can even hypothesize that this could represent a development in Annamese Chinese. There is, of course, no means of verifying or excluding this idea, but it would seem more likely than grammaticalization of a Chinese morph only in pre-Viet-Muong and not the source variety of Chinese. A last thought is that, as suggested in Section 3.6 about conjunctive words, the borrowing of comparative words may have filled a lexical gap as some evidence in some data shows parataxis as a means of comparative structures, but again, this is not a testable hypothesis.

3.8 Modal words

Modal words are not included in the data of WOLD, so there is no statistical point of reference for borrowing of such words. It is, nevertheless, reasonable to assume that loanwords expressing modality would most likely be borrowed in a bilingual community or other significant language contact situation. As for native etyma, there is no clear evidence of reconstructable modal elements at the Proto-Vietic level, whether preverbal words or sentence-final particles. There is the Central Vietnamese negative morph *nỏ* ‘no/not’, for which there is a reconstructed Proto-Vietic *-nɔh. However, with attestations in only a variety of Muong and four Pong lects, it can at best only be connected to a later sub-branch stage within Vietic.

In contrast, there are several probable ECLs expressing types of modality (condition, obligation, ability, passive voice, and emphasis), as shown in Table 21. Not all of these are in widespread usage in modern Vietnamese, but considering the lack of native modal terms, these loanwords are, or were, significant as they serve core modality functions. The ECL meaning ‘to stop’ and the prohibitive meaning ‘don’t’ is a probable grammaticalized function in Vietnamese. This meaning extends at least back to the 1651 de Rhodes’ dictionary, so it is not a recent development. However, I find no attestations of it in the Muong Bi data, and it is not included in Nguyễn V. T.’s (2005) comparative Muong data, so it might not date to the Proto-Viet-Muong stage. Careful sifting of earlier Nôm texts might or might not show the use of this word, which is normally a spoken functor. Regarding *nỏi* ‘able’, the semantics are slightly different from the Chinese source, suggesting possible grammaticalization in Vietnamese, but the form matches the expected form well. Finally, *được* ‘get/able’ matches well the function in Chinese, and the segments are reasonable, but the tone height is unexpected, but not enough to exclude it as there are no viable alternative etymological sources.

Table 21: ECL modal words

Gloss	ECL	SV	Muong	Chinese	OC	MC
able	<i>nỏi</i>	<i>nại</i>	<i>nỏi</i>	耐 <i>nài</i> ‘able to endure’	*nʰə-s	nojH
by, due to	<i>bởi</i>	<i>bị</i>	<i>pỏi</i>	被 <i>bèi</i>	*m-pʰ(r)ajʔ-s	bjeH
don’t (prohibitive)	<i>đừng</i>	<i>đình</i>	(<i>chở</i>)	停 <i>tíng</i>	*Cə.[d]ʰeŋ	deng
get/able/(passive)	<i>được</i>	<i>đắc</i>	(<i>ản</i>)	得 <i>dé</i>	*tʰək	tok
must	<i>tua</i> (<i>archaic</i>)	<i>tu</i>	NA	須 <i>xū</i>	*[s]o	sju
sentence particle (emphatic)	<i>thay</i>	<i>tai</i>	NA	哉 <i>zāi</i>	*[ts]ʰə	tsoj

One problematic item is worth noting. The preverbal *bị*, from the Chinese *bèi* 被 passive marker, is a common adversative (i.e., indicates a negative effect) passive marker in Vietnamese, but it is relatively recent development, not one seen in early Nôm texts. In Table 19, a preposition with passive-like features *bởi* ‘by/due to’ has ECL phonological features, though it is in need of checking textual data to see how early it can be attested, making it somewhat tentative.

Lastly, Vietnamese *phải* has a partially passive-like function, in addition to the senses of ‘correct’ and ‘must’. It is even listed with this sense in de Rhodes’ 1651 Vietnamese-Portuguese-Latin dictionary. This word has a superficial comparable form to Old Chinese *m-pʰ(r)ajʔ-s (MC bjeH, SV *bị*) of Chinese *bèi* 被. However, the passive function in Sinitic was not developed at the stage of Old Chinese, but rather has been hypothesized to have developed only in the late first to early second millennium (Peyraube 1996:177). Instead, this term likely followed a grammaticalization pattern in Southeast Asia involving words meaning ‘correct’, ‘must’, and/or ‘contact’ which developed a passive-like function. This is the case with Khmer *trəw* ‘correct/must’ and Thai *thuək* ‘correct, conforming; to contact’, both of which are also passive markers in those languages (see Matisoff 1991:425-426, but his claim of shared origin of Vietnamese *được* and Thai *thuək* is not supported by historical phonological data). This has developed even in the Vietic language Thavung, with a distinct etymon *cəh* meaning ‘correct’ and

also marking the passive voice. Thus, comparable evidence suggests it is an instance of chance phonological similarity of a distinct etymon.

4 Summary of the data and the extent and limits of early structural impact

The situation presented in the previous sections both highlights core native Austroasiatic and Vietic elements in Vietnamese grammatical vocabulary and shows clear evidence of early grammatical lexical borrowing from Sinitic in the first several centuries of the first millennium CE. The categories of grammatical categories of retentions and borrowing can be summarized as follows.

1. **Categories of significant lexical retentions:** A solid set of numeral terms ('1' to '10'), core pronouns (1s, 2s, 3s, 2p), question words ('what', 'where', 'who'), locational words (several terms), and time words ('day', 'month', 'year', etc.) are native retentions. These are precisely the types of vocabulary commonly considered in establishing linguistic affiliation and thus underscore the Vietic and Austroasiatic origins of Vietnamese, despite the long-term impact of language contact with and lexical borrowing from Sinitic and later stages of Chinese languages.
2. **Categories with significant quantities of loanwords:** Sinitic has contributed to pre-Viet-Muong Vietic a modest number of locational words and a significant number of measure words and connective and modal terms. ECLs with temporal functions include several of the twelve-animal zodiac calendar, which thus represent a kind of cultural borrowing, but together with other early loanwords related to time, the semantic domain of time was notably lexically impacted in this part of Vietic from this early period.
3. **Areas of uncertainty:** Some issues are uncertain or unanswerable. **Chance similarity** is still possible, and so with few exceptions, absolute certainty cannot be claimed. The issue of **grammaticalization** highlights the challenges in identifying loanwords. Some of the Austroasiatic etyma and ECLs have undergone grammaticalization at later periods after speciation of Vietic or Viet-Muong, and the precise timing of the grammaticalization of classifiers and passive-voice markers in Chinese is still under consideration. Moreover, some of the types of grammaticalization are seen in various languages in Southeast Asia. Thus, we cannot always know with complete certainty whether words were borrowed as grammaticalized morphs. Also, the **timing of some developments** in Viet-Muong are unclear, some clearly occurring in the second millennium, but in other cases, evidence for structural changes in the first millennium of the ECL period is lacking or is unclear.

Overall, despite any uncertainties, the data shows Vietnamese grammatical vocabulary contains a core of Vietic etyma (numerals, pronouns, locational terms, etc.) with the borrowing of a significant number of grammatical ECLs with the kinds of shared functional elements (measure words, aspectual words, comparative words, etc.) that likely facilitated communication in a bilingual community.

This suggests a question: Can a language borrow grammatical vocabulary but not undergo morpho-syntactic or semantico-syntactic change? For pre-Proto-Viet-Muong, there is no direct evidence, such as textual data, to answer this question. Only in the early second millennium is there textual data of archaic Vietnamese to show possible impact of language contact with Sinitic, and thus any changes had to have occurred before that, but with no means of determining timing (i.e., anytime from 1 CE to the turn of the second millennium). Regardless, I argue that these grammatical ECLs could have largely fit into the existing Vietic morphosyntactic structures and may not have impacted Vietic syntactic structures for several centuries approaching the development of the Viet-Muong sub-branch. Also, crucially, there are instances in which the Sinitic elements have been fit into the typological structures of Viet-Muong. An example is the post-nominal modifiers for locative nouns (Section 3.3). While there is no existing textual data for that early period of pre-Proto-Viet-Muong, it seems possible, and even probable, that other instances of structural adaption of grammatical ECLs occurred.

Furthermore, numerous grammatical domains in Vietnamese have seen little to no influence due to contact with Sinitic in that early period (e.g., negation terms (unlike Tai, see Pittayaporn et al. 2014), core numerals (again, unlike Tai), clause-final particles, etc.). Also, Vietnamese lacks many key Chinese typological features (e.g., phrase-final nominalizing particles, post-nominal modifiers, the A-

not-A question pattern, clause-initial time adverbs, etc.). Vietnamese also has typological features that are unlike those in varieties of Chinese (e.g., post-clausal adverbial elements, intensifiers and negation words in both pre- and post-verbal positions, etc.). As both core pronouns and clause-final particles have pragmatic functions, their lack of borrowing from Sinitic offers a sense of the limits on the lexical borrowing and the nature and intensity of Sinitic-Vietic bilingualism as well as the limits of the ultimate linguistic impact of Sinitic on late Vietic, certainly in the pre-Viet-Muong period. Thus, it is reasonable to speculate that in the first several centuries of Sinitic-Vietic contact, there was a substantial amount of bilingualism, and that the evidence supports a scenario of a socioculturally robust Vietic speech community speaking a language with Austroasiatic typological and structural features and native grammatical vocabulary for centuries into the period of Chinese settlement in the region.

To provide additional context for the borrowing of grammatical morphs and questions of related syntactic structure, it is also useful to consider the differing language contact circumstances in the pre-Viet-Muong period in the first millennium versus the developments in Viet-Muong moving into the second millennium. In the first millennium, the pre-Viet-Muong Vietic still would have had a typical early Austroasiatic typological template (i.e., non-tonal, polysyllabic, before regional typological convergence) before the period of Southeast Asian typological convergence. By the mid-second millennium, Viet-Muong had completely lost presyllabic material (though Vietnamese retained cluster onsets into the 1800s, and Muong lects still have them), it had fully developed tones, and it had undergone substantial typological convergence in the Southeast Asian language area. These features are listed in Table 22 (adapted from Alves 2020:54), in which hypothesized differences between Vietic and Vietnamese are highlighted.

Table 22: Linguistic structural features of Vietic versus Vietnamese

Linguistic	Vietic	Modern Vietnamese
<i>Information structure</i>	<ul style="list-style-type: none"> • Topic-comment • Middle voice only, no explicit lexical marking of the passive voice 	<ul style="list-style-type: none"> • Topic-comment • Commonly employs lexical marking of the passive voice
<i>Clauses</i>	<ul style="list-style-type: none"> • SVO/AVP 	<ul style="list-style-type: none"> • SVO/AVP
<i>Noun-phrase structure</i>	<ul style="list-style-type: none"> • Noun + quantity and modifiers 	<ul style="list-style-type: none"> • Quantity + noun + modifiers
<i>Locational terms</i>	<ul style="list-style-type: none"> • Prepositional structure, but locational nouns with postposed modifiers 	<ul style="list-style-type: none"> • Prepositional structure, but locational nouns with postposed modifiers
<i>Modality</i>	<ul style="list-style-type: none"> • Unknown 	<ul style="list-style-type: none"> • Preverbal modals, modal sentence-final particles
<i>Phonology</i>	<ul style="list-style-type: none"> • Clusters • Sesquisyllables • No tones, possible phonation 	<ul style="list-style-type: none"> • No clusters (but retained in Vietnamese into the 1800s, still retained in Muong lects) • No presyllables (but textual evidence of some in the 1200s) • Complex tone system with phonation
<i>Morphology</i>	<ul style="list-style-type: none"> • Derivational prefixes and infixes • Alternating reduplication 	<ul style="list-style-type: none"> • No affixes, only compounding • Alternating reduplication

Connecting lexical borrowing and syntactic structural aspects, we can see a correspondence between grammatical ECLs (i.e., measure words and passive markers) and the change in position of quantity phrases in noun phrases and the addition of lexical marking of the passive voice. The number and types of native Vietic measure words are limited, while a significant number of ECL measure words (but not classifiers until later) were borrowed into pre-Proto-Vietic-Muong. No native numbers were replaced, with a few early ECL numbers borrowed with highly restricted usage. In the era under consideration, Vietic likely had very different noun phrase structure from that of modern languages: strictly post-nominal elements and no grammaticalized classifiers, as comparative evidence in Austroasiatic indicates (Alves 2020). It is uncertain when or how the restructuring of the Vietic noun phrase occurred,

though the pre-nominal position of numbers and classifiers is attested in Vietnamese Nôm texts by the 1300s.

The borrowing of several ECL conjunctive, comparative, and modal words also suggests the possibility that Sinitic had a larger number of such lexemes than in Vietic at that time. Indeed, in a review of Austroasiatic grammatical vocabulary, preverbal modal verbs, sentence particles, and classifiers are all lacking, whereas there are full sets of pronominal, interrogative, and locative terms (Alves, Jenny, and Sidwell et al. 2020). Also, as noted, parataxis—juxtaposed clauses without explicit lexical marking for semantico-semantic interpretation—may have left positions to fill with such lexical items, but such a supposition will require additional data to clarify.

Recent borrowing of similar types of grammatical words can be seen among modern Austroasiatic languages in the region: So Thavung from Thai (see Srisakorn 2008), Semai from Malay (Alves fieldnotes 1998), and Chut lects from Vietnamese (see Nguyễn V. L. 1993 for Ruc, Babaev and Samarina 2018), and certainly many others. All these minority languages with less sociocultural status than the national languages they are in contact with have thus far retained their Austroasiatic typological structural profiles. In contrast with these examples, Vietic was not a minority language and was, as indicated by archaeohistorical and comparative linguistic studies, a language spoken by a community in an emerging proto-urban area (see Alves 2021). Again, native numbers, pronouns, and many locative terms were retained in Viet-Muong, so this Vietic speech community had some sociocultural status with their Sinitic-speaking neighbors.

5 Evidence of Annamese Chinese and the development of Viet-Muong

Twenty years ago, I wrote (Alves 2001:222) that the impact of Chinese on Vietnamese was “primarily of lexical influence with some accompanying phonological influence” in light of Thomason and Kaufman’s hierarchy of borrowing scale (Thomason and Kaufman 1988:74-75). I also posited that various structural typological features of modern Vietnamese reflect regional typological changes, not only the impact of language contact with Chinese. Based on the current data, this overall position is still supported, though with somewhat more evidence of structural influence of Sinitic.

I have also suggested in some publications (e.g., Alves 2009) that the language contact in the Viet-Muong period leading to borrowing, including grammatical loanwords, was influenced to a good extent by literary transmission of words, paralleling the Japanese situation, and I argued this at the time as evidence for widespread bilingualism in the first millennium in northern Vietnam was lacking. However, since then, Phan (2010, 2013) has presented arguments based on historical and linguistic evidence precisely supporting a large Chinese community in northern Vietnam in the first millennium. He posits that his coined “Annamese Chinese” developed as a substantial Chinese variety in northern Vietnam, but that this speech variety completely shifted to Viet-Muong in the early first millennium. I have since published data which further supports such a hypothetical Chinese community (hypothetical as there is no concrete historical textual description of it, only general mention of Chinese groups periodically migrating), including both lexical borrowing and structural adaptation (Alves 2016, 2017, 2018a, 2018b, 2020). This study provides further deep support via the early borrowing of grammatical morphs, in a period likely before widespread literacy in Chinese among Vietic speakers.

Indeed, at this point, the linguistic evidence in support of a culturally influential Chinese community in northern Vietnam throughout the first millennium continues to grow, making it increasingly difficult to account for the ECL data without such a community. This bilingual and bicultural scenario would eventually have had an impact on practices of literacy in northern Vietnam in the Viet-Muong speech community. We do not have a clear understanding of the development of literacy in northern Vietnam in the first millennium. Nevertheless, literacy undoubtedly became more prominent over the centuries, as evidenced by the construction of the first university in Vietnam, the Văn Miếu 文廟 ‘Temple of Literature’ in Hanoi in 1070 CE, built a century **after** Vietnamese political independence from China. Assuming a Chinese community was still part of the language ecology of northern Vietnam at that point, we can expect a lingering bilingual situation of Viet-Muong and Annamese Chinese. But we can also assume that growing literacy-based cultural practices contributed in various secondary ways to the lexical borrowing and codification of character readings of Late

Middle Chinese, but such possible impact of literacy on the sociolinguistic conditions of lexical borrowing occurred after the ECL period.

Altogether, it was the Viet-Muong speech community that borrowed the largest quantity of Sinitic vocabulary, and this community is the most typologically divergent of the Vietic sub-branches. The other sub-branches of Vietic borrowed Sinitic content words, and a few grammatical words made their way into Vietic broadly. However, the so-called “Sinification” of Viet-Muong generally and Vietnamese specifically was far from immediate, and rather took more than several centuries for Viet-Muong to undergo structural convergence with Annamese Chinese. And if one considers that (a) tones in Viet-Muong may have developed only towards the end of the first millennium (Alves 2019), (b) the complete loss of presyllabic material only occurred in the early second millennium (Shimizu 2015, Gong 2019), and (c) onset clusters with [r] and [l] lingered into the 1800s (Vu 2019), and they are still retained in varieties of Muong (Nguyễn V. T. 2005), the concept of “Sinicization”, certainly does not account for the entire typological situation in Viet-Muong. Instead, Viet-Muong is the result of very long-term language contact with Sinitic but also neighboring languages, leading to shared regional exchange and typological tendencies, all on top of an Austroasiatic typological template which also likely conditioned the ways in which some linguistic features evolved.

Acknowledgments

I would like to thank Paul Sidwell, Shimizu Masaaki, and Edith Aldridge for providing comments and suggestions for parts of the paper. I am, of course, responsible for any errors.

Article Contents

1. Grammatical features and words in Vietnamese
2. Grammatical vocabulary in Vietnamese
3. Categories of function words Vietic and Early Chinese loanwords
 - 3.1 Pronouns and Interrogative Terms
 - 3.2 Numbers and quantity expressions
 - 3.3 Locational words
 - 3.4 Time words and calendar terms
 - 3.5 Measure words
 - 3.6 Conjunctive words
 - 3.7 Comparative terms
 - 3.8 Modal words
4. Summary of data and the limits of early structural impact
5. Evidence of Annamese Chinese and the development of Viet-Muong

References

- Aldridge, Edith. 2015 (online). Old Chinese syntax: basic word order. in *The Encyclopedia of Chinese Language and Linguistics*, General Editor Rint Sybesma. http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000307
- Alves, Mark J. 1998. Sentence data collected from Semai speakers in the Cameron Highlands, Malaysia, June, 1998. Unpublished notes.
- Alves, Mark J. 2001. What’s so Chinese about Vietnamese? In *Papers from the Ninth Annual Meeting of the Southeast Asian Linguistics Society*, ed. by Graham W. Thurgood, 221-242. Arizona State University, Center for Southeast Asian Studies.
- Alves, Mark J. 2005. Sino-Vietnamese grammatical vocabulary and triggers for grammaticalization. *The 6th Pan-Asiatic International Symposium on Linguistics*, 315-332. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.
- Alves, Mark J. 2007a. Categories of grammatical Sino-Vietnamese vocabulary. *Mon-Khmer Studies* 37:217-229.

- Alves, Mark J. 2007b. Sino-Vietnamese grammatical borrowing. in *Grammatical Borrowing in Cross-Linguistic Perspective*, ed. by Yaron Matras and Jeanette Sakel, 343-361. New York: Mouton de Gruyter.
- Alves, Mark. 2009. Loanwords in Vietnamese. In *Loanwords in the World's Languages: A Comparative Handbook*, ed. by Martin Haspelmath & Uri Tadmor (eds.), 617–637. Berlin: Walter de Gruyter.
- Alves, Mark J. 2015a. Grammatical Sino-Tai vocabulary and implications for ancient Sino-Tai. Handout at the ICSTLL, 21 August 2015. DOI: 10.13140/RG.2.1.3610.8009.
- Alves, Mark J. 2015b. Pacoh. in *The Handbook of Austroasiatic Languages, Vol. 2.*, ed. by Mathias Jenny and Paul Sidwell, 881-908. Boston: Brill.
- Alves, Mark J. 2016. Identifying early Sino-Vietnamese vocabulary via linguistic, historical, archaeological, and ethnological data. *Bulletin of Chinese Linguistics* 9 (2016):264-295.
- Alves, Mark J. 2017. Chinese Loanwords in Vietnamese Pronouns and Terms of Address and Reference. *Proceedings of the 29th North American Conference on Chinese Linguistics (NACCL-29)*, Vol. 1, ed. by Lan Zhang, 286-303. Memphis, Tennessee: University of Memphis.
- Alves, Mark J. 2018a. Early Sino-Vietnamese Lexical Data and the Relative Chronology of Tonogenesis in Chinese and Vietnamese. *Bulletin of Chinese Linguistics* 11.1–2:3–33.
- Alves, Mark J. 2018b. Lịch sử của chữ 共 trong tiếng Việt [History of the character 共 in Vietnamese]. *Ngôn Ngữ* 10 (2018):3-22.
- Alves, Mark J. 2020. Initial steps in reconstructing Proto-Vietic syntax. in *Austroasiatic Syntax in Areal and Diachronic Perspective*, ed. by Mathias Jenny, Paul Sidwell, and Mark Alves, 46-81. Boston: Brill.
- Alves, Mark J. 2021. The Đông Sơn Speech Community: Evidence for Vietic. *Crossroads* 20:1-41.
- Alves, Mark J., Mathias Jenny, and Paul Sidwell. 2020. Austroasiatic affixes and grammatical lexicon. in *Austroasiatic Syntax in Areal and Diachronic Perspective*, ed. by Mathias Jenny, Paul Sidwell, and Mark Alves, 287-334. Boston: Brill.
- Ansaldo, Umberto. 2010. Surpass comparatives in Sinitic and beyond: typology and grammaticalization. *Linguistics* 48.4:919-950.
- Babaev, Kirill V. and Irina V. Samarina. 2018. *Materials of the Russian-Vietnamese linguistic expedition: May language* (in Russian). Moscow: Yask Publishing House.
- Baxter, William H. & Laurent Sagart. 2014a. Baxter-Sagart Old Chinese reconstruction, version 1.1 (20 September 2014).
- Baxter, William H. & Laurent Sagart. 2014b. *Old Chinese: A New Reconstruction*. Oxford: Oxford University Press.
- Behr, Wolfgang. 2009. Classifiers, lexical tone, determinatives: a look at their emergence and diachrony in Old Chinese and beyond. Handout from Diachrony of CLF Workshop, NIAS Wassenaar, 2009.
- Coedés, Georges. 1935. L'Origine du cycle des douze animaux au Cambodge . *Toung Pao* 31:315-329.
- Diffloth, Gérard. 1986. Proto-Austroasiatic Creaky Voice. *Mon-Khmer Studies Journal* 15:139-154.
- Haspelmath, Martin and Uri Tadmor. 2009. World Loanword Database. Web. Accessed 1 June 2021. <http://wold.livingsources.org/>
- Ferlus, Michel. 1982. Spirantisation des obstruantes médiales et formation du système consonantique du vietnamien. *Cahiers de Linguistique - Asie Orientale* 11.1:83–106.
- Ferlus, Michel. 1992. Histoire abrégée de l'évolution des consonnes initiales du vietnamien et du sino-vietnamien. *Mon-Khmer Studies* 20:111–125.

- Ferlus, Michel. 2013. The sexagesimal cycle, from China to Southeast Asia. 23rd Annual Conference of the Southeast Asian Linguistics Society, May 2013, Bangkok, Thailand. <halshs-00922842v2>
- Ferlus, Michel. 2014. Arem, a Vietic language. *Mon-Khmer Studies* 43.1: 1–15.
- Gong, Xun. 2019. Chinese loans in Old Vietnamese with a sesquisyllabic phonology. *Journal of Language Relationship* 17.1-2: 55-72.
- Guhanyu Changyongzi Zidian* 古汉语常用字字典 [Dictionary of Commonly Used Ancient Chinese Characters]. 1997. Beijing: Shangwu Yinshuguan 商務印書館.
- Guhanyu Cidian: Dongfang xuesheng gongjushu* 古汉语词典: 东方学生工具书 [Ancient Chinese Dictionary: Oriental Student's Reference Book]. 2005. Shanghai: Zhongguo Chubanshichuan, Dongfang Chubanshichuan 中国出版集团, 东方出版中心.
- Heine, Bernd and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. New York: Cambridge University Press.
- Jenny, Mathis. 2021. The national languages of MSEA: Burmese, Thai, Lao, Khmer, Vietnamese. in *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, ed. by Paul Sidwell and Mathias Jenny, 599-622. New York: de Gruyter Mouton.
- Matisoff, James A. 1991. Areal and universal dimensions of grammaticalization in Lahu. in *Approaches to Grammaticalization. Volume 2: Focus on Types of Grammatical Markers*, ed. by Elizabeth Closs Traugott and Bernd Heine, 383-453. John Benjamins.
- Meinsterernst, Barbara. 2011. From obligation to future? A diachronic sketch of the syntax and the semantics of the auxiliary verb *dāng* 當. *Cahiers de Linguistique Orientale* 40.2:3-31.
- Nguyễn, Tài Cẩn. 1995. *Giáo Trình Lịch Sử Ngữ Âm Tiếng Việt (sơ thảo)* [A Textbook of Vietnamese Historical Phonology (Draft)]. Hà Nội: Nhà xuất bản Giáo dục.
- Nguyễn, Văn Lợi. 1993. *Tiếng Rục* [The Rục language]. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.
- Nguyễn, Văn Khang, Bùi Chi, and Hoàng Văn Hành. 2002. *Từ điển Mường-Việt* [A Mường-Vietnamese dictionary]. Hà Nội: Nhà Xuất Bản Văn Hoá Dân Tộc.
- Nguyễn, Văn Tài. 2005. *Ngữ âm tiếng Mường qua các phương ngôn* [Phonetics of Muong through dialects]. Hà Nội: Nhà Xuất Bản Từ Điển Bách Khoa.
- Norman, Jerry. 1985. A note on the origins of the Chinese duodenary cycle. *Linguistics of the Sino-Tibetan Area: The State of the Art*, ed. by Graham Thurgood, James A. Matisoff, and David Bradley, 85-89. Pacific Linguistics, Series C-87. Canberra: The Australian National University.
- Peyraube, Alain. 1996. Recent Issues in Chinese Historical Syntax. in *New Horizons in Chinese Linguistics*, ed. by C.-T., James Huang and Y.-H. Audrey Li, 161–213. Dordrecht: Kluwer.
- Phan, John. 2010. Re-imagining “Annam”: A new analysis of Sino-Viet-Muong linguistic contact. *Chinese Southern Diaspora Studies* 4 (2010):3-24.
- Phan, John. 2013. *Lacquered Words: The Evolution of Vietnamese under Sinitic Influences from the 1st century BCE through the 17th Century CE*. Dissertation. Cornell University.
- Pittayaporn, Pittayawat, Jakrabhop Iamdanush and Nida Jampathip. 2014. Reconstruction of Proto-Tai negators. *Linguistics of the Tibeto-Burman Area* 37.2:151-180.
- Pulleyblank, Edwin G. 1977-1978. The final consonants of Old Chinese. *Monumenta Serica* 33 (1977-1978):180-206.
- Pulleyblank, Edwin G. 2000. Morphology in Old Chinese. *Journal of Chinese Linguistics* 28.1:26-51.
- de Rhodes, Alexandre. 1991 (translation into Vietnamese from 1651 publication). *Từ Điển Annam-Lusitan-Latinh (Thường Gọi là Từ Điển Việt-Bồ-La)* [Annamese-Portuguese-Latin Dictionary]. Ho Chi Minh City: Nhà Xuất Bản Khoa Học Xã Hội.
- Schuessler, Axel. 2007. *ABC etymological dictionary of Old Chinese*. Honolulu: University of Hawaii Press.

- Shimizu, Masaaki. 2015. A Reconstruction of Ancient Vietnamese Initials Using Chữ Nôm Materials. *NINJAL Research Papers* 9:135–158.
- Shorto, Harry L. 2006. *A Mon-Khmer comparative dictionary*, ed. by Paul Sidwell, Doug Cooper, and Christian Bauer eds. Canberra: Pacific Linguistics.
- Sidwell, Paul. nd. Austroasiatic numerals. Web.
<https://sites.google.com/view/paulsidwell/austroasiatic-numerals>. Accessed 28 June 2021
- Srisakorn, Preedaporn. 2008. So (Thavung) Grammar. Dissertation. Thailand: Mahidol University.
- Tadmor, Uri, Martin Haspelmath, and Bradley Taylor. 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27.2:226-246.
- Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
- Thompson, Laurence C. 1984-85. A Vietnamese grammar. *Mon-Khmer Studies Journal* 13-14:1-367.
- Vương, Lộc. 2002. *Từ Điển Từ Cổ* [A Dictionary of Ancient Words]. Hà Nội: Nhà Xuất Bản Đà Nẵng, Trung Tâm Từ Điển Học.
- Vũ, Đức Nghiệu. 2014. Cấu trúc danh ngữ tiếng Việt trong văn bản Phật thuyết đại báo phụ mẫu ân trọng kinh (Noun phrase structure in the text Phật thuyết đại báo phụ mẫu ân trọng kinh). *Ngôn Ngữ* 2014.1:3-19.

THE SPLIT OF PROTO-TAI VOICED STOP CONSONANTS IN NUNG OF TRANG DINH DISTRICT VIETNAM

Ayaka HIRANO
Osaka University

itsumo.kokoroni.taiyouwo.0v0@gmail.com

Abstract

In most modern Tai varieties, Proto-Tai voiced stops are devoiced, and Proto-Tai voiced stops developed into another class of sounds, but they did not split into multiple consonants. For example, Proto-Tai *b either became /p/ or /p^h/ in most modern Tai languages. However, in Trang Dinh Nung, a variety of the Nung language spoken in the Trang Dinh district, Lang Son province, Vietnam, each Proto-Tai voiced stops split into two consonants. For example, Proto-Tai *b split into both /p/ and /p^h/ in Trang Dinh Nung. This paper is aimed at showing that this phenomenon in Trang Dinh Nung is related to tone and constitutes evidence that the phonetic feature of the B tone in Proto-Tai had a low pitch.

Keywords: Tai languages, tone, consonant split

ISO 639-3 codes: cmn, ltc, nut

1 Introduction

This paper is aimed at demonstrating that each Proto-Tai voiced stop split into two consonants in Trang Dinh Nung (a variety of the Nung language used in Trang Dinh district) and showing that this phenomenon is related to tone. In Section 1, I discuss how Proto-Tai voiced stops have changed and are realized in modern Tai varieties, and I provide background information about Trang Dinh Nung, which is the subject of this paper. In Section 2, I describe the phonological system of Trang Dinh Nung, and I discuss etyma with Proto-Tai voiced stops to clarify the consonant split in Trang Dinh Nung. In Section 3, I discuss the relationship between the consonant split and tone. In Section 4, I discuss the phonetic feature of Proto-Tai B tone on the ground of the consonant split in Trang Dinh Nung.

1.1 Proto-Tai voiced stops in modern Tai varieties

Modern Tai varieties can be classified into three types from the perspective of how Proto-Tai voiced stops change. In modern Tai varieties, Proto-Tai voiced stops (1) became simple voiceless stops, (2) became aspirated stops, or (3) were retained as voiced stops. Few modern Tai varieties belong to the type 3.¹ Instead, Proto-Tai voiced stops are devoiced in most modern Tai varieties, such as *b > /p/ or *b > /p^h/.

In Trang Dinh Nung, however, the Proto-Tai voiced stops split into two consonants. For example, *b has split into /p/ and /p^h/. To the best of my knowledge, this split pattern in Trang Dinh Nung is uncommon in modern Tai varieties.²

1.2 Trang Dinh Nung

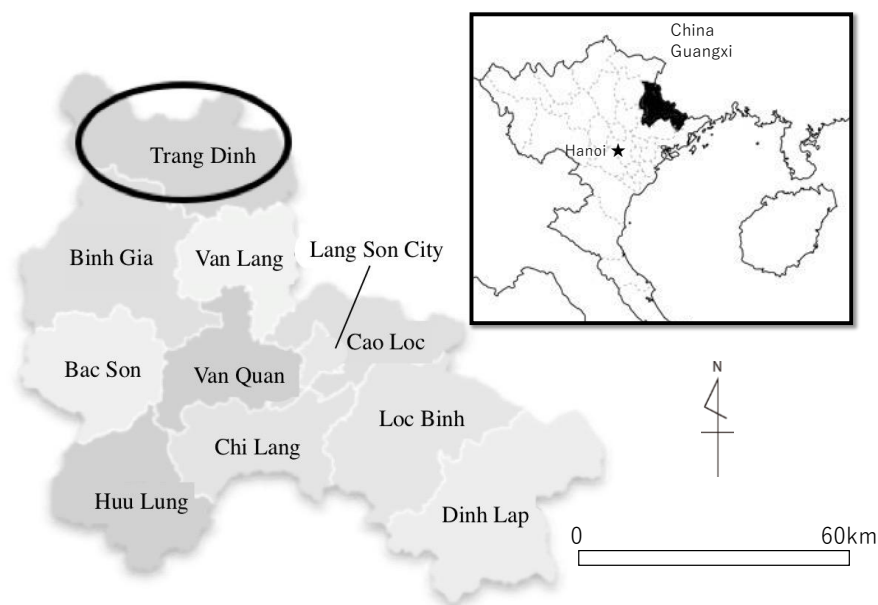
The Nung language is spoken by the Nung people, who live mainly in northeast Vietnam. In the 2019 Vietnam Population and Housing Census (General Statistic Office of Vietnam 2020), the Nung population totaled 1,083,298. Trang Dinh district is part of Lang Son province, Vietnam, as shown in

¹ The Proto-Tai voiced stops “have been devoiced in most modern Tai varieties, except for a few dialects on the Sino-Vietnamese border” (Pittayaporn 2009:110).

² Pittayaporn (2009:110) indicates that most modern Tai varieties reflect Proto-Tai voiced stops “either as plain /p-/ , /t-/ , /c-/ , and /k-/ , or as aspirated /p^h-/ , /t^h-/ , /c^h-/ and /k^h-/”.

Figure 1. The Nung language belongs to the Central Tai group of Tai languages (Li 1960). The Nung ethnic group contains subgroups, and whose names correspond with their original homelands in China. According to Phan and Khổng (eds.) (1999), there are three subgroups of Nung in Trang Dinh district: Nung Chao, Nung An, and Nung Phan Slinh. More than ninety percent of the Nung in Trang Dinh district are Nung Chao. According to Fang (1989:163), the ancestors of Nung Chao migrated from Longzhou (龍州), which is located in present Longzhou county.³ The Trang Dinh Nung data in this paper are from my fieldwork in Trang Dinh district, Lang Son province, Vietnam. The informant is a Nung Chao man, who was born and raised in Trang Dinh district.

Figure 1: Map of Lang Son Province



The syllable structure of Trang Dinh Nung is $C_1(C_2)V(C_3)/T$. Table 1 shows the Trang Dinh Nung consonants that can occur as C_1 . Two Trang Dinh Nung consonants, /b/ and /d/, correspond to the Proto-Tai implosive *b and *d, respectively. In other words, Proto-Tai voiced stops are devoiced in Trang Dinh Nung, as they are in most modern Tai varieties.

³ According to Fang (1989: 163), the Nung An people had migrated from Jie'anzhou (結安州), and the Nung Phan Slinh people had migrated from Wanchengzhou (萬承州). Jie'anzhou and Wanchengzhou are located in present Tiandong (天等) county and Daxin (大新) county, respectively.

Table 1: *C₁ consonants in Trang Dinh Nung*

	labial	dental	palatal	velar	glottal
voiceless stops	p	t		k	ʔ
aspirated stops	p ^h	t ^h		k ^h	
implosives	ɓ	ɗ			
affricate			tɕ		
nasals	m	n	ɲ	ŋ	
voiceless fricatives	f	s			h
voiced fricatives	v	z			
lateral fricatives		ɬ			
lateral		l			

In Trang Dinh Nung, /w/ and /j/ can occur as C₂. Table 2 shows the combination of C₁ + C₂ in Trang Dinh Nung. The velar sounds can occur with /w/, and the bilabial sounds can occur with /j/. Regarding /ʔw/, it is found only in onomatopoeia.

Table 2: *Combinations of C₁ + C₂ in Trang Dinh Nung*

kw	k ^h w	ŋw	(ʔw)
pj	ɓj	p ^h j	mj

Table 3 shows the consonants that can occur as C₃. The C₃ consonant /w/ can occur only after /ã/, the other C₃ consonants can occur after any vowels.

Table 3: *C₃ consonants in Trang Dinh Nung*

p	t	k
m	n	ŋ
w	j	w

Trang Dinh Nung has six tones, which are developed from Proto-Tai tones, as shown in Table 4. Although the Proto-Tai D tone has split according to vowel length in many modern Tai varieties, it has not split in Trang Dinh Nung.

Table 4: *Tonal split pattern of Trang Dinh Nung (Adapted from Gedney 1972:202)*

	A	B	C	DS	DL
Voiceless friction sounds *f-, * ^h m-, *p ^h -, ...					
Voiceless unaspirated stops *p-, *t-, *k-, ...	1 [33 ɿ]	3 [35 ɿ]	5 [213 ɿ]	3 [35 ɿ]	3 [35 ɿ]
Glottal *ɓ-, *ɗ-, *ʔ-, ...					
Voiced *v-, *m-, *b-, ...	2 [32 ɿ]	4 [11 J]	6 [32ʔ ɿʔ]	4 [11 J]	4 [11 J]

2 Reflexes in Trang Dinh Nung

Table 5 shows the Proto-Tai consonant reflexes in Trang Dinh Nung⁴. One Proto-Tai consonant corresponds to one Trang Dinh Nung consonant in principle, but *b, *d, *j, *g, and *ɣ, which are highlighted in gray in Table 5, have two reflexes in Trang Dinh Nung.

Table 5: Consonant reflexes in Trang Dinh Nung

		labial	alveolar	palatal	velar	uvular	glottal
stops	voiceless	*p > p	*t > t	*c > tɕ	*k > k	*q > k ^h	
	voiced	*b > p, p ^h	*d > t, t ^h	*j > tɕ, s	*g > k, k ^h	*g > k	
glottalized		*ʔ > ʔ	*dʔ > dʔ	*ʔj > z			*ʔ > ʔ
fricatives	voiceless		*s > ʃ	*ç > tɕ	*x > k ^h	*χ > k ^h	*h > h
	voiced		*z > ʃ	*ʒ > tɕ	*ɣ > k, k ^h		
nasals	voiceless	* ^h m > m	* ^h n > n	* ^h ɲ > ɲ	* ^h ŋ > ŋ		
	voiced	*m > m	*n > n	*ɲ > ɲ	*ŋ > ŋ		
liquids and glides	voiceless	* ^h w > p ^h	* ^h r > h * ^h l > l				
	voiced	*w > f	*r > ʃ *l > l				

Among the five Proto-Tai consonants that split into two series in Trang Dinh Nung, *ɣ is an exception: it is a fricative, not a stop, at least in the Proto-Tai reconstruction. However, the split of *ɣ in Trang Dinh Nung also occurs for the same reason as the split in Proto-Tai voiced stops, as shown in Section 2.

To show that the ‘1-into-2 split’ phenomenon in Trang Dinh Nung is uncommon in Nung varieties, I compare etyma in Western Nung from Gedney’s word list (Hudak 2008) with etyma in Trang Dinh Nung. Western Nung (Hudak 2008) has 23 initial consonants, as shown in Table 6. The voiceless stops /b/ and /d/ in Western Nung correspond to Proto-Tai *ʔ and *d, respectively. Thus, it is clear that in Western Nung, the Proto-Tai voiced stops developed into voiceless stops, similar to most modern Tai varieties. In Western Nung, “the only recorded initial cluster is [kw] in kwaay² – stubborn” (Hudak 2008:36).

⁴ The Proto-Tai reconstructions by Pittayaporn (2009) are adopted in this paper. Although there are some differences between the Proto-Tai reconstruction by Li (1977) and that by Pittayaporn (2009), the differences do not affect the conclusion in this paper.

Table 6: Onsets of Western Nung (Data from Hudak 2008)

	labial	dental	palatal	velar	glottal
voiceless stops	p	t	c	k	ʔ
voiced stops	b	d			
aspirated	p ^h	t ^h	c ^h	k ^h	
nasals	m	n	ɲ	ŋ	
voiceless fricatives	f	s	ʃ		h
voiced fricatives	v	ð			
lateral		l			
approximants			y		

Table 7 shows the codas in Western Nung. Unlike in Trang Dinh Nung, Western Nung has /ʔ/ as a coda.

Table 7: Codas of Western Nung (Data from Hudak 2008)

p	t	k	ʔ
m	n	ŋ	
w	y	ɰ	

Western Nung has six tones, which have developed from Proto-Tai tones, as shown in Table 8. Unlike in Trang Dinh Nung, the Proto-Tai D tone is split according to vowel length.

Table 8: Pattern of the tonal split in Western Nung (Data from Hudak 2008)

	A	B	C	DS	DL
Voiceless friction sounds *f-, * ^h m-, *p ^h -, ...					
Voiceless unaspirated stops *p-, *t-, *k-, ...	1 [14 ʌ]	2 [21 ɹ]	3 [22ʔ 4ʔ]	6 [55 ɿ]	2 [21 ɹ]
Glottal *ɸ-, *ɗ-, *ʔ-, ...					
Voiced *v-, *m-, *b-, ...	4 [44 ɿ]	5 [31 ɹ]	6 [55 ɿ]	4 [44 ɿ]	5 [31 ɹ]

The ‘1-into-2’ splits in Trang Dinh Nung are related to Proto-Tai tone. Tables 9 through 13 show etyma with Proto-Tai *b, *d, *ɟ, *g, and *ɣ. In these tables, Proto-Tai voiced stops (and *ɣ) are each split into two series according to Proto-Tai tones in Trang Dinh Nung, such as *b > /p, p^h/, *d > /t, t^h/, and *ɟ > /tɛ, s/. In contrast, Proto-Tai voiced stops do not split and are simply devoiced in Western Nung. Regarding *ɣ, it developed into /h/ in Western Nung.

*Table 9: Etyma with Proto-Tai *b*

Proto-Tai (Pittayaporn 2009)	Trang Dinh Nung	Western Nung (Hudak 2008)
*be: ^A ‘raft’	pe ²	peε ⁴
*be:ŋ ^A ‘expensive’	peŋ ²	peŋ ⁴
*bi: ^B ‘elder sibling’	p ^h i ⁴	p ^h ii ⁵
*bu:k ^D ‘pomelo’	pūk ⁴	

*Table 10: Etyma with Proto-Tai *d*

Proto-Tai (Pittayaporn 2009)	Trang Dinh Nung	Western Nung (Hudak 2008)
*da: ^A ‘to smear’	ta ²	
*da: ^B ‘river’	t ^h a ⁴	taa ⁵
*di: ^B ‘place’	t ^h i ⁴	tii ⁵
*daw ^C ‘cane’	tăw ⁶	
*da:k ^D ‘land leech’	tak ⁴	taak ⁵

*Table 11: Etyma with Proto-Tai *j*

Proto-Tai (Pittayaporn 2009)	Trang Dinh Nung	Western Nung (Hudak 2008)
*jim ^A ‘to taste’	teim ²	cim ⁴
*jaŋ ^B ‘to weigh’	săŋ ⁴	caŋ ⁵
*je ^B ‘to soak’	se ⁴	cii ⁵
*ja:ŋ ^C ‘elephant’	teəŋ ⁶	caəŋ ⁵
*jɤ:k ^D ‘rope’	teək ⁴	cik ⁵

*Table 12: Etyma with Proto-Tai *g*

Proto-Tai (Pittayaporn 2009)	Trang Dinh Nung	Western Nung (Hudak 2008)
*ge:ŋ ^A ‘stink bug’	keŋ ²	
*gu: ^B ‘pair’	k ^h u ⁴	kuu ⁵
*gaw ^C ‘owl’	kăw ⁶	
*gap ^D ‘narrow’	kap ⁴	

Table 13: *Etyma with Proto-Tai *y*

Proto-Tai (Pittayaporn 2009)	Trang Dinh Nung	Western Nung (Hudak 2008)
*yo:A ‘neck’	ko ²	hoo ⁴
*ye:ŋ ^B ‘shin’	k ^h ɛŋ ⁴	
*yam ^B ‘night’	k ^h äm ⁴	ham ⁴
*yo:l ^C ‘to hammer’	kɔŋ ⁶	

Table 14 summarizes the splits of Proto-Tai voiced consonants in Trang Dinh Nung. Proto-Tai voiced stops in the syllables which do not have the B tone are realized as simple voiceless consonants, whereas Proto-Tai voiced stops in the syllables with the B tone are aspirated in Trang Dinh Nung. For instance, *J, it is realized as /tɕ/ in the syllables which do not have the B tone and as /s/ in those with the B tone. As for *ɣ, although it is not a stop, but a fricative in Proto-Tai reconstruction, it also splits into /k/ and /k^h/, similar to the voiced stop *g. We can therefore hypothesize that *y merged into *g during the early stage. After that, it split into /k/ and /k^h/.

Table 14: *Pattern of Proto-Tai voiced consonant splits in Trang Dinh Nung*

	A, C, D	B
*b	p	p ^h
*d	t	t ^h
*J	tɕ	s
*g	k	k ^h
*ɣ	k	k ^h

3 Relationship between aspirated sounds and tone height

The ‘1-into-2’ split in Trang Dinh Nung is very important for developing a better understanding of the Proto-Tai B tone. A similar split occurs in modern Beijing Mandarin — the ‘ping song ze bu song [平送仄不送]’ (level-aspirated / non-level, unaspirated) phenomenon. Middle Chinese voiced stops in the syllables with level tone became devoiced and aspirated, whereas the same initials in the syllables with other tone categories are realized as unaspirated in modern Beijing Mandarin. For example, in Middle Chinese, 平 píng ‘level’ and 病 bìng ‘sick’ have the voiced stop *d, and 平 píng ‘level’ has a level tone, whereas 病 bìng ‘sick’ has a departing tone. In modern Mandarin, 平 píng ‘level’ is /p^hiŋ¹/; that is, *b is realized as /p^h/. In contrast, 病 bìng ‘sick’ is /piŋ⁴/; that is, *b is realized as /p/, as shown in Table 15.

Table 15: *‘Ping song ze bu song’ phenomenon in Chinese*

Character	Middle Chinese (Baxter and Sagart 2014)	Beijing Mandarin
平 ‘level’	*bjaeng / level tone	píng /p ^h iŋ ¹ /
病 ‘sick’	*bjaeng / departing tone	bìng /piŋ ⁴ /

The ‘ping song ze bu song’ phenomenon might have occurred because of pitch height. Although it is impossible to describe the pitch height of each tone in Middle Chinese concretely, some scholars such as Pulleyblank (1978:178) have pointed out that, based on historical documents, the level tone in Middle Chinese had a low pitch. Chen (2015:100) notes the tendency of voiced stops to cause breathy voice and the relationship between breathiness and low tone height. Thus, we can hypothesize that the level tone in Middle Chinese preserved the breathiness of voiced initials because it was low.

4 Proto-Tai B tone

Although not all scholars have categorized Proto-Tai as a tonal language, some scholars suggest that in Proto-Tai, pitch height was part of a tonal contrast. For example, Pittayaporn (2009:271) argues that the pitch height of the B tone was low. In contrast, Liao (2016) argues that the earlier stage of Proto-Tai had non-tonal structures (p. 120) and proposes that the B tone in the later stage of Proto-Tai should be high-falling due to influence from Middle Chinese as well as the typological collocation in Mainland Southeast Asian languages between tone pitch and phonation voice (p. 102). Considering the split of voiced stops in Mandarin, which is discussed in Section 3, the ‘1-into-2’ split in Trang Dinh Nung might suggest the Proto-Tai B tone was low.

5 Conclusion

Proto-Tai voiced stops became simple voiced stops or voiceless aspirated stops in most modern Tai varieties. In contrast, Proto-Tai voiced stops split into two consonants depending on the Proto-Tai tones in Trang Dinh Nung. In the syllables which do not have the B tone, Proto-Tai voiced stops are realized as simple voiceless consonants, whereas Proto-Tai voiced stops in the syllables which have the B tone are realized as aspirated in Trang Dinh Nung.

This ‘1-into-2’ split in Trang Dinh Nung is very important to improving our understanding of the phonetic features of the Proto-Tai B tone. It may be evidence that the pitch height of the B tone in Proto-Tai was low.

References

- Baxter, William H. and Sagart, Laurent. 2014. *Old Chinese: A New Reconstruction*. Oxford: Oxford University Press.
- Chen, Zhongmin [陈忠敏]. 2015. Qisangyin yu Didiao. [气嗓音与低调]. [Breathy Voice and Low Tone]. *Journal of Chinese Linguistics*. 43.1A:90–118.
- Fang, Honggui [范宏贵]. 1989. Woguo Zhuangzu yu Yuenan Daizu, Yizu de Gujin Guanxi [我国壮族与越南岱族、依族的古今关系]. [Ancient and Modern Relationship between Zhuang people in our country and Tay, Nung people in Vietnam]. Fang Honggui [范宏贵] and Gu Youshi [顾有识]. (eds.) 1989. *Zhuangzu Lungao*. [壮族论稿]. [Essays of Zhuang people]. 160–175. Nanning: Guangxi Renmin chubanshe.
- Gedney, William J. 1972. A checklist for Determining Tones in Tai Dialects. *Studies in Linguistics: in Honor of George L. Trager*. ed. by M. Estellie Smith. 423–437. The Hague: Mouton.
- Gedney, William J. 1989. Speculations on early Tai tones. *Selected Papers on Comparative Tai Studies*. ed. by R.J. Bickner et al. 207–228. Michigan: Center for South and Southeast Asian studies, the University of Michigan.
- General Statistics Office of Vietnam. 2020. *Kết quả Toàn bộ Điều tra Dân số và Nhà ở vào năm 2019*. [Completed Results of the 2019 Vietnam Population and Housing Census]. Hanoi: Statistical Publishing Office.
- Hudak, J. T. 2008. *William J. Gedney's comparative Tai source book*. Honolulu: University of Hawaii Press.
- Li, Fang Kuei. 1960. A tentative classification of Tai dialects. *Culture in History Essays in Honor of Paul Radin*. ed. by Stanley Diamond. 951–959. New York: Columbia University Press.
- Li, Fang Kuei. 1977. *A handbook of Comparative Tai*. Honolulu: University of Hawai'i Press.
- Liao, Hanbo. 2016. *Tonal development of Tai languages*. M.A. thesis, Payap University.
- Phan Hữu Dật and Không Diển (eds.). 1999. *Địa chí Huyện Tràng Định Tỉnh Lạng Sơn*. [Topography of Trang Dinh District, Lang Son Province]. Lang Son: People's committee of Trang Dinh District.
- Pittayaporn, Pittayawat. 2009. *The phonology of Proto-Tai*. Doctoral dissertation, Cornell University.
- Pulleyblank, E. G. 1978. The Nature of the Middle Chinese Tones and Their Development to Early Mandarin. *Journal of Chinese Linguistics*. 6.2:173–203.

THE TAI MEUAY OF BOLIKHAMXAY PROVINCE (LAOS), THEIR ETHNONYM, AND THEIR LANGUAGE

Jean PACQUEMENT
Agrégé de grammaire
jeanpacquement@gmail.com

Abstract

This article on Tai Meuay first discusses their ethnonym, attested in Vietnam (Nghệ An and Thanh Hóa) and Laos (Bolikhambxay). After introducing the language data available in the literature, it analyzes Tai Meuay data collected in the districts of Khamkeuth, Pakkading, and Viengthong of Bolikhambxay Province. It also takes into account speakers' perception of their ethnic group and language. Regarding the Tai Meuay tonal diversity, this article identifies a main type with a 123-4 split in the A, B, C and DL columns of Gedney's tone diagram, as against marginal types characterized by no split in DL. Although the considered varieties share lexical items with Tai dialects of Nghệ An, tonal features (e.g., absence of split in DS) and phonological features (e.g., the reflex of the initial *y-) indicate that Tai Meuay is affiliated with Tai Daeng within the Southwestern Tai branch of the Tai language family.

Keywords: Tai Meuay, Tai Thaeng, Tai Daeng dialects in Laos and Vietnam, ethnomimics, tone systems

ISO 639-3 codes: tyr

1 Introduction and literature overview

This article on Tai Meuay,¹ both as an ethnonym and as a dialect belonging to the Southwestern Tai branch of the Tai language family, focuses on Tai Meuay varieties spoken in a province of Laos, Bolikhambxay Province.² Earlier discussions regarding the ethnonym 'Tai Meuay' in the literature relate it to a few areas of Vietnam (Đặng 2010, Ferlus 2008, Guignard 1912, Robert 1941, Vi 1996) and Laos (Chamberlain 1984, Seidenfaden 1967).

Starting from Thanh Hóa Province in Vietnam, the French colonial administrator R. Robert (1941:8, 10) explains the context of the appellation "Tay Mroy" in his ethnographic notes about the "Tay Dèng" (Tai Daeng) of Lang Chánh District. The "Jò" (Tai Yo), who live in Thường Xuân District of Thanh Hóa and *Quỳ Châu* District of Nghệ An, are the southern neighbors of the Tai Daeng. The Tai Yo call the Tai Daeng "Tay Mroy", and even the Tai Daeng often use that appellation to refer to themselves. "Tay Mroy" thus appears to be an exonym of the Tai Daeng and is neither their main autonym nor the name of the language they speak. That use of the appellation 'Tai Meuay' to refer to

¹ For a discussion regarding the term 'Meuay' and its use, see section 2. We owe the Romanized form used in this article, 'Tai Meuay', to James R. Chamberlain (1984). 'Tai' and 'Tay' representing [ta:j] and [taj] respectively in the Romanized orthography of Vietnamese and in some usages of Romanization followed in Laos, the initial component of the phrase 'Tai Meuay' should be spelled 'Tay'. However, Chamberlain uses 'Tai' instead of 'Tay' because the Romanized form 'Tai' is the one conventionally used when referring to the Tai language family. When Romanized forms for either the ethnonym of the Tai Meuay or their language name are referred to in relationship with the authors who use them, they are enclosed in double quotation marks.

² For toponyms of Laos, this article gives Romanized forms of their names commonly used in that country (e.g., Bolikhambxay, Houaphanh, Khammouane – names of provinces –; Khamkeuth, Pakkading, Viengthong – names of districts –, etc.). There are indeed some variations in spelling, and one can meet with forms such as Borikhambxay, Houaphan, Khammouan, Khamkeut, instead of the forms given above. As for places located in Vietnam, toponyms are given in Vietnamese script.

the Tai Daeng is confirmed by the fieldwork conducted by Sầm Công Danh, a graduate student at the University of Social Sciences and Humanities (Hanoi, Vietnam), in Thường Xuân District.

Some Tai Daeng of Xamtay District (Houaphanh Province, Laos), especially those living in Muong Pao Subdistrict, say that they migrated from places they call [muaŋ^{A4} dɛ:ŋ^{A3}]³ – the Tai Daeng name of Yên Khương Commune in Lang Chánh – and [muaŋ^{A4} mɔ:t^{DL4}] – the Tai Daeng name of Bát Một Commune in Thường Xuân (Sầm Công Danh, personal communication, March 24, 2021). As a result, the fact that Erik Seidenfaden (1967:91) mentions “Thai Mũoi” together with “Red Thai” (Tai Daeng) among the Tai groups inhabiting the region of Houaphanh does not come as a surprise. However, according to extensive interactions with Tai Daeng speakers in Houaphanh Province, not only in Samtay District but also in Viengxay District and Sốp Bau District, the ethnonym ‘Tai Meuay’ seems not to be in use nowadays in Houaphanh Province.

When it comes to the Western part of Nghệ An Province, Théodore Guignard (1912:XXII), a missionary priest of the *Missions étrangères de Paris* based at Cành Tráp (in the present-day Tương Dương District) from 1892 to 1906, mentions “Thay Mườy” as a language name, and he lists “Thay Mườy” among a few dialects of the “Thay” language such as “Thay” itself, “Lao”, “Phù Thay”, and “Phuôn”. Vi Van An (1996:30–32) mentions “80,000 Tai living in three districts: Con Cuong, Tuong Duong and Ky Son”. They belong to three main groups: the “Tai Muong”, or people who “belong to the *Muong*”; the “Tai Thanh”, or people whose ethnonym “could possibly have originated from their old homeland”, who might thus be linked to either “Thanh Hoa” or “Muong Thanh (Dien Bien Phu)”; and the “Tai Muoi”, or people having migrated from “Muong Muoi, which today belongs to the district of Thuan Chau, province Son La”. According to Đặng Nghiêm Vạn (2010:80, 82), “Mường Muôi” was the capital of the Tai Dam after the 13th century, and the “Tày Muoi” are people who were displaced from “Mường Muôi”, following a rebellion against Lê Lợi at the beginning of 15th century.

The Tai dialects of that region having been documented by Michel Ferlus (2008), one can see that, in his classification, there are only two groups of dialects for that region.

- The first group consists of “Tay Yo” and “Tay Muong”, the latter being also called “Tay Pao” in Tương Dương District. The appellation “Tay Muong” can actually refer to both subdialects.
- The second group comprises Tai Daeng varieties, also called “Tay Thanh” or “Tay Meuy”.

If we follow Ferlus’ classification, the use of ‘Tai Meuay’ in Western Nghệ An, ‘Tai Meuay’ being written either “Tai Muoi” or “Tay Meuy”, is the same as in Robert’s 1941 study: it is an appellation among other ethnonyms referring to the Tai Daeng.

The last area for which the literature provides accounts about the Tai Meuay is the region of Lakxao (Khamkeuth District) in Laos,⁴ which used to belong to Khammouane Province, but has been under Bolikhamxay Province since 1986.⁵ The speakers whose varieties are dealt with by Chamberlain

³ In order to note a syllable used in a variety belonging to the Southwestern Tai branch of the Tai family, whether it is discussed as a separate word or as a part of a compound or a sentence, this article uses a transcription based on the International Phonetic Alphabet. As most syllables which are discussed are expected to have cognates in other Tai languages, we have chosen not to note the tone of a given syllable in reference to the tone system of the variety in which it is used. Instead, we give the category of that syllable in the tone diagram devised by William J. Gedney (1972), which helps summarize Tai dialects’ tone systems. If we take the example of the syllable [muaŋ^{A4}], its category, which is given in the superscript indication, is ‘A4’, with ‘A’ indicating its etymological tone, as reconstructed for Proto-Tai, and ‘4’ showing the consonant type to which its onset belongs. As the category of a given syllable in Gedney’s tone diagram remains the same (there are very few exceptions), one can easily find the tone of that syllable in varieties whose tone systems have already been documented. Figure 1 in subsection 3.1 reproduces Gedney’s tone diagram.

⁴ Lakxao, which has developed since 1975, is the present-day center of Khamkheuth District. Nape used to be the main center of the region at the time of the French colonization. The toponym ‘Lakxao’ means ‘Kilometer 20’ ([lak^{DS1} sa:w^{A4}]), the starting point being the historical administrative building of Nape, locally known as the ‘bungalow’.

⁵ Roughly speaking, the present-day Bolikhamxay Province, created in 1986, comprises the territory of the former Borikhane Province, as well as areas formerly belonging to Khammouane Province (Khamkeuth District), to Vientiane Province, or to Xiengkhouang Province (a few subdistricts of Viengthong District).

(1984:68–69) come from two villages of Khamkeuth District. The appellation ‘Tai Meuy’ appears to be the only ethnonym of those informants and is the name of their language.

The Tai Meuy spoken in Khamkeuth District was further investigated by a few researchers.

- A variety spoken in the village of Phônxy (Khamkeuth)⁶ was documented in 1993 by Thongpheth Kingsada and Michel Ferlus, who recorded both “the Matisoff 200-wordlist” and “the Gedney tone checklist” for Pranee Kullavanijaya and Theraphan L-Thongkum.⁷ The two linguists of Chulalongkorn University (Thailand) used the language data of that recording in an article in which they deal with dialects belonging to the Central and Southwestern branches of the Tai family (Kullavanijaya and L-Thongkum 1998).
- Sasithorn Onlao (2010a) investigated a part of Khamkeuth District which she identifies as the “Nam Paw basin”, for her master’s thesis which she did at Mahasarakham University (Thailand). Her study introduces six Southwestern Tai dialects spoken in the area, including “Tai-thaeng”, “Tai-mot”, “Tai-moei”, “Phoo-thai”, “Nho”, and “Laos”. The Tai Meuy variety she discusses is a variety spoken in Namphao Village.⁸ The data collected by Onlao are referred to in an article in which she focuses on the tone merger and split patterns of the dialects she deals with (Onlao 2010b).
- Souksada Southxay (2016-2017), a native of the previously mentioned Phônxy (Khamkeuth) Village, discussed “570 basic words of Meuy language” in the dissertation she submitted for the completion of her bachelor’s degree (Department of Lao Language and Culture, Faculty of Letters, National University of Laos). Furthering her studies in folk literature at the master’s degree level, she analyzed a corpus of thirty stories (Southxay 2019-2020).

In addition, Tai Meuy living in other areas of the present-day provinces of Bolikhamxay and Khammouane are mentioned in the following accounts.

- Chamberlain (1991:103) refers to “Sop Vieng” as “a Tai Moey /mëy C1/ village”. Sôp Vieng is located “near the old LS 28 airstrip at Ban Done”, that means in the present-day Chomthong Subdistrict (Viengthong District, Bolikhamxay Province).
- Chamberlain (1996:11) speaks of the “Moey” in Nakai Plateau (Khammouane Province) as one of the “Tai speaking groups [which] have recently moved to the plateau from Khamkeut district in Borikhamxay”.
- Joachim Schliesinger (2003:175) lists a few villages inhabited by Tai Meuy in the districts of Paksane and Pakkading.

As the previously mentioned studies, this article will discuss data elicited from Tai Meuy speakers of Khamkeuth District. Tai Meuy data collected in two other districts of Bolikhamxay Province, Pakkading and Viengthong, will also be included in this article.

2 The ethnonym ‘Meuy’: a linguistic analysis

Starting with Robert (1941:8), because he explains the context of the ethnonym “Tay Muoy”, his account lacks linguistic information regarding the syllable “Muoy” he gives and its meaning. Robert’s transcription of the Tai Daeng varieties spoken in “Muong Chéng” ([muan^{A4} cɛ:ŋ^{C2}], the Tai Daeng name of Lang Chánh) and “Muong Dèng” ([muan^{A4} dɛ:ŋ^{A3}]) is rather accurate for lexical items and

Ironically, Khammouane, the village to which Khammouane Province owes its name, being located in Khamkeuth District, is nowadays in Bolikhamxay Province.

⁶ As a matter of coincidence, there are Tai Meuy locations in Khamkeuth District, e.g., Phônxy, Phônxyay, Nakhua, having the same names as Tai Meuy locations in Pakkading District. As this article mentions Phônxy and Phônxyay near Lakxao, as well as Phônxy and Phônxyay in Pakkading District, the former Phônxy and Phônxyay are referred to as Phônxy (Khamkeuth) and Phônxyay (Khamkeuth) and the latter Phônxy and Phônxyay as Phônxy (Pakkading) and Phônxyay (Pakkading).

⁷ Thongpheth Kingsada and Michel Ferlus’ recording is now an open access resource in the Pangloss Collection (<https://pangloss.cnrs.fr/>).

⁸ Namphao (or “Nam Paw”, as spelled by Sasithorn Onlao), in which -phao represents the Lao syllable [p^ha:w^{A4}], is the name of both the river Namphao in Khamkeuth District and of the village Namphao named after it.

their tones in his wordlist and lexicon (Robert 1941:128–139, 141–164),⁹ but it is less reliable when it comes to ethnonyms. For example, Robert uses the spelling “Jò” to refer to the ethnonym of the Tai Yo [taj^{A4} jɔː^{C4}], but the expected form in his system of transcription is actually “Jɔ̌”. As for “Muoy”, the absence of tonal mark supports either an A4 syllable or a B123 syllable of the Southwestern Tai tone system. The interpretation of this article is that “Muoy” is an A4 syllable of Southwestern Tai, not a B123 syllable. When writing “Muoy”, Robert actually renders in Tai Daeng the Vietnamese designation ‘Mười’ of the Tai Meuy since A4 syllables of Southwestern Tai correspond to A2 syllables of Vietnamese.

With respect to the form “Muòy” given by Guignard (1912:XXII), its tone mark consistently characterizes C1, C2, and C3 syllables of Southwestern Tai in his Lao-French dictionary and has nothing to do with the way it is used in Vietnamese.¹⁰ Chamberlain (1984:68) is the first to explicitly posit “/muaj C1/”, a C1 syllable in the Southwestern Tai tone system. Ferlus (2008:299) also gives a C1 syllable, “[məːj^{C1}],¹¹ in which one will notice the reduction of the diphthong. Giving the exact tone of the ethnonym ‘Meuy’ in Southwestern Tai remains all the more relevant as it cannot be determined from the usage in Vietnamese and in Lao.

- In Vietnamese, we find, as mentioned above, ‘Mười’, an A2 syllable in the Vietnamese tone system. The diphthong is not simplified. The expected form ‘Mười’, a Vietnamese B1 syllable being the equivalent of a Southwestern Tai C1 syllable, is however the correct form, according to Vi (1996:38, and personal communication, November 11, 2018) of the Vietnam Museum of Ethnology, who is himself a member of a Thái group of Vietnam.¹²
- In Lao, one consistently finds ມຸ້ຍ [mɔ̌ːj^{C4}]. Whereas Southwestern Tai C1 syllables are pronounced with a rising tone in Tai Meuy, they are pronounced with a low-falling tone in Lao. As a result, Lao speakers are likely to interpret a C1 syllable of Tai Meuy such as [mɔ̌ːj^{C1}] as a C4 syllable of Lao. The diphthong is simplified.

2.1 The diphthong in the ethnonym ‘Meuy’ and its reduction

As could be seen from the forms mentioned above, the vowel in the term ‘Meuy’ is either the diphthong /-ua-/ or the monophthong /-ɔː-/. Three phonological rising diphthongs are found in Southwestern Tai languages: /-ia-/ , /-ua-/ , and /-ua-/. There are a few Southwestern Tai languages in which those diphthongs are simplified, such as Tai Don, Phu Thai, and others.¹³ In addition, Ferlus (2008:310) mentions the case of the “Tay Muong” spoken in Tương Dương District (Nghê An Province) and “Tay Maen”, better known in the literature as “Mène”, a Tai dialect of Khamkeuth District described as “originally spoken in Nghê An” (Chamberlain 1991): diphthongs in “Tay Muong” and “Tay Maen” are simplified in front of semivowels.¹⁴

When it comes to the Tai Meuy varieties spoken in Bolikhamxay Province, the data we have show that there are very few words in which the diphthong is simplified. Then, the diphthong which is

⁹ Robert’s transcription is clear and consistent for the tones in the A123, B4, C123, C4, and DL4 boxes of the tone diagrams used to summarize tone systems of dialects belonging to the Southwestern Tai branch of the Tai family. For a more detailed introduction to tone diagrams, see subsection 3.1.

¹⁰ Guignard (1912:XX) calls his transcription of Lao “quốc ngữ laotien”. That Quốc ngữ for Lao displays the same tone marks as the Quốc ngữ for Annamite, but uses them in a different way.

¹¹ Ferlus (2008:309) lists the word “[məːj^{C1}]” among a few lexical units of Tai Yo having no cognates in other languages of the Tai family. That remark reminds us that, even though it is possible to posit a C1 syllable of Southwestern Tai for that word on the basis of its use in a few languages such as Tai Yo and Tai Meuy, it cannot be taken into account in discussions regarding the Southwestern Tai branch of the Tai language family as a whole or, from another perspective, Proto-Southwestern Tai etyma.

¹² In Vietnam, ethnographic and linguistic accounts use the form ‘Thái’ to refer to groups speaking Southwestern Tai dialects. As for the form ‘Tày’, it is used for groups speaking Central Tai dialects.

¹³ In Phu Thai dialects, the rising diphthongs /-ia-/ , /-ua-/ , and /-ua-/ are consistently replaced by the long monophthongs /-eː-/ , /-ɔː-/ , and /-oː-/.

¹⁴ According to Frédéric Pain (personal communication, January 25, 2021), diphthongs in the Tai Muong of Tương Dương, also called Tai Pao, can be simplified in front of [-w] and [-j], but such a reduction is not “an absolute rule”, and a same speaker can say either [kua^{C1}] or [koː^{C2}] for the lexical item ‘banana’.

simplified is always /-ua-/, and its reduction occurs in front of [-j], e.g., [mɤ:j^{B4}] ‘to be tired’, [ʔɤ:j^{C3}] ‘elder sister’, as against [muaj^{B4}] ‘to be tired’, [ʔuaj^{C3}] in Lao. Finally, only the ethnonym ‘Meuay’, which is used as a separate word or as a final component in a phrase, displays either a diphthong or a monophthong. In the ethnonym, /-ua-/ and /-ɤ:-/ somehow lose their phonological status and become free variants, hence [muaj^{C1}] or [mɤ:j^{C1}] ‘Meuay’; [taj^{A4} muaj^{C1}] or [taj^{A4} mɤ:j^{C1}] ‘Tai Meuay’; [sa:w^{A1} muaj^{C1}] or [sa:w^{A1} mɤ:j^{C1}] ‘Meuay young woman’. It should be noted that, although speakers use freely [muaj^{C1}] or [mɤ:j^{C1}], an overwhelming majority of the twelve speakers from different locations with whom we could discuss the term ‘Meuay’ for this research use [muaj^{C1}] when repeating the word or when asked to pronounce it in citation form.

2.2 Another use of the term ‘Meuay’

This article has so far dealt with the term ‘Meuay’, a C1 syllable posited by Chamberlain (1984) and Ferlus (2008), as an ethnonym only. In addition, we have accounts which relate that term to places of origin of the Tai Meuay, either in Vietnam (Đặng 2010, Vi 1996) or in Laos (Schliesinger 2003).¹⁵ However, neither such accounts nor the discussion of the forms [muaj^{C1}] and [mɤ:j^{C1}] in the preceding subsection have taken into account another use of that term.

Apart from being used as a separate word or as the final component of a phrase, positions in which the term ‘Meuay’ is indeed the ethnonym of the Tai Meuay, it is found in the initial position of a phrase, especially when used as the initial component in pronominal compounds. The term ‘Meuay’ then denotes plurality, as in [mɤ:j^{C1} ʔem^{C3}] ‘we (as many people)’, [mɤ:j^{C1} t^haw^{A4}] ‘you (as many people)’, [mɤ:j^{C1} sa:^{A4}] ‘they (as many people)’.¹⁶ In that position, all the speakers are found to use the form [mɤ:j^{C1}], with the reduction of the diphthong, except one speaker who consistently uses [muaj^{C1}]. When asked to pronounce the syllable [mɤ:j^{C1}], used as the initial component of a pronominal compound, in citation form, a seven-twelfths majority of the speakers use [muaj^{C1}].

Some of the speakers when discussing the term ‘Meuay’ make the following statement: “the (Tai) Meuay are us” – [(taj^{A4}) muaj^{C1} mɤ:n^{B4} mɤ:j^{C1} tu:^{A2}] ((Tai –) Meuay – be – plural – us). That statement as well as the translation of ‘Meuay’ as [paʔ sa:^{A4} son^{A4}] ‘citizen, people’, which was proposed by one speaker in another discussion, suggest that Tai Meuay speakers do not separate the ethnonym ‘Meuay’, pronounced [muaj^{C1}] or [mɤ:j^{C1}], from the use of [mɤ:j^{C1}] to denote plurality.

2.3 The syllable [muaj^{C1}] and its semantic equivalent [puak^{DL4}]

This article, which already follows Chamberlain’s 1984 account by using the Romanized form ‘Tai Meuay’, will also follow his account by positing the phonological syllable “/muaj C1/” that he gives, with the diphthong /-ua-/, for both phonetic syllables [muaj^{C1}] and [mɤ:j^{C1}], for the three positions they can occupy in a sentence (an initial component of a phrase, a final component of a phrase, or a separate word), and for both the meanings (an ethnonym and a pluralizer) discussed above.

The use of [muaj^{C1}] as a pluralizer and the related meaning ‘group, people’ it implies (Ferlus 2008:299, 309) are worth taking into account in the present discussion, because they do not support toponymic interpretations of ‘Meuay’ such as the ones referred to above. Furthermore, [muaj^{C1}] in that use is the semantic equivalent of [puak^{DL4}] – pronounced [p^huak^{DL4}] in Thai and Lao –, and [puak^{DL4}] is used to name some ethnic groups as well. According to Chamberlain (personal communication, March 18, 2021), [puak^{DL4}] “was used for Kra peoples and for some Austroasiatic groups, apparently as a marker of lower social status, as in its use in reference to the Ksing Mul in the Tai Dam feudal system”.

As this article proposes to understand [muaj^{C1}] in comparison with its semantic equivalent [puak^{DL4}], one will note that, apart from the fact that the term [muaj^{C1}] has the same uses as [puak^{DL4}],

¹⁵ According to Schliesinger (2003:175), the place of origin of the “Tai Meuy” is “Muang Meuy, west of Hua Phan province in the most eastern part of Luang Prabang province”. A few speakers interviewed during this research stated that they had migrated from Luang Prabang three hundred years ago. However, they could not say from which part of Luang Prabang Province they arrived and did not mention a place called “Muang Meuy”.

¹⁶ Ferlus (2008: 307) notes the same use of “[mɤ:j^{C1}]” in Tai Yo.

a remark by Robert suggests that, in the relationship between the Tai Yo and the Tai Daeng, the latter could be in particular cases the ones with an inferior status. According to Robert (1941:10), the “Jò” occupied the best lands in some areas of Thường Xuân, and their chiefs would bring in “Tay Dèng” to cultivate higher valleys. The use of the appellation “Tay Muoy” by the Tai Yo when they refer to the Tai Daeng, as it is noted by Robert (1941:8), would thus be explained by a local context in which the Tai Daeng would be a labor force available for the Tai Yo.

3 The Tai Meuay language data for Bolikhamxay Province in the literature

The overview of the literature in the introduction of this article (see section 1) mentions fifteen accounts in which the Tai Meuay are dealt with in one way or another. Five of those accounts (Chamberlain 1984, Kullavanijaya and L-Thongkum 1998, Onlao 2010a, Onlao 2010b, Souththixay 2016-2017) are linguistic accounts dealing specifically with varieties spoken by the Tai Meuay in Khamkeuth District. As Onlao’s (2010b) article is nothing more than a summary of her master’s thesis (Onlao 2010a), and as we could not consult Souththixay’s 2016-2017 dissertation, the following presentation builds on the other three accounts.

3.1 Chamberlain’s 1984 account

It is a short introduction dealing with the Tai Meuay varieties spoken in two villages, Keng Bay and Nava, of Khamkeuth District. In an area described as “linguistically very rich” (Chamberlain 1984:62), Tai Meuay is “quite widespread” (Chamberlain 1984:68). The present research confirms such a qualification, especially if one compares Tai Meuay with other dialects of the same region which are spoken in a few villages or even in one village only.¹⁷

Chamberlain’s account introduces Tai Meuay along with other dialects, a good number of them being spoken in Khamkeuth District or in neighboring districts belonging to the province of Khammouane as it was before 1986. All the dialects he discusses belong to the two groups he had earlier proposed in his classification of the Southwestern Tai branch of the Tai language family, the P group and the PH group (Chamberlain 1972, 1975).¹⁸ The five dialects of the P group Chamberlain takes into account are “Tai Meuay”, “Tai Khang”, “Tai Kuan”, “Tai Maen”, “Tai Pao” (Chamberlain 1984:66–70). Information related to Tai Daeng, a P language which is not spoken in that region, is included “for comparison with Tai Meuay” (Chamberlain 1984:69). The dialects of the PH group included in Chamberlain’s account are four dialects of the “PH group Neua-Phuan languages”, namely, “Tai Nheuung”, “Tai Kaloep”, “Tai Nho”, “Phu Tai” (Chamberlain 1984: 70–76), as well as four dialects of the “PH group Lao-Southern Thai languages”, including “Yooy”, “Kaleung”, “Tai Bo”, “Yo” (Chamberlain 1984: 76–82). Chamberlain lists a few “phonological characteristics” and “lexical characteristics” for each of those dialects.

¹⁷ Among the languages which are discussed in this article, Tai Meuay is spoken in many locations of Khamkeuth District. As for Tai Thaeng (see subsection 4.4), it is less widely spoken. When it comes to languages spoken in one village of Khamkeuth District only, a good example is Saek, which is spoken in Nakadôk Village. Chamberlain (1998) is the first who mentioned Nakadôk and its distinct Saek dialect. Although a few Saek families can be found in other villages of that area and in Lakxao City itself, all of them come from Nakadôk.

¹⁸ According to Chamberlain (1975:62), the Southwestern Tai branch of the Tai language family was divided around the 8th century into two groups, which subsequently evolved independently from each other. Among the features of the two groups’ divergent evolution, Chamberlain focuses on the devoicing of voiced initial stops reconstructed for Proto-Tai (*b-, *d-, *j-, *g-) and names both groups according to the reflexes [p-] or [p^h-] of the Proto-Tai consonant *b-, hence the P group and the PH group. Other specialists of Tai historical linguistics disagree with Chamberlain’s dating of the previously mentioned devoicing sound shift and thus reject his classification. Gedney (1991:208), for example, states that “the trouble with this classification is that it uses as its basic criterion something very late in the history of these languages, but Chamberlain wants to make it very early”.

- Starting with the phonological characteristics listed by Chamberlain, one will note that the five dialects of the P group share, as expected, the reflex [k-] in words reconstructed with the initial *g-.¹⁹ The reflex [k^h-] in words reconstructed with the initial *ɣ- characterizes Tai Meuay. As for the rhyme *-au, its reflex in Tai Meuay is [ɛ:].
- Among lexical characteristics, one will notice the lexical items “paa (paah)” (‘to go’), “kii” (‘to eat’), “ʔet” or “ʔeʔ” (‘to do’).

Tone systems of the varieties dealt with for each dialect are introduced separately. They are summarized in tone diagrams adapted from the tone diagram devised by William J. Gedney (1972:434). Figure 1 shows Gedney’s tone diagram with its five columns, representing the tones reconstructed for Proto-Tai, and its four rows, referring to initial consonant types. Devised to display “a maximum of possible tonal distinctions resulting from the various types of tonal splits that has been described”, Gedney’s tone diagram helps compare Southwestern Tai varieties’ tone systems and identify tone splits and mergers which are relevant for dialect differentiation.

Figure 1: *Gedney’s tone diagram for Tai dialects (following Gedney 1972:434).*

		Proto-Tai Tones				
		A	B	C	D-short	D-long
Initials at time of tonal splits	1 Voiceless friction sounds					
	2 Voiceless unaspirated stops					
	3 Glottal					
	4 Voiced					
		Smooth Syllables			Checked Syllables	

Chamberlain (1972, 1975) changed the layout of the columns in Gedney’s diagram to A, B, C, DL (D-long), and DS (D-short). Chamberlain’s layout is followed in all the tone diagrams which were created for this article.

Chamberlain (1984:67) proposes two tone diagrams for Tai Meuay on the basis of the data he elicited from two male speakers, one aged 65 from Nava Village and another one in his seventies from Keng Bay Village. In Figure 2, we have adapted his tone diagram for the Keng Bay variety in order to display an interpretation using tone numerals²⁰ of the tones in that Tai Meuay variety.

¹⁹ With regard to the reflex [tɛ-] in words reconstructed with the initial *dz- (Chamberlain 1984:66), Tai Meuay shares it with languages of the P group, such as Tai Dam, Tai Don, and Tai Daeng. However, the Tai Maen and Tai Pao varieties studied by Chamberlain, although they belong to the P group, are characterized by the reflex [s-].

²⁰ In the tone diagrams which were created for this article, tones are represented with tone numerals. This notation of tones, devised by Yuen Ren Chao (1930), focuses on the pitch, which it indicates on a five-point scale, with 1 being the lowest pitch and 5 being the highest. In order to describe a particular tone’s pitch and contour, tone numerals appear as sequences of numbers representing the starting point, change points (if any), and the end point of the concerned tone’s fundamental frequency (F0) curve.

Figure 2: *The tone diagram for Keng Bay Tai Meuay with tone numerals (following Chamberlain 1984:67).*

Initial\Tone	A	B	C		DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[423]	[554]	[445]		[554]	[34]
2 Voiceless unaspirated stops, *p, etc.						
3 Glottal, *ʔ, ʔb, etc.						
4 Voiced, *b, m, l, z, etc.	[443]	[224]	[221]		[224]	

Summarizing tonal features of the Keng Bay variety, we have six tones in the A, B, and C columns and three tones in the DL and DS columns. In his introduction to the “P Group Languages”, Chamberlain (1984:68) notes a “tone splitting” which “took the form 123-4” in most dialects of that group, and we can observe such a 123-4 split in the columns displaying tone splits (A, B, C, DL). In comments below his tone diagram for Keng Bay Tai Meuay, Chamberlain further notes the “B-DL coalescence” (B = DL, with B123-4 and DL123-4, B123 [554] = DL123 [554] and B4 [224] = DL4 [224]), as well as the “lack of splits in DS column” (DS1234 [34]). The three types of shading displayed in our adaptation of Chamberlain’s tone diagram reflect the latter two sets of features, which will be summarized as B = DL (with B123 = DL123, B4 = DL4) and DS1234. Chamberlain’s diagram for Keng Bay is also characterized by a creakiness in the low-falling tone of C4 [221]. One will eventually note that, although both B4 [224] and C123 [445] have rising tones, the pitch heights of those tones are different, and there is no coalescence.

Figure 3 shows Chamberlain’s tone diagram for the Nava variety, with an interpretation using tone numerals of the tones in that variety.

Figure 3: *The tone diagram for Nava Tai Meuay with tone numerals (following Chamberlain 1984:67).*

Initial\Tone	A	B	C		DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[434]	[443]	[44]		[443]	[33]
2 Voiceless unaspirated stops, *p, etc.						
3 Glottal, *ʔ, ʔb, etc.						
4 Voiced, *b, m, l, z, etc.	[335]		[115]			

The Nava variety has five tones in the A, B, and C columns and two tones in the DL and DS columns. Chamberlain’s comments related to the tone diagram for Keng Bay Tai Meuay apply to the tone diagram for the Nava variety as well. There is the same lack of splits in DS column (DS1234 [33]). However, when it comes to the B-DL coalescence, that feature is not the same for the two varieties: in the Nava variety, neither B nor DL display a split (B1234 [443] = DL1234 [443]). Two types of shading in our adaptation of Chamberlain’s tone diagram for the Nava variety reflect those sets of tonal features.

As commented by Chamberlain (personal communication, March 18, 2021), “there seemed to be two dialects”. The absence of splits characterizing the B and D columns in the tone diagram given for Nava Tai Meuy is noteworthy, because it can also be found in tone diagrams available in the literature for a “Tai Pao” variety (Chamberlain 1984:67) and a “Mène” variety (Chamberlain 1991:108), both spoken in the present-day districts of Khamkeuth and Viengthong of Bolikhamxay Province,²¹ as well as for the “Tay Muong” variety spoken in the Tưng Dương District of Nghệ An Province (Ferlus 2008:310). In addition, tone shapes in the Nava variety tone diagram are rather similar to tone shapes in tone diagrams available for those dialects. However, tone diagrams for the Tai Pao and the Mène of Bolikhamxay and for “Tay Muong” of Nghệ An regularly display a split in the DS column, unlike tone diagrams for the Tai Meuy varieties of Keng Bay and Nava.

3.2 *Kullavanijaya and L-Thongkum’s 1998 article*

As previously mentioned (see section 1), the “Tai Moei” variety discussed in these authors’ article is a variety spoken in Phônscy Village (Khamkeuth). The speaker, a female aged 39, was recorded on the 6th of January 1993 by Thongpheth Kingsada and Michel Ferlus.

Whereas Chamberlain (1984:68) notes the prevalence of a 123-4 split for each column in languages belonging to the P Group, Kullavanijaya and L-Thongkum (1998:284) speak of the 123-4 split for the A and C columns in “Tai Dam, Tai Don, Tai Lue, and Tai Daeng (including Tai Phoeng and Tai Moei)”.²² As a matter of fact, their focus on the A and C columns is relevant when one compares those languages of the P Group with languages such as Lao, which has A1-23-4 and C1-234, or Thai, which has A1-234 and C123-4.

In addition, according to the authors, “Tai Daeng and closely related dialects, Tai Phoeng and Tai Moei, have a further step of tone merger, i.e., B4 merges with C123”. That B4-C123 coalescence is noted for Tai Daeng by Gedney (1989:421, 423, 425) and is always confirmed in interactions with Tai Daeng speakers, whether in Houaphanh Province or in Thanh Hóa Province. As for Chamberlain (1984:69), he speaks of a “C123-B4/DL4 coalescence” for Tai Daeng, but he does not mention it for Tai Meuy, that particular coalescence being absent in the tone diagrams he gives for the Keng Bay and Nava varieties.

The authors do not propose a tone diagram for the “Tai Moei” of Phônscy (Khamkeuth) they refer to in their article. However, their article (Kullavanijaya and L-Thongkum 1998:285–286) contains the following fourteen lexical items used in that variety whose tones are noted with tone numerals: [bɿ²⁴³] ‘leaf’, [bin²⁴³] ‘to fly’, [din²⁴³] ‘soil’, [daŋ²⁴³] ‘nose’, [daau²⁴³] ‘star’, [phom²⁴³] ‘hair’, [bɯən²⁴³] ‘moon’, [sai bu²⁴³] ‘navel’ for the A123 box; [baan⁴⁵] ‘village’, [phum⁴⁵] ‘bee’ for the C123 box; [bək³³] ‘flower’, [ka duk³³] ‘bone’, [pik³³] ‘wing’ for the DL123 box; [naak²⁴] ‘otter’ for the DL4 box.²³ As the lexical items given for the DL123 and the DL4 boxes support a split in the DL column, the tone merger and split pattern in Phônscy (Khamkeuth) could be of the Keng Bay type rather than of the Nava type.

3.3 *A first summary of the criteria provided by the literature*

Chamberlain’s 1984 account and Kullavanijaya and L-Thongkum’s 1998 article provide criteria that help identify Tai Meuy varieties, and a linguistic analysis will thus check the following characteristics in each variety.

²¹ While the Mène and Tai Pao varieties can be said to be spoken in Khamkeuth and Viengthong, Mène is definitely more widespread than Tai Pao.

²² “Tai Phoeng, a branch of Tai Daeng, can be found in Muong Kham, Xiangkuang province” (Kullavanijaya and L-Thongkum 1998: 285).

²³ The main purpose of the authors when they list such lexical items is to show that, although the dialects they identify as “Tai Phoeng” and “Tai Moei” appear to be “closely related” to Tai Daeng, “the different patterns of consonant changes can be used as criteria for separating Tai Phoeng and Tai Moei from Proper Tai Daeng and from each other” (Kullavanijaya and L-Thongkum 1998: 285). For example, “Tai Moei” lexical items such as [bɿ²⁴³] ‘leaf’, [bɯən²⁴³] ‘moon’, [din²⁴³] ‘soil’, [phom²⁴³] ‘hair’, etc., whose respective initials are reconstructed as *ɣb-, *ɣbl-, *ɣd-, and *ph- in Proto-Southwestern Tai, do not display the reflexes [v-] (for both *ɣb- and *ɣbl-), [l-] (for *ɣd-), and [f-] (for *ph-), which characterize many varieties of Tai Daeng.

- Phonological characteristics (e.g., the reflex of the initial *ɣ- and the reflex of the rhyme *-au);
- Lexical characteristics (e.g., Tai Meuay’s specific lexical items for ‘to go’, ‘to eat’, or ‘to do’);
- Tonal characteristics (e.g., 123-4 split in columns displaying tone splits, especially in the A and C columns; B-DL coalescence; lack of splits in DS column; B4-C123 coalescence).

For two of the tonal characteristics which have just been mentioned (B-DL coalescence, B4-C123 coalescence), literature provides a counterexample with Onlao’s 2010 study.

3.4 Onlao’s 2010 dissertation

The Tai Meuay data in Onlao’s study were elicited from a female speaker, aged 65, living in Namphao, a village on the outskirts of Lakxao City, where a bridge on the road from Lakxao to Nakai crosses the Namphao River. According to the information gathered in the area during this research, Namphao cannot be considered as a Tai Meuay location. However, as many Tai Meuay families have come to live in Lakxao City and the villages which belong to it (e.g., Nongpong and Sômsanouk), a few Tai Meuay speakers can be found in Namphao.

The description of the tones which the author proposes for the Namphao variety is based on measurements of their fundamental frequency (F0) in the Hertz (Hz) scale, which she summarizes by giving a figure displaying the F0 curves of the tones in the appendix of her dissertation (Onlao 2010a:189). Figure 4 shows a tone diagram for Namphao, with an interpretation, using tone numerals, of the tones in that variety.

Figure 4: *The tone diagram for Namphao Tai Meuay with tone numerals based on the F0 measurements of the tones, as given by Onlao (2010a:189).*

Initial\Tone	A	B	C		DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[²²⁴]	[⁴⁴²]	[⁴⁴⁵]		[³²⁴]	[³⁵]
2 Voiceless unaspirated stops, *p, etc.						
3 Glottal, *ʔ, ʔb, etc.						
4 Voiced, *b, m, l, z, etc.	[¹²¹]	[¹¹²]				

The Namphao variety has five tones in the A, B, and C columns and two tones in the DL and DS columns. While a 123-4 split can be observed in A, B, and C, there are no splits in either DL or DS (DL1234 [324], DS1234 [35]). As we have a split in B (B123-4) and no split in DL (DL1234), there is no B-DL coalescence in this tone diagram (B ≠ DL). In addition, the tone shape in DL1234 [324] cannot be related to either the tone shape in B123 [442] or the tone shape in B4 [112]. When it comes to the tones in the B4 and C123 boxes, although the tones in B4 [112] and C123 [445] are rising tones, their tone numerals indicate quite different pitch heights, and there is no coalescence. A distinctive feature of this tone diagram, according to Onlao (2010a:86), is the B4-C4 coalescence (B4 [112] = C4 [112]). Could that B4-C4 coalescence noted for Namphao Tai Meuay possibly be used as a criterion when dealing with other varieties? In the tone diagram given for the Namphao variety, the sets of tonal features given as B4 = C4 and DS1234 are reflected by two types of shading.

4 The investigation and the Tai Meuay language data of this research

From October 2010 to May 2012, the author of this article was based in Kaysone Phomvihane City (Savannakhet Province, Laos) and had the opportunity to work on students’ linguistic autobiographies together with a Lao instructor teaching French at Savannakhet University. Although the final article

dealing with that study focuses on Phu Thai students of the provinces of Khammouane and Savannakhet (Pacquement and Phongphanith 2012), four Tai Meuy students belonging to Bolikhamxay Province – one from Khamkeuth District, two from Viengthong District, and one from Pakkading District – participated at an earlier stage of the research and provided data. The student from Khamkeuth District provided a great deal of additional information about her village, Vangko, other villages in the surroundings, and the Tai Meuy language spoken there during in-depth interviews conducted in May 2012 and further exchanges which took place later in December 2015.

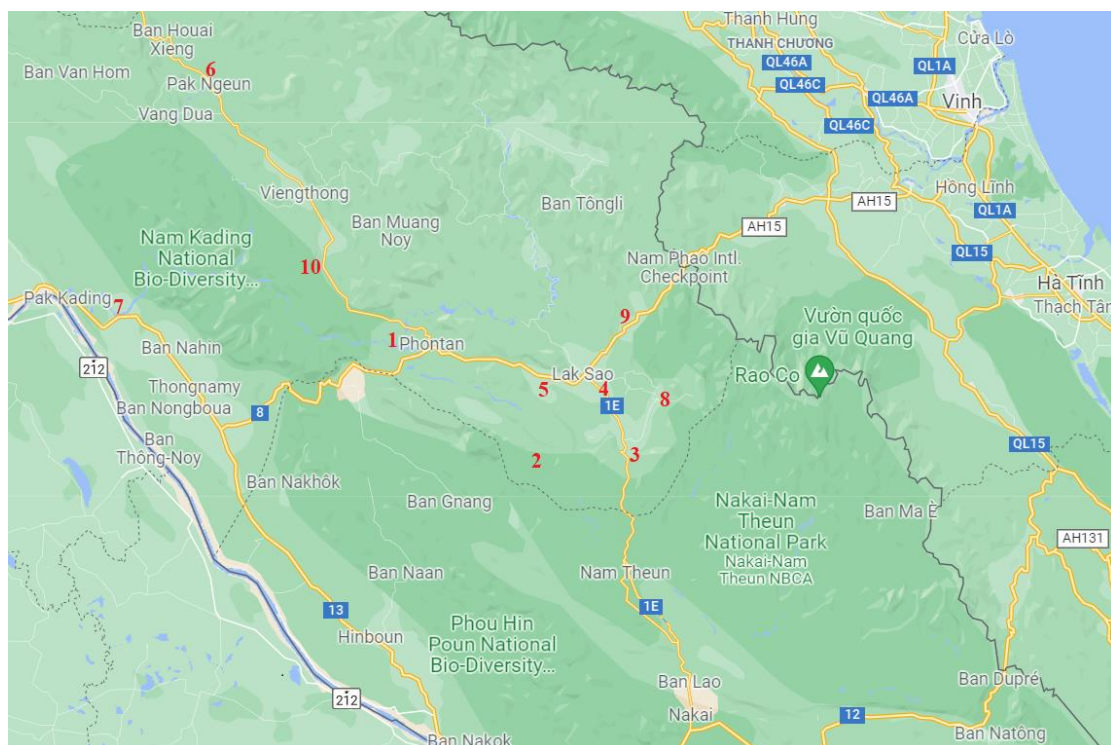
The author of this article eventually visited the region of Lakxao in July 2016 and could go and stay there again in November 2016, November 2017, and March 2018. He also visited a few locations in Viengthong District (November 2017) and Pakkading District (January 2020). Whereas in Khamkeuth District, as well as in some parts of Viengthong District, Tai Meuy is, as noted by Chamberlain (1984:68), “quite widespread”, in Pakkading District, Tai Meuy is spoken in a few villages only.

This section discusses the data collected for this research. That data collection consisted of informal interviews, in which various topics, such as body parts, colors, kinship, nature, weather, as well as food gathering and cooking processes, were discussed. One purpose of those exchanges, which were recorded, was to elicit representative monosyllabic words for each of the twenty boxes in Gedney’s tone diagram. Whenever such words appeared during the interviews, speakers were sooner or later invited to pronounce them in citation style at least two times.

The data analysis showed both a language uniformity with respect to phonological and lexical features and a language diversity with respect to tonal features. The focus in this section is on tonal characteristics. The PRAAT program (Boersma and Weenink 2021) was used to obtain measurements of fundamental frequency for each tone in the monosyllabic words elicited from speakers. Auditory judgements by some speakers, those whom we could meet again to discuss the data after the tone analysis, were also taken into account.

Map 1 shows both the locations whose varieties are dealt with in the literature (see section 3), i.e., Keng Bay (1), Nava (2), Phônxy (Khamkeuth) (3), Namphao (4), and the locations for whose varieties this article proposes tone diagrams, i.e., Phayat (5), Chomthong (6), Phônxy (Pakkading) (7), Thônkghe (8), Nong-o (9), Hinngôn (10).

Map 1: From google maps (<https://www.google.com/maps>) @2021 Google.



4.1 Phayat (Khamkeuth)

The interview in Phayat Village, which took place in July 2016, was with a female speaker in her forties. The tone diagram for the variety she speaks is given in Figure 5. That speaker was interviewed again in March 2018.

Figure 5: *The tone diagram for Phayat Tai Meuy.*

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[³¹³]	[³⁴²]	[²⁴⁵]	[³⁴³]	[²⁵]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, ʔb, etc.					
4 Voiced, *b, m, l, z, etc.	[³⁴⁴]	[¹²⁵]		[¹²⁴]	

The Phayat variety has five tones in the A, B, and C columns and three tones in the DL and DS columns. The 123-4 split is found in the columns displaying tone splits. The DS column, which has no split, has a rising tone (DS1234 [²⁵]). On the basis of the measurements of the fundamental frequency obtained for the tones in the B123 and DL123 boxes, and in the B4 and DL4 boxes, there is a B-DL quasi-coalescence: B ≈ DL, with B123 [³⁴²] ≈ DL123 [³⁴³], and B4 [¹²⁵] ≈ DL4 [¹²⁴]. In addition, B4 merges with C4 coalescence (B4 [¹²⁵] = C4 [¹²⁵]). The tones in B4 and C123 are both rising tones (B4 [¹²⁵] and C123 [²⁴⁵]), but their contours are different, and there is no coalescence. The last two tonal features (B4 = C4 and B4 ≠ C123) were confirmed by the speaker’s auditory judgement. Various types of shading in the tone diagram for Phayat reflect the following sets of tonal features: B ≈ DL (with B123 ≈ DL123, B4 ≈ DL4), B4 = C4, and DS1234.

4.2 Chomthong (Viengthong)

The interview in Chomthong, nowadays the headquarters of the subdistrict bearing the same name in Viengthong District, took place in November 2017 with a female speaker in her forties. Figure 6 shows the tone diagram for the variety she speaks.

Figure 6: *The tone diagram for Chomthong Tai Meuy.*

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[³¹³]	[⁵⁴¹]	[⁵⁵]	[⁵⁴¹]	[²²]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, ʔb, etc.					
4 Voiced, *b, m, l, z, etc.	[³²]	[³³⁴]	[³³⁵]	[³³⁴]	

The tone numbers in the B4 and C4 boxes (B4 [³³⁴] ≈ C4 [³³⁵]) and the F0 measurements represented in [³³⁴] and [³³⁵] being close to each other, the 6 tones in the A, B, and C columns turn out to be five; and there are three tones in the DL and DS columns. Then, the tone splits and mergers in the tone

diagram for Chomthong can be related to those displayed in the tone diagram for Phayat. Tone shapes, however, are different. For example, the tone in the DS column is a low-level one (DS1234 [22]). When considering the B-DL coalescence, we have B123 [541] = DL123 [541], and B4 [334] = DL4 [334] for Chomthong, as opposed to B123 [342] ≈ DL123 [343], and B4 [125] ≈ DL4 [124] for Phayat. As for the tones in B4 and C123, they have different contours (B4 [334] ≠ C123 [55]) in Chomthong, and there is no coalescence. The various types of shading displayed in the tone diagram for Chomthong reflect the sets of tonal features summarized as B = DL (B123 = DL123, B4 = DL4), B4 ≈ C4, DS1234.

4.3 Phônscy (Pakkading)

With respect to Phônscy (Pakkading), this article takes into account the varieties spoken by two speakers. The first speaker who provided data is a former student of the National University of Laos in her twenties. She left her native place, Phônscy (Pakkading), at the age of 10, and then lived in another part of Bolikhamxay Province (Thaphabat District), where Lao varieties are spoken, but her family continued to speak the Tai Meuy of Phônscy (Pakkading). She was interviewed in August 2018 at Nongkhai (Thailand), where some of her relatives run a restaurant. The tone diagram for the variety she speaks is given in Figure 7.

Figure 7: The first tone diagram for Phônscy (Pakkading) Tai Meuy.

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[⁵ 214]	[⁵ 53]	[⁴ 45]	[⁴ 41]	[³ 3]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, etc.					
4 Voiced, *b, m, l, z, etc.	[⁴ 33]	[³ 35]		[³ 25]	

This first tone diagram for Phônscy (Pakkading) has five tones in the A, B, and C columns and three tones in the DL and DS columns. Although the concerned speaker has lived in Lao-speaking environments since the age of 10, her tone diagram displays a 123-4 split in the A, B, C, and DL columns, the columns which have tone splits. Lao being characterized by quite different splits in the A and C columns (A1-23-4, C1-234), the language she spoke during the interview was definitely not Lao. As the tone diagram for her variety can be related to the tone diagrams for the Tai Meuy varieties of Phayat and Chomthong discussed above, it can be inferred that she still speaks Tai Meuy. The tone in the DS column is a mid-level one (DS1234 [33]). When it comes to the B and DL columns, which both display a 123-4 split (B123-4 and DL123-4), measurements of the fundamental frequency for the tones in the B and DL columns support a B-DL quasi-coalescence (B ≈ DL): the tones in the B123 and DL123 boxes are both falling tones, albeit with different tone numbers (B123 [553], DL123 [441]); as for the tones in the B4 and DL4 boxes, they can be more easily related to each other (B4 [335] ≈ DL4 [325]). As in Phayat, B4 merges with C4 (B4 [335] = C4 [335]). One will further note that, in this first tone diagram for Phônscy (Pakkading), the tone in the B4 box, a mid-rising tone, can be compared with the tone in the C123 box (B4 [335], C123 [445]). However, we could not meet that speaker again to determine whether there could be a merger. In order to clearly show the relationship of this first tone diagram for Phônscy (Pakkading) to the tone diagrams for Phayat and Chomthong, the types of shading it displays reflect only the following three sets of tonal features: B ≈ DL (B123 ≈ DL123, B4 ≈ DL4), B4 = C4, and DS1234. As for the rather complex tone in the A123 box (A123 [5214]), which is characterized by a voice quality feature, pressed voice, in the third quarter of its duration, it actually brings us to the second tone diagram for Phônscy (Pakkading).

The second speaker who provided data for the variety spoken in Phônscy (Pakkading) variety is a female speaker in her fifties whose family seems not to be related to that of the first speaker. The interview took place in January 2020 in the village of Phônscy (Pakkading) itself. Figure 8 shows the tone diagram for the variety she speaks.

Figure 8: *The second tone diagram for Phônscy (Pakkading) Tai Meuay.*

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[⁵⁴ 13]	[⁵⁴]	[³⁵]	[⁵³]	[²¹²]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, ʔb, etc.					
4 Voiced, *b, m, l, z, etc.	[²²¹]	[²¹⁴]	[¹¹²]	[²¹³]	

This second tone diagram for Phônscy (Pakkading) has six tones in the A, B, and C columns and three tones in the DL and DS columns. The tone in the DS column is a low-falling and rising tone (DS1234 [212]). There is a B-DL quasi-coalescence ($B \approx DL$), both columns displaying the same split (B123-4 and DL123-4), and, whether one considers the tones in the B123 and DL123 boxes or those in the B4 and DL4 boxes, the two tones in each pair can be related to each other (B123 [54] \approx DL123 [53], B4 [214] \approx DL4 [213]). This tone diagram lacks the B4-C4 coalescence (B4 [214] \neq C4 [112]). Furthermore, unlike the previous tone diagram for the same location, the two rising tones in the B4 and C123 boxes have different contours (B4 [214] \neq C123 [35]), and there seems to be no coalescence. The tone mergers and splits in this second tone diagram for Phônscy (Pakkading) thus appear to be of the same type as the ones in Chamberlain’s tone diagram for Keng Bay Tai Meuay. Accordingly, the various types of shading it displays reflect the sets of tonal features given as $B \approx DL$ (with B123 \approx DL123, B4 \approx DL4) and DS1234. When it comes to the tone in the A123 box (A123 [5413]), this second speaker is found to pronounce it with the same voice quality feature as the previous speaker. That voice quality feature, pressed voice, occurring in both cases in the third quarter of the tone duration, appears to be a distinctive characteristic of the variety spoken in Phônscy (Pakkading).

4.4 *Thôngkhe*

A more intriguing tone diagram is the one for *Thôngkhe* Tai Meuay in Figure 9. The interview, which took place in July 2016, was with a male speaker in his fifties. That speaker was interviewed again in November 2017 and March 2018.

The *Thôngkhe* variety has nine tones in the A, B, and C columns and three tones in the DL and DS columns. As expected, there is only one tone in each of the A4, B123, B4, C123, DL123, DL4, and DS1234 boxes. However, the A123 box contains three tones, and the C4 box two tones. The three tones in the A123 box can be found in A1, A2, and A3 words, as well as with all the possible initial consonants of A1, A2, and A3 words. Each of the two tones in the C4 box can be found with the initials associated with C4 words.

Figure 9: *The tone diagram for Thônkghe Tai Meuy.*

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[⁵¹] [⁴²³] [¹¹³]	[³⁴³]	[³⁵]	[³⁴³]	[³⁴]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, etc.					
4 Voiced, *b, m, l, z, etc.	[²³³]	[²²⁵]	[^{41?}] [⁴²⁵]	[²³⁵]	

With respect to the tone mergers and splits, the tone diagram for Thônkghe displays the same pattern as the tone diagram proposed by Chamberlain for Keng Bay Tai Meuy. The DS column having no splits, its tone is a rising one (DS1234 [34]), as in the Keng Bay and Phayat varieties. There is a B-DL quasi-coalescence ($B \approx DL$), for which we have a B123-DL123 coalescence (B123 [343] = DL123 [343]), and a B4-C4 quasi-coalescence ($B4 [225] \approx DL4 [235]$). While the tone numbers in the B4 and C123 boxes (B4 [225], C123 [35]) could have supported a kind of coalescence, the speaker's auditory judgement did not support it. When he was invited to listen to B4 and C123 words, such as [ka:^{B4}] 'cost, fee' and [ka:^{C2}] 'young plant', [na:^{B4}] 'paternal grandmother' and [na:^{C1}] 'grass', [mɿ:^{B4}] 'to be tired' and [mɿ:^{C1}] 'Meuy',²⁴ which he himself had pronounced several times in citation form in previous interviews, he always identified them correctly. As for the tones in the B4 and C4 boxes, they have quite different contours. The various types of shading displayed in the tone diagram for Thônkghe Tai Meuy reflect the tonal features given as $B = DL$ (with B123 = DL123, $B4 \approx DL4$) and DS1234.

In order to explain the three tones ([51], [423], and [113]) in the A123 box,²⁵ we propose to start from the tone identified by Kullavanijaya and L-Thongkum (1998:285–286) for the same box (A123 [243]). Consisting of a rising part and a falling part, that tone's contour is frequent in many Tai Daeng varieties for words in the A123 box.

- In the speech of the speaker interviewed in Thônkghe, the rising part being either very short or absent, the falling tone, high falling in his speech [51], makes sense.
- In what might be a secondary development, the second half of that falling tone can become slightly rising, hence the second tone shape [423], also found in the A123 box of Chamberlain's tone diagram for Keng Bay Tai Meuy.
- Although a tone in the A123 box consisting only of the rising part of the tone shape [243] could not be found in the Tai Meuy varieties which were investigated, there is such a rising tone in the A123 box of the tone diagram we give for a Tai Thaeng variety that could be investigated during this research (see Figure 10). However, the rising tone in the A123 box of the variety spoken by the speaker of Thônkghe (A123 [113]) is much lower than the one we have in the Tai Thaeng variety (A123 [245]).

The case of the C4 box is a case for which one cannot focus on F0 measurements only. The two tone shapes we find in the C4 box ([41?] and [425]) can be found in some Tai Daeng dialects as well. Many

²⁴ As the concerned speaker pronounces the ethnonym 'Meuy' either [muaj^{C1}] or [mɿ:^{C1}] – whether he uses that word in natural speech or pronounces it in citation form –, the parts of the recordings he was invited to listen to in order to identify [mɿ:^{B4}] 'to be tired' and [mɿ:^{C1}] 'Meuy' were only the parts in which he was pronouncing [mɿ:^{C1}].

²⁵ These three tones in the A123 box are not specific to the speech of that speaker only and actually concern other speakers of all generations in that particular area, including the female student from Vangko mentioned at the beginning of section 4.

Tai Daeng varieties spoken in Houaphanh and Thanh Hóa have a falling tone in the C4 box, and that tone can be found to be slightly glottalized. For Tai Meuyay, Chamberlain gives a low-falling tone in the C4 box of his tone diagram for the Keng Bay type (C4 [221]), its second half being characterized by a creakiness. In the variety spoken by the speaker of Thônkhe, we have a high falling tone with a slight glottal constriction [41ʔ]. When the glottal constriction is absent, the second half of the tone becomes rising, hence the contour in the falling-rising tone [425].

Focusing on the two falling tones in the C4 boxes of the tone diagrams for the Keng Bay (C4 [221]) and Thônkhe varieties (C4 [41ʔ]), one will note that such a tone shape in C4, which cannot be found in the tone diagrams for other Tai Meuyay varieties, appears to be specific to the Tai Meuyay varieties of Thônkhe and Keng Bay. However, we find a similar contour for the tone in the C4 box, with the same tone numbers as those in our interpretation for Keng Bay Tai Meuyay (see Figure 2), in the Tai Thaeng variety investigated in Nong-o Village,²⁶ whose tone diagram is given in Figure 10 for comparison. The Tai Thaeng data on which it is based were elicited from a male speaker in his forties during an interview which took place in November 2017.

Figure 10: *The tone diagram for Nong-o Tai Thaeng.*

Initial\Tone	A	B	C	DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[²⁴⁵]	[²³]	[¹²⁴]	[³⁴]	[²⁴]
2 Voiceless unaspirated stops, *p, etc.					
3 Glottal, *ʔ, ʔb, etc.					
4 Voiced, *b, m, l, z, etc.	[²²]	[¹²⁴]	[²²¹]	[¹²⁴]	

This tone diagram will not be discussed in detail. We will simply note that, with a B-DL quasi-coalescence and the lack of splits in the DS column, the tone mergers and splits in this particular Tai Thaeng variety have the same pattern as those in the Tai Meuyay varieties spoken in Keng Bay and Thônkhe. However, such a proximity can only highlight the main difference between the former (Tai Thaeng) and the latter two (Tai Meuyay) varieties: the tone diagram available to us for Thai Thaeng is precisely characterized by a B4-C123-DL4 coalescence, supported by tone numbers and the F0 measurements they represent. When it comes to Tai Meuyay varieties, the use of the PRAAT program and speakers' auditory judgements could not support an indisputable B4-C123 merger in any Tai Meuyay variety. In the tone diagram for Nong-o Tai Thaeng, the sets of tonal features given as $B \approx DL$ ($B123 \approx DL123$, $B4 = DL4$), $B4 = C123 = DL4$, and $DS1234$ are reflected by three types of shading.

4.5 Hinngôn

The last tone diagram discussed in this article is the one for Hinngôn, a village belonging to Naxouang Subdistrict in Viengthong District.²⁷ The interview in Hinngôn took place in November 2017 with a female speaker in her forties. Figure 11 shows the tone diagram for the variety she speaks.

²⁶ The ethnonym and language name 'Thaeng' ([tʰe:ŋ^{A1}]) is spelled 'Thanh' in accounts dealing with ethnic groups and Tai dialects in Vietnam (see section 1). When it comes to Khamkeuth District, Tai Thaeng, a dialect of the P group, is mainly spoken in a few villages around Nape and Nong-o. Outside that area, Tai Thaeng is also spoken near Thônkhe in Nathone Village.

²⁷ From Khamkeuth District to Xiengkhouang Province, Naxouang is the first subdistrict of Viengthong District.

Figure 11: *The tone diagram for Hinngôn Tai Meuy.*

Initial\Tone	A	B	C		DL	DS
1 Voiceless friction sounds, *s, hm, ph, etc.	[²²⁵]	[⁴¹]	[³⁴⁴]		[⁴⁴²]	[³⁴]
2 Voiceless unaspirated stops, *p, etc.						
3 Glottal, *ʔ, ʔb, etc.						
4 Voiced, *b, m, l, z, etc.	[⁴⁴]	[⁴²]				

The Hinngôn variety has five tones in the A, B, and C columns and two tones in the DL and DS columns. The main difficulty with this tone diagram lies in the B and DL columns. On the basis of the F0 measurements for the tones in those two columns, we obtain a first tone for B1, B2, B3, a second one for B4, and a third one for DL1, DL2, DL3, DL4. Their tone numbers being [41], [42], [442] respectively, could there be some kind of coalescence?

As the author of this article could not visit Hinngôn again to meet the concerned speaker and further assess the relationship between the three tones in the B123, B4, and DL1234 boxes, we must first take a close look at the F0 measurements which the tone numbers given in the tone diagram represent.

- The tone numbers [41] for the B123 box represent a tone whose starting point is 265Hz.²⁸ During its first quarter, the F0 curve rises to reach a change point (284Hz). Then, from the second quarter to the last quarter, it gradually falls to 197Hz.
- The tone numbers [42] for the B4 box represent a tone with a similar contour. The F0 curve first rises from 266Hz to 277Hz, the latter value of the F0 being reached in the middle of the second quarter. Then it falls to a comparatively higher end point (234Hz).
- The tone numbers [442] in the DL1234 box represent a tone whose F0 curve rises from 276Hz to 294Hz during its first half. Its change point is in the exact middle of the F0 curve, and it then gradually falls during its second half to 249Hz.

Taking into account the significant difference between the falling part of the tone in the B123 box (with a F0 falling from 284Hz to 197Hz) and that of the tone in B4 (with a F0 falling from 277Hz to 234Hz), one can reasonably conclude that there is no coalescence between these two tones. In addition, with a split in B (B123-4) and no split in DL (DL1234), there is definitely no B-DL coalescence (B ≠ DL). One will also note the following tonal features: B4-C4 coalescence (B4 [42] = C4 [42]), no B4-C123 coalescence (B4 [42] ≠ C123 [344]), and lack of splits in DS column (DS1234 [34]). On the basis of the tone mergers and splits, the tone diagram for Hinngôn can thus be compared with the one which was created for the Namphao variety studied by Onlao (2010a).

However, when one takes into account the tone contours in the B123, B4, DL1234 boxes of both tone diagrams, such a comparison based only on tone mergers and splits seems to have its limits.

- In Sasithorn Onlao's study, the three tones in the B123, B4, and DL1234 boxes have completely different contours (B123 [442], B4 [112], DL1234 [324]) and cannot be related to one another.
- With respect to the tone diagram for Hinngôn, although F0 measurements do not support a coalescence between the tones in the B123, B4, and DL1234 boxes, especially between the tone in the B123 box and the tone in the B4 box, their tone shapes are comparable, because all these

²⁸ As for other varieties, Yuen Ren Chao's five-point pitch scale was applied to all the F0 measurements which were obtained for the tones of that speaker, those measurements ranging from 197Hz (end point of the tone in the B123 box [41]) to 330 Hz (end point of the tone in the A123 box [225]).

three tones are falling or at least have a significant falling part (B123 [41], in B4 [42], and in DL1234 [442]).

With such similarities of the tones in the B123, B4, and DL1234 boxes, the tone diagram for Hinngôn appears to have some relation to the tone diagram proposed by Chamberlain for Nava. It should nevertheless be borne in mind that the Nava tone diagram shows a B-DL coalescence, with no splits in the B and DL columns, and does not have a B4-C4 coalescence.

Accordingly, the various types of shading displayed in the tone diagram given for Hinngôn Tai Meuy reflect the sets of tonal features given as falling tones with different tone numbers in B123, B4, and DL1234 (three shades of the same color), B4 = DL4, and DS1234.

5 The speakers' perspective: ethnographic and sociolinguistic information

This research was also an opportunity to listen to speakers' views on their ethnic group and language.

5.1 Ethnographic information: 'Tai Meuy Katip Nyeu' and 'Tai Meuy Katip Noi'

Tai Meuy speakers in Khamkeuth District, especially those in the area surrounding Lakxao City, identify two main groups of Tai Meuy: the 'Tai Meuy Katip Nyeu' – [taj^{A4} muaj^{C1} kaʔ ti:p^{DL2} ɲɯ:^{B1}] – and the 'Tai Meuy Katip Noi' – [taj^{A4} muaj^{C1} kaʔ ti:p^{DL2} nɔ:j^{C4}]. These designations refer to the size ([ɲɯ:^{B1}] 'big, large', [nɔ:j^{C4}] 'small') of a basket called [kaʔ ti:p^{DL2}], which Tai Meuy women carry on their backs using a forehead band.²⁹

With respect to the speakers whose varieties were taken into account for the tone analysis proposed in the preceding section, all of them, but the one interviewed in Chomthong (Viengthong District),³⁰ clearly stated that they belong to one of these two groups.

- The male speaker interviewed in Thônkghe (Khamkeuth District) is a Tai Meuy Katip Nyeu. More generally, in the surroundings of Thônkghe, Tai Meuy are Tai Meuy Katip Nyeu.
- The female speaker interviewed in Hinngôn (Viengthong District) stated that her family and her husband's family are Tai Meuy Katip Nyeu, and that there are Tai Meuy Katip Nyeu in the surrounding area.
- The female speaker interviewed in Phayat (Khamkeuth District) is a Tai Meuy Katip Noi. The Tai Meuy in Lakxao City, as well as in the area of Thônkghe, say that Phayat Village is the only Tai Meuy Katip Noi location in Khamkeuth District.
- The two female speakers interviewed for the variety spoken in Phônxy (Pakkading) are Tai Meuy Katip Noi. Other villages near Phônxy (Pakkading), such as Phônxyay (Pakkading), Nakhuanai and Nakhuanok, also have Tai Meuy speakers. However, only a few speakers in that area say they are Tai Meuy Katip Noi. Others know and use only the designation 'Tai Meuy'.

This research did not investigate possible linguistic differences between Tai Meuy Katip Nyeu varieties on the one hand and Tai Meuy Katip Noi varieties on the other. The fact that the tone diagram for Hinngôn is very different from the one for Thônkghe, with speakers being Tai Meuy Katip Nyeu in both locations, suggests that it may be actually difficult to identify common features for at least the Tai Meuy Katip Nyeu varieties. We will simply say that the varieties spoken by the two groups are mutually intelligible. As for the differences mentioned by individual speakers, they are related to their auditory perception of a few segmental features (e.g., length of syllables) and prosodic features (e.g., sentence intonation) rather than to the use of specific lexical items.

²⁹ The Tai Daeng have the same basket and refer to it in the same way. A similar basket, called [kaʔ dɯ:p^{DL3}], is used by both the Tai Meuy and the Tai Daeng, but the weaving is slightly different, and the [kaʔ dɯ:p^{DL3}] is generally of a larger size than the [kaʔ ti:p^{DL2}].

³⁰ Talking about Chomthong and the Tai Meuy of that area, that speaker did not refer to the Tai Meuy Katip Nyeu and the Tai Meuy Katip Noi. When asked specifically about those two designations, she said that she had never heard them.

5.2 Sociolinguistic information: ‘Tai Meuy Đông’ and ‘Tai Meuy Lao’

When the author of this article visited the village of Phônscy (Khamkeuth) in November 2017, Tai Meuy speakers there, who are Tai Meuy Katip Nyeu as in Thônkghe Village, explained that they identify two ways of speaking Tai Meuy.

- The first way is called ‘Tai Meuy Đông’, the term [doŋ^{A3}] referring to forest areas. It is the way Tai Meuy speak their dialect in their original locations, which are supposed to be in remote areas.
- The second way is called ‘Tai Meuy Lao’. It is the way Tai Meuy Katip Noi and some Tai Meuy Katip Nyeu speak their language when they live with Lao people.

The distinction between these two ways of speaking Tai Meuy was later confirmed by speakers belonging to other villages, such as Thônkghe and Vangko.

According to Tai Meuy speakers in Phônscy (Khamkeuth), Thônkghe and Vangko, Tai Meuy Lao is spoken by all the Tai Meuy Katip Noi, such as those of Phayat Village in Khamkeuth District and those living in Pakkading District. It is also spoken by Tai Meuy Katip Nyeu in villages located on the main roads, such as Phônxy (Khamkeuth), a Tai Meuy village on the main road from Lakxao to Nakai where Tai Meuy Katip Nyeu are said to speak Tai Meuy Lao.³¹

While only some Tai Meuy Katip Nyeu actually speak Tai Meuy Đông, many of them live in the area comprising the villages of Phônscy (Khamkeuth), Thônkghe, and Vangko. Tai Meuy Katip Nyeu speaking Tai Meuy Đông in that area say that they originate from a village called Namuong.³² It will be noted that, although villages such as Phônscy (Khamkeuth), Thônkghe, Vangko, and others are not on the main road between Lakxao and Nakai, they are rather recent locations.³³

With respect to the two ways of speaking Tai Meuy referred to as Tai Meuy Đông and Tai Meuy Lao, their difference is mainly related to word choice. One aspect of the difference between the Tai Meuy Đông and Tai Meuy Lao ways of speaking involves the use of pronouns. Tai Meuy varieties have the following first- and second-person pronouns: [ʔem^{C3}] and [ka:^{A4}] ‘I’, [thaw^{A4}] and [muŋ^{A4}] ‘you’.

- A speaker uses [ʔem^{C3}] ‘I’ to mean that he or she does not enjoy a high status. That speaker will then use [muŋ^{A4}] ‘you’ with a person of a higher status and [thaw^{A4}] with a person of equal or lower status.
- A speaker uses [ka:^{A4}] ‘I’ when he or she wants to stress his or her higher status. That speaker will then use [muŋ^{A4}] ‘you’ with a person of the same status and [thaw^{A4}] with a person of lower status.

While people speaking in the Tai Meuy Đông way use all those four pronouns, those who speak the Tai Meuy Lao way still use the four pronouns, but they tend to avoid [ka:^{A4}] ‘I’, [thaw^{A4}] and [muŋ^{A4}] ‘you’, especially with outsiders.

Another aspect of the difference between the Tai Meuy Đông and Tai Meuy Lao ways of speaking concerns the use of a few lexical items. Here are two examples.

- For ‘rain’, we have [fɾn^{A1}] in Tai Meuy Đông (as in Tai Daeng) and [fon^{A1}] in Tai Meuy Lao (as in Lao).

³¹ Tai Meuy Lao should not be considered as a recent development of the Tai Meuy language: a speaker of Phônscy (Khamkeuth) in his fifties recalled he was born in another village, Nakhua, now an abandoned location, where the way of speaking was already identified as Tai Meuy Lao forty years ago.

³² Namuong is located on a gravel path which starts in Wangko and ends in the vicinity of Nape. From Namuong, there is a pathway to Vietnam, and walking to Vietnam is said to take three hours. However, the only official checkpoint in the area is the one on the road from Lakxao to the border via Nape.

³³ The Tai Meuy Katip Nyeu speaking Tai Meuy Đông who came to live in Phônscy (Khamkeuth) live with some Tai Meuy Katip Nyeu speaking Tai Meuy Lao, themselves from other villages. As for the Tai Meuy Katip Nyeu speaking Tai Meuy Đông who came to live in villages such as Thônkghe and Vangko, they live with other Tai groups, the most prominent one being the Tai Bo (Chamberlain 1984:81–82; 1996:11–12).

- For ‘calf (of the leg)’, we have [pi:^{A2} khe:ŋ^{B4}] in Tai Meuy Đông, the form [pi:^{A2}] being specific to Tai Meuy, and [bi:^{A3} khe:ŋ^{B4}] in Tai Meuy Lao (as in Lao).

6 Conclusion: the position of Tai Meuy in the Southwestern Tai branch of the Tai language family

In presenting the Tai Meuy language data from this research, this article has focused on tonal characteristics. Although the tonal features discussed in section 4 suggest a linguistic diversity of Tai Meuy in the area considered for this study, all the tone diagrams proposed for Tai Meuy varieties display both a 123-4 split in columns displaying tone splits and an absence of splits in their DS columns, giving a first glimpse of a Tai Meuy language uniformity. With regard to the phonological and lexical characteristics identified in the literature, they were found in all the locations which were investigated without significant variation, suggesting a more consistent Tai Meuy language uniformity. While the previous sections of this article left out phonological and lexical features, they will now be taken into account in this conclusion, which focuses on the position of Tai Meuy in the Southwestern Tai branch of the Tai language family.

In Chamberlain’s 1984 account, as well as in Kullavanijaya and L-Thongkum’s 1998 article, Tai Meuy is posited as a dialect of the P group related to Tai Daeng. The data elicited for this research confirm that Tai Meuy varieties are to be related to Tai Daeng. The main arguments for that placement are the following: the absence of splits in the DS column; the reflex [k^h-] of the initial *ɣ-; the reflex of the rhyme reconstructed as *-au.

- Tone systems of all the Tai Meuy varieties dealt with in the literature (section 3) as well as in this research (section 4) exhibit a tonal coalescence in the DS column. According to Chamberlain (1984:67), “lack of splits in DS column seems to be areal feature of the Hua Phan dialects”. As Tai Daeng dialects in Thanh Hóa Province (data elicited by the author of this article) are also characterized by a tonal coalescence in the DS column, it is reasonable to assume that it is a Tai Daeng feature.
- When it comes to initial consonants, Chamberlain (1984:66) notes the reflexes [k-] and [tɛ-] in Tai Meuy for words reconstructed with the initials *g- and *dz- respectively. Tai Meuy varieties share those reflexes with other languages belonging to the P group of the Southwestern Tai branch in the Tai family. However, in the P group, the reflex [k^h-] in words reconstructed with the initial *ɣ- characterizes Tai Daeng, such as [k^hɛ:ŋ^{B4}] ‘lower leg, shin’, in contrast with [kɛ:ŋ^{B4}] in Tai Yo and Tai Dam. The Tai Daeng reflex of the initial *ɣ- is found in all the Tai Meuy varieties studied in this research.³⁴
- When it comes to the rhyme reconstructed as *-au, Chamberlain (1984:66) notes that the Tai Meuy reflex is [ɾ:]. As Tai Daeng and Tai Yo have the same reflex [ɾ:] for the rhyme *-au, it helps separate Tai Daeng, Tai Yo, and dialects related to both languages, from languages such as Tai Dam, in which the rhyme is [-au]. In connection with the rhyme *-au, the cognate of the Tai Dam interrogative [dau^{A3}] ‘which’ is [ɾ:^{A4}] in both Tai Daeng and Tai Yo. All the Tai Meuy varieties mentioned in this article have the reflex [ɾ:] for the rhyme *-au and use the interrogative [ɾ:^{A4}] of Tai Daeng and Tai Yo.

Tai Meuy has been related to Tai Yo, as in the Pangloss Collection (<https://pangloss.cnrs.fr/>), and to Tai Dam, as in *Glottolog 4.3* (<https://glottolog.org/>). However, neither the Pangloss Collection nor *Glottolog 4.3* have given references with supporting arguments. The relationship of Tai Meuy with

³⁴ Considering the initial consonant [k^h-] in Tai Daeng and in Tai Meuy varieties leads to another important difference of Tai Daeng and Tai Meuy with Tai Yo. While Tai Yo lexical items such as [ha:w^{A1}] ‘white’, [haw^{C1}] ‘enter’, or [hɛ:n^{A1}] ‘arm’ have the initial consonant [h-], their cognates in Tai Daeng and in Tai Meuy varieties are [k^ha:w^{A1}], [k^haw^{C1}], and [k^hɛ:n^{A1}]. As a matter of fact, words whose initials were reconstructed by Li Fang-kuei (1977) as *x- or *kh-, and later by Pittayawat Pittayaporn (2009) as *x-, *ɣ-, or *q-, are never found to display in Tai Daeng and Tai Meuy varieties the initial [h-], a reflex specific to Tai Yo and other Tai dialects of Nghê An and Bolikhamxay related to Tai Yo such as Mène.

Tai Dam, which appears to be only a relationship between languages belonging to the P group, might be based on Vietnamese scholars' accounts linking Tai Meuay with a location in the Tai Dam area: "Muong Muoi, which today belongs to the district of Thuan Chau, province Son La" (Vi 1996:32).

Even if Tai Meuay varieties are definitely to be related to Tai Daeng, they appear to have a kind of relationship with Tai Yo and other Tai dialects of Nghệ An and Bolikhamxay which can be related to Tai Yo. That relationship, which can be described in terms of language contacts, mainly involves lexical items.

- **Verbs:** [ki:^{A4}] 'eat', [pa:^{A4}] 'go', [pe:^{A4}] 'be', [ʔa:^{A4}] 'take', [ʔe:^{A4}] 'do'.
- **Pronouns:** [ka:^{A4}] 'I', [sa:^{A4}] 'he, she, it'. It will be noted that, although Tai Meuay shares [ka:^{A4}] as a pronominal word with Tai Yo, [ka:^{A4}] has a different use in Tai Yo and expresses reciprocity.
- **Other items shared by Tai Meuay varieties and Tai Yo:** [ʔuk^{DS3} ʔik^{DS3}] 'brain', in contrast with [ʔe:k^{DL3}] in Tai Daeng; a tense marker expressing the future tense ([kham^{C1}]); the pluralizer [muaj^{C1}], which in that case tends to be pronounced [my:j^{C1}]; the phrase referring to the moon [ma:k^{DL1} buan^{A3}], as against [to:^{A2} buan^{A3}] in Tai Daeng.

In contrast with the relationship of Tai Meuay with Tai Yo and other Tai dialects of Nghệ An and Bolikhamxay related to Tai Yo, the relationship of Tai Meuay with Tai Daeng appears to be a deep linguistic and ethnolinguistic relationship. Chamberlain (1984:68) makes the following remark.

"The lineage names are identical to those of the Red Tai. There is a red band around the top of the woman's sarong which is worn tied above the breasts, also like the Red Tai."

This article has noted that the Tai Meuay share with the Tai Daeng the basket called [kaʔ ti:p^{DL2}]. A last example of that deep relationship is the fact that both Tai Daeng and Tai Meuay share a lexical particularity: they use different words for the lexical item 'mother' ([ʔe:^{C4}] in Tai Daeng, [me:^{C4}] in all Tai Meuay varieties) on the one hand and to refer to female animals ([me:^{B4}]) on the other. In Lao and most Tai languages spoken in Laos, the form [me:^{B4}] is used for both meanings.

Acknowledgements

First of all, I would like to thank all the people of Bolikhamxay Province who helped me in one way or another during this research, especially Viloth and Tiengkham Chanthavyxay. I will also mention the following individuals: Bounhômé Damnôngboun, Dengtoy Kimmixai, Khatthaya Vongsavanh, Khenmy Alounxay, Kinn Chandavong, Souksada Souththixay, Syvone Thammavong, Xiengpany Sythaochan.

Second, I am grateful to Sipaseuth Phongphanith and Vanh Thongmany, who accompanied me on some of my visits to Lakxao and helped me interview speakers there, to Sâm Công Danh and Lương Thị Dung, who did their best to provide me with a picture of the linguistic situation in Western Nghệ An, and to Vincent Berment, who shared with me his extensive knowledge of Lao lexicography and took the trouble to scan some documents for me.

Last but not least, I am indebted to James R. Chamberlain for his valuable comments, which definitely helped me better understand what he calls the issue of the "Meuay".

References

- Boersma, Paul and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.51, retrieved 11 August 2021 from <http://www.praat.org/>.
- Chamberlain, James R. 1972. The origin of the Southwestern Tai. *Bulletin des Amis du Royaume Lao* 7-8:233-244.
- Chamberlain, James R. 1975. A new look at the history and classification of the Tai languages. In Jimmy G. Harris and James R. Chamberlain (eds.), *Studies in Tai Linguistics in Honor of William J. Gedney*, 49-66. Bangkok: Central Institute of English Language.

- Chamberlain, James R. 1984. The Tai dialects of Khammouan province: their diversity and origins. *Science of Language Papers* 4:62–95. Bangkok: Chulalongkorn University (Department of Linguistics, Faculty of Arts).
- Chamberlain, James R. 1991. Mène: a Tai dialect originally spoken in Nghê An (Nghê Tinh), Vietnam: preliminary linguistic observations and historical implications. *Journal of the Siam Society* 79.2:103–123.
- Chamberlain, James R. 1996. Indigenous Peoples and Hydropower Development: A Case Study in the Use of Ethnolinguistic Analysis in Environmental Assessment. Paper given at Rencontre Internationale sur la Préservation des Patrimoines Immatériels des Groupes Minoritaires de la République Démocratique Populaire Lao, Vientiane, October 7–11, 1996.
- Chamberlain, James R. 1998. The origin of Sek: implications for Tai and Vietnamese history. *Journal of the Siam Society*. 86.1–2:27–48.
- Chao, Yuen Ren. 1930. A system of tone letters. *Le Maître Phonétique* 30:24–27.
- Đặng, Nghiêm Vạn. 2010. *Communauté nationale et pluri-ethnicité au Viêt-Nam*. Translated by Nguyễn Hương. Paris: Revue Péninsule (Cahiers de Péninsule 10).
- Ferlus, Michel. 2008. The Tai Dialects of Nghê An, Vietnam (Tay Daeng, Tay Yo, Tay Muong). In Anthony V.N. Diller, Jerold A. Edmondson and Yongxian Luo (eds.), *The Tai-Kadai Languages*, 298–316. London and New York: Routledge.
- Gedney, William J. 1972. A Checklist for Determining Tones in Tai Dialects. In M. Estellie Smith (ed.), *Studies in Linguistics in Honor of George L. Trager*, 191–205. The Hague: Mouton.
- Gedney, William J. 1989. A comparative sketch of White, Black, and Red Tai. In R.J. Bickner, J. Hartmann, T.J. Hudak & P. Peyasantiwong (eds.), *Selected Papers on Comparative Tai Studies*, 415–462. Ann Arbor: Center for South and Southeast Asian Studies, the University of Michigan.
- Gedney, William J. 1991. Comments on linguistic arguments relating to inscription one. In James R. Chamberlain (ed.), *The Ram Khamhaeng Controversy*, 193–226. Bangkok: The Siam Society.
- Guignard, Théodore. 1912. *Dictionnaire laotien-français*. Hongkong: Imprimerie de Nazareth.
- Kullavanijaya, Pranee and Theraphan L-Thongkum. 1998. Linguistic criteria for determining Tai ethnic groups: Case studies on central and south-western Tais. In S. Burusphat (ed.), *The International Conference on Tai Studies*, 273–298. Salaya, Nakhon Pathom, Thailand: Institute of Language and Culture for Rural Development, Mahidol University.
- Li, Fang-kuei. 1977. *A handbook of comparative Tai*. Honolulu: The University Press of Hawaii.
- Onlao, Sasithorn. 2010a. *The Grouping of Tai Language in Nam Paw Basin, Khamkeut District, Bolikhamxai Province, Lao People's Democratic Republic*. M.A. thesis, Mahasarakham University. (In Thai)
- Onlao, Sasithorn. 2010b. Grouping of Tai Languages in Nam Paw Basin, Khamkeut District, Bolikhamxai Province, Lao People's Democratic Republic by Using Criteria of Tonal Sound Changes. *Journal of Mekong Societies* 6. 3:47–72. (In Thai)
- Pacquement, Jean and Sipaseuth Phongphanith. 2012. Multilinguisme, plurilinguisme et compétence linguistique chez les Phu Thăi du centre du Laos et du nord-est de la Thaïlande: le cas des étudiants phu thăi de l'Université de Savannakhet. *Synergies Pays Riverains du Mékong* 4:129–139.
- Pittayaporn, Pittayawat. 2009. *The Phonology of Proto-Tai*. Ph.D. dissertation. Department of Linguistics, Cornell University.
- Robert, R. 1941. *Notes sur les Tay Dèng de Lang Chánh (Thanh-hoá – Annam)*. Institut Indochinois pour l'Étude de L'Homme (Mémoire N°1). Hanoi: Imprimerie d'Extrême-Orient.
- Schliesinger, Joachim. 2003. *Ethnic groups of Laos. Volume 3. Profile of Austro-Thai-speaking peoples*. Bangkok: White Lotus.

- Seidenfaden, Erik. 1967. *The Thai peoples. Book I: The origins and habitats of the Thai peoples with a sketch of their material and spiritual culture*. Bangkok: The Siam Society.
- Southixay, Souksada. 2016-2017. *The study of Muery language fundamental words in Phonsy Village Khamkert District Borlikhamxay Province*. Undergraduate dissertation, Department of Lao Language and Culture, Faculty of Letters, National University of Laos. (In Lao)
- Southixay, Souksada. 2019-2020. *Tai Meuy Amazing Stories*. M.A. thesis, National University of Laos. (In Lao)
- Vi, Van An. 1996. About names and settlement history of the Tai Groups settling in Nghe An Province (Central Vietnam). *Tai Culture* 1.1:30–40.

AUSTROASIATIC DISPERSAL: THE AA “WATER-WORLD” EXTENDED. WERE THE PROTO-AUSTROASIATICS COASTAL MIGRANTS?

Paul SIDWELL
University of Sydney
paulsidwell@gmail.com

Abstract:

This paper continues ideas introduced by Sidwell (2020) concerning Austroasiatic prehistory and the Neolithic transformation in Mainland Southeast Asia to propose a fundamentally new vision of the AA dispersal. I speculatively propose that a significant proportion of early AA speakers were oriented to estuarine environments—as opposed to inland or upland ones as has been commonly assumed. I abandon my earlier (Sidwell 2009, 2010 and elsewhere) proposed AA homeland on the middle Mekong, in favor of a Red River Delta locus of dispersal, from which settlers emigrated variously upstream (Northern AA speakers) and coastally around Indo-China to the Malay peninsula and India. This coastal movement may have been facilitated by improved watercraft, perhaps mediated by interaction with early Austronesians. The hypothesis is consistent with the aquatic component in the early AA lexicon (Blench 2018), and the emerging archaeological indications of the coastal spread of Neolithic rice farmers.

Keywords: Austroasiatic, homeland, migration, reconstruction

ISO 639-3 codes: vie, kuf, bru, bdq, sti, lbo, kjg, mlf, puo, pll, vbm, mnw, cbn, khm, cog, khr, pcj, srb

1 Introduction: the AA homeland reviewed

Over the past century or so, many ideas about the Austroasiatic (AA) homeland and migrations have been expounded, with sharply conflicting assumptions, and reliance on divergent kinds of evidence. Below, we briefly survey the range of proposals, before moving on to the new hypothesis. The extant range of AA homeland proposals are grouped below into seven broad categories, based on the geographical assumptions and the kinds of evidence regarded as significant. These are listed as follows:

1) Central China: It may be assumed that as Neolithic cereal farmers, AA correlates with ancient rice domestication about the middle Yangtze, potentially in relation to other Asian language families. The idea is particularly associated with the observation by Norman & Mei (1976) that the name of the Yangtze River in ancient Chinese resembles a generic AA word for ‘river’ (although the thrust of the N&M paper favors a Southern China hypothesis). Scholars such as Starosta (2005), Sagart (2003, 2011) and others, have posited a general origin of Asian languages in central China, citing broad typological correlations between multiple families:

The shared linguistic typology just described can be due to a very old genetic relationship between AA and STAN¹ [...] or to diffusion. In either case, a period of geographical closeness between languages ancestral to AA and STAN must be assumed.
(Sagart 2011:355)

While the broad explanatory power of the central China hypothesis also partly explains its attractiveness, no convincing body of lexical coincidences has emerged to support the hypothesis.

¹ STAN = Sino-Tibetan-Austronesian.

2) SW China: This model puts AA inland and upland, at or near the convergence of the great rivers emerging out of Tibet: the Salween, Mekong, Yangtze/Jinsha. The basic assumption is that communities migrated downstream to their present locations; ostensibly reasonable as it potentially explains the movement of AA speakers into both Indo-China and Eastern India. It is articulated by Blust (1996) and Peiros (2011), in particular.

3) Southern China: The region of Yunnan, Guangdong, Guangxi, and Fujian were inhabited by *Yue* or barbarians, later variously absorbed or driven south by the Chinese since Han times. The hypothesis was particularly supported by Norman & Mei (1976) comparing a handful of *Yue* and Chinese words with AA etyma; although a widely cited paper, the claims were essentially refuted by Sagart (2008). Schuessler (2006) controversially proposes hundreds of Chinese-Khmer lexical parallels (ignoring advances in AA reconstruction), suggesting that AA speakers were present in southern China and maybe as far north as Shandong. The southern China hypothesis is also supported by archaeologists and geneticists (e.g., Bellwood 2021, Higham 2021, Lipson et.al. 2018, McColl et. al. 2018) as evidence is strongly emerging that cereal agriculture came into Indo-China from southern China during the Neolithic.

4) Indo-China (central or northern): The diversity of AA branches in Indo-China, with half the family more or less along the Mekong, suggests a locus of origin and riverine dispersal vectors (e.g., Sidwell 2009, Sidwell & Blench 2011). Forms of the hypothesis go back to Schmidt's (1906) Austric proposal, which saw the overlap of AA and Chamic as indicating dispersal of Austric from Indo-China. Historically, northern Indo-China overlaps with *Yue* China, and the Red River Delta is ambiguous in this respect, so a hard line is not drawn between these here. Any hypotheses reliant on interpretation of language geography and diversity are dependent on phylogeny, and this remains controversial in regard to AA.

5) Bay of Bengal coast-hinterland: The center of diversity argument shifts eastward when one places special emphasis on Munda as a primary branch, and this is proposed by van Driem (2001, 2007) and Diffloth (2005, 2009). These scholars also note that the environment around the Bay of Bengal includes animals and plants indicated in the proto-AA lexicon, as well as giving medial proximity to both India and Indo-China. Presumably the locus ranges from the Ganges-Brahmaputra delta to the Irrawaddy Delta, and the Patkai Range in between. Although appealing on grounds of simplicity, this hypothesis lacks archaeological or other interdisciplinary support.

6) Eastern India: Among South Asia-oriented scholars there has been a strong tendency to identify India as the homeland of AA, and the Munda as an especially ancient population there. For example, Peterson (2017) makes a typological case that AA constitutes a prehistoric substrate in the eastern half of northern South Asia, investigating substrate hypotheses that go back to Kuiper (1948). Van Driem (2001) characterizes Munda as the most linguistically conservative AA branch, echoing Pinnow's (1963, 1966) discussions of Munda morphology and confidence that proto-AA was extensively prefixing and suffixing. Typological arguments have been taken further, with Donegan & Stampe (1983, 2002, 2004) arguing that Munda independently shifted in grammatical and rhythmic structure, and:

This suggests that the Austroasiatic people may have dispersed from South Asia rather than South-East Asia, and the shift of Munda from rising to falling rhythm, after the eastern language had moved eastward, may have been the cause rather than the effect of the profound polarization of South and South-East Asian language structures.
(Donegan & Stampe 2004:27)

However, the typological arguments have been strongly challenged recently (see, for example Anderson 2020, Jora & Anderson, this volume) and great antiquity of Munda in India is open to challenge.

7) Punjab: Witzel (1999 etc.) has argued that Vedic shows signs of a local substrate with AA features (especially particular prefixes), which he calls “Para-Munda” and regards as an ancient AA language, implying AA speakers in the Punjab. He also speculates that an AA may have influenced language in the Himalayas and eastern Afghanistan and had links to Sumarian. He speculates, “If indeed so, the speakers of (Para-)Austro-Asiatic would have been builders of a number of great civilizations, from Mesopotamia to Pakistan/India, Burma and Cambodia.” (Witzel 1999:12) While this is not an explicit claim of homeland location, Witzel places AA in western India prior to 1500 BC, with a clear implication that the language subsequently spread eastward.

It is clear that there is no obviously correct solution the AA homeland problem; well informed scholars have come to widely divergent and entrenched views. We would hope that if real progress were being made in recent years/decades, that views and evidence would be converging on some kind of consensus, with the clarity of the Austronesian out of Formosa hypothesis (see Blust 2019 for a recent overview). Another aspect of the problem is that secondary literature--such as reference works and interdisciplinary papers--have mostly failed to reflect the diversity of views of the subject of the AA dispersal. Specific claims and hypotheses are cited without critical engagement, gaining authority by dint of repetition and re-citation. Regrettably, these days AA studies is a small and geographically dispersed field, and without a critical mass of interested scholars and students it is difficult to sustain constructive discussions that can move things forward.

The homeland hypotheses reviewed here rely on a diversity of evidence and inferences, and while these can offer attractive narratives, mostly they are articulated without assessing the strengths of counter claims. We see this quite starkly, for example, in the fact that the India and China homeland claims have been supported by both lexical and typological comparisons. Yet those comparisons are not tabled against those of the competing hypotheses nor their strengths and weaknesses discussed. Furthermore, these are dependencies between the kinds of evidence that are used: for example, what are the probabilities associated with lexical similarities given particular typological alignments? What are the probabilities particular typological parallels will occur independently among neighbors, or would emerge from selective lexical comparisons? These are interesting questions and also go to general questions of how we assess the strengths of claims made about reconstructed and ancient languages (i.e., paleolinguistics).

Given this state of AA homeland views, we find maps like those presented here as figures 1 and 2, with arrows showing suggested migration paths in directly opposite directions. Such maps published before 2019, as far as I can tell, all have something in common: they imply primary movement overland and along rivers, with only the Nicobarese crossing substantial waters. The principle is so neat that the homeland problem seems to resolve to a discussion about whether peoples moved upstream or downstream here or there, such that we might draw our arrows going clockwise or anti-clockwise depending on our favored origin point. This may make for straightforward narrative, but it does little to actually test competing claims on their strengths, and it boxes any discussions into an inland riverine perspective at the expense of other possibilities.

Figure 1: “The Southeastern Riverine hypothesis for the Austroasiatic dispersal” (Sidwell & Blench 2011:339, figure 6)

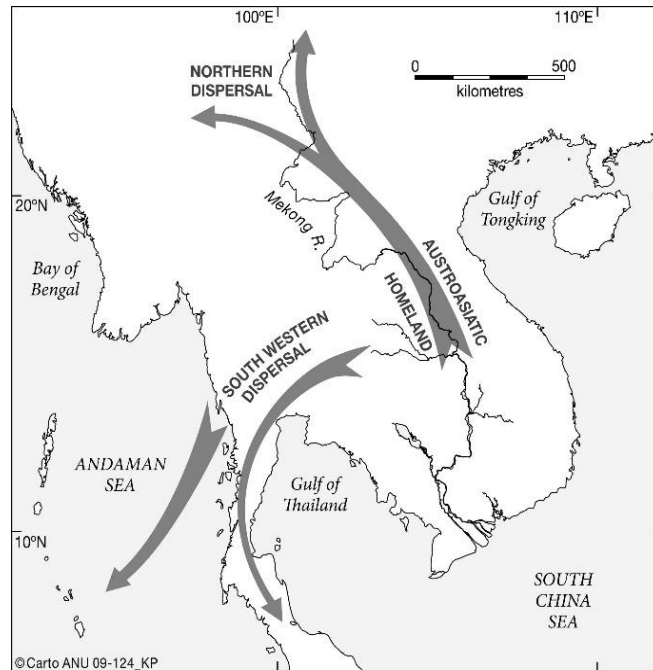
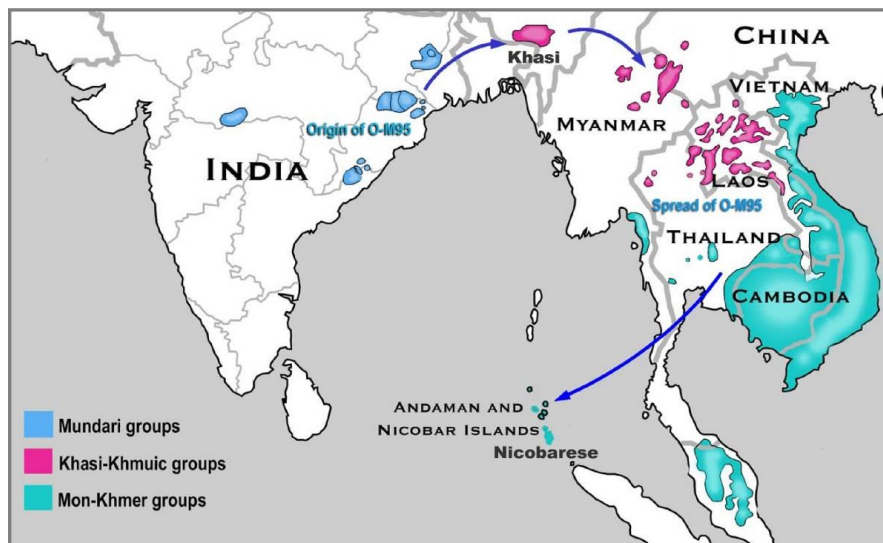


Figure 2: “...schematic representation of the routes of migration of the different Austro-Asiatic linguistic subgroups of India.” (Kumar et al. 2007:49, figure 1).



Another commonality among considerations of AA homeland is that ancient AA speakers were cereal cultivators before they dispersed into their recognized branches, based on the distribution of cereal related lexicon (see Sidwell & Rau 2014 for an overview). If this is accepted, it has strong implications for the possible AA center of dispersal. A very old chronology, say more than 7ky BP, effectively forces a homeland identification somewhere in China, proximal to the Yangtze, since we can confidently identify millet and rice in central China at such a time depth (in fact domestication of millet occurred further north than rice and may have led the spread of cereal cultivation chronologically (Stevens et al. 2021)). Such early dates also seem to rule out an Indian homeland as archaeology now indicates that rice and millet cultivation was practiced in the Indus Valley around 4400–4200 BP, and came somewhat later to the Ganges (Bates et al. 2017, Petrie et al. 2016).

A very old chronology for AA has found independent statistical support since the 1990s, tending to put proto-AA in a 7~9ky BP window. This includes the glottochronology of Peiros (1998, 2004) and the calibrated chronology by Greenhill using the 200-word data set that this author compiled (Sidwell 2015). To these we can also add Diffloth's (2005) 7,000-plus year chronology based on intuitive reckoning. While these dating estimates would seem to support a central China homeland, we have the problem that AA shows a rice vocabulary which is independent from other Asian language families (see Sagart 2011), and scant indication of any lexical parallels between AA and Sino-Tibetan or other families beyond some odd lexical similarities (e.g., Norman & Mei 1976).²

I take the view that the results of the various statistical studies have to be treated with great caution. Rates of lexical change may have been highly variable in the past, and the very low cognate scores that underlie the older dates may reflect uncontrolled factors such as word tabooing, and founder effects arising from small groups migrating to remote locations and their interactions with autochthonous populations. Consequently, analyses that include, for example, Munda or Nicobarese, may be skewed in ways we do not understand, and removing such outlying branches from consideration significantly reduces the age estimate.

Thomas (1973) applied the glottochronological replacement rates specified by Gleason (1955)³ to the lexicostatistical set of Thomas & Headley (1970), which did not include Nicobarese, Munda, or Pakanic, and found a root date squarely within the Indo-Chinese late-Neolithic. Thomas found that the calculations, "...could point to a mass dispersal from some central Mon-Khmer homeland some time during the 2nd millennium B.C." (1973:139). Specifically, a date of 3800 BP is indicated for 20% of cognacy, which is in the lower range of the inter-branch percentages, and thus can be regarded as indicative.

2 AA homeland and the Red River Delta

If we do locate the AA dispersal in time around, say, 4000 BP, a dispersal center in southern China or Indo-China becomes a serious possibility. Studies in recent decades across multiple disciplines support the view that agriculturalists from southern China began to migrate into the Red River Delta (RRD) and other parts of northern Indo-China from the last centuries of the fifth millennium BP, bearing the "Neolithic package" that included rice, millet, pigs, dogs, incised pottery, and supine burials, attested at sites such as Man Bac. This yielded a genetically mixed population, hence the term "Two-layer hypothesis":

The traditional archaeological evidence, including ceramic vessels and stone adzes, is thus unanimous in identifying the settlement of mainland Southeast Asia by at least 2000 BC by migrants from the north....
(Higham 2021:26)

... cranial and ancient DNA analyses of the excavated Neolithic skeletons (c. 1900 BCE) from this site reveal a remarkable admixture between an indigenous and morphologically Australo-melanesian population that was represented earlier at Con Co Ngu'a, and an immigrant Neolithic East Asian population that was morphologically related to modern Vietnamese. The latter entered Vietnam from the north, according to the ancient mitochondrial DNA record from Man Bac and cranial comparisons with Neolithic populations in China.

Indeed, it is clear that [...] the centuries around 2500-2000 BCE witnessed some remarkable cultural and biological changes in Southeast Asia (Hanihara et al. 2012; Higham 2013). South Chinese Neolithic populations with food production based on rice, millet, pigs and dogs pressed southwards, in the process settling alongside or simply amalgamating the indigenous hunter-gatherer populations (Bellwood 2015:55).

The resulting Phùng Nguyên culture (4000–3500 BP) of northern Vietnam is thus a prime candidate for being either proto-AA or late-AA speaking, and the direct ancestor of the Vietic languages of the region

² This is notwithstanding the fact that various claims for lexical parallels have been made. For example, Peiros (1998) presents some 15 AA-HmongMien comparisons in support of his wider Austric model.

³ Thomas specifies 1965, apparently in error.

spoken there today (cf. Alves 2021 for further discussion). Among archaeologists who advance the two-layer hypothesis, such as Charles Higham and Peter Bellwood, it is supposed that the incoming farmers were AA speaking, with the local indigenes shifting to AA speech. This assumption leaves open the possibility that AA originated somewhere further north, and that is not disputed here; in this paper we are concerned with where, when, and how the AA culture and language dispersed across the region. There may have been a pre-AA language that migrated from one location to another before ultimate dispersal. Alternatively, AA may have been indigenous to the RRD or some nearby territory, with inward settlers assimilating to the local language. I personally doubt that we will ever have a clear answer to this problem, and prefer to focus on the dispersal question.

The RRD as a locus of AA dispersal also implies that it is the Vietic homeland. While the Vietic homeland is not a matter of consensus, various scholars do explicitly identify the pre-Common Era Đông Sơn culture of the Red River Valley with Vietic, such as O’Harrow (1979), Ferlus (2009), Trần (2011), Taylor (2013), and Alves (2021) and continuity from Phùng Nguyên to Đông Sơn, both cultural and demographic, appears to be supported by a diversity of evidence. And as with Mon and Khmer, there is no evidence that indicates their migration from some inland origin; rather, Vietic speakers seem to have been living in the Red River delta, exploiting the rich estuarine environment for farming and gathering, for as long as we can tell.

The RRD, as an estuarine environment, is also consistent with the reflections offered by Blench’s (2018) paper: “Waterworld: lexical evidence for aquatic subsistence strategies in Austroasiatic”. That paper emerged after Blench and I cooperated on a (2011) paper combining linguistics and archaeology to propose a 4000 BP AA dispersal from somewhere along the Mekong, accepting traditional inland/riverine assumptions.

Blench (2018) explores in detail specific lexical indications for the AA homeland, focusing on likely utilization of aquatic environments by early AA speakers:

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. [...] 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. [...].
(Blench 2018:192)

Blench assembles evidence for proto-language items for:

Fauna: fish (general), catfish, crab, crocodile, eel, heron, otter, pelican, prawn/shrimp, tortoise, turtle, turtle (freshwater)
Fishing: to poison fish, scoop net, fish trap
Boats: three roots for boat
Geographical features: large river/sea, river valley, ditch/canal

Such lexical indications are arguably consistent with estuarine/coastal environments; potentially far richer in terms of diverse food sourcing than inland waterways, and demonstrably consistent with Mán Bac and other Phùng Nguyên sites that have been identified and documented in Northern Vietnam. Doubtless, there are many other locations that fit these broad indications if one includes all the coastline from the Pearl River delta to the Mahanadi-Brahmani Delta of India, but the RRD has a special resonance in this context of the two-layer hypothesis, and so is a strong candidate the AA locus of dispersal.

A specific linguistic argument can also be made for the primacy of the RRD. Ferlus (2009) argues that Vietic preserves evidence of the morphological formation of the word for ‘pestle’ (Vietnamese *chày*) from ‘to husk’ (Vietnamese *xay*) ultimately from ‘to dig, hollow, excavate’ by <r> infixation (*tʃe? > *tʃe: > *tʃre: > je:) speculating that it reflects the specific innovation in Southeast Asia of the wooden husking mortar and pestle, which proved easier to make and husked grains with less shattering compared to stone mortars. While Ferlus’ point is to suggest diffusion from ancient or pre-Vietic across AA, locating the AA dispersal in the Indo-Chinese Neolithic permits us to correlate the dispersal of ‘pestle’ with the dispersal of AA itself, and even hint at explanation for that dispersal, i.e., a special

case of the farming/language dispersal hypothesis (Renfrew 1987 and passim). The wooden pestle may have been a key innovation for early AA speakers adopting cereal cultivation, and carried widely as AA dispersed, even to be found in Munda. This strikes me as the best explanation for the distribution of the ‘pestle’ etymon across the family, as exemplified below:

Vietic: Vietnamese *chày*
 Katuic, Kantu **ntre:**, Bru **ntri:**
 Bahnaric: Bahnar **hđrəj**, Stieng **rənaj**, Laven **ʔrəj**
 Khmuic: Khmu **cn^dreʔ**, Thin **ŋgrəʔ**, Ksingmul **hagɛ:**
 Palaungic: Palaung **gre:**, Wa **ŋiʔ**
 Monic: Mon **riʔ**, Nyah Kur **ŋri:**?
 Khmeric: Khmer **ʔŋrɛ:**
 Pearic: Chong **kəhi:**^{R1}
 Munda: Kharia **eŋdi**, Gorum **in(d)ri**, Sora **ɔŋrij**

The root reconstructed *tʃeʔ ‘to dig, hollow, excavate’ by Ferlus is phonologically marked, as the *tʃ segment is rare in the reconstructed proto-Vietic sound system, and generally not reconstructed for proto-AA. Ferlus speculates that *tʃ was part of a layer of “Dongsonian vocabulary” that Vietic speakers took on as they migrated into the RRD region, correlating with the demographic “two-layer hypothesis”, I take no position here on the origin of *tʃ in AA. The dispersal of its reflexes is arguably aligned with the primary dispersal of AA.

Arriving at this point in our discussion, the question arises of by what routes and means did early AA groups disperse? What are the pros and cons of the RRD as the principal AA locus of dispersal? Is there something about the RRD that suggests a new way of looking at the problem? And is there anything else we need to consider?

3 Maritime Migration?

Among AA branches, only the Nicobarese were traditionally considered in terms of maritime migration, and then as an anomaly in a language family of mainly inland dwellers, reflecting the strong bias towards conceptualizing AA migrations as inland/riverine events. That changed recently with the “Munda Maritime Hypothesis” (MMH) of Rau & Sidwell (2019), which proposes that pre-Munda speakers migrated from SEAsia to Odisha (India) across or around the Bay of Bengal. Just as the Nicobarese necessarily reached their islands by sailing over open waters in Neolithic times, the MMH suggests that the pre-Mundas did something similar. Secondly, the MMH also asks if the pre-Aslian speakers also arrived on the Malay Peninsula by coastal navigation. If correct, these speculations imply the possibility of an ancient AA coastal/maritime culture on the shores of the Andaman Sea.

While the MMH marshals multiple lines of argumentation—linguistic, archaeological, geographical, and genetic—the main thrust is that the distribution of Munda languages is best explained by dispersals upstream and outward from the Mahanadi-Brahmani Delta (MBD). Assuming that the MBD was not itself the origin point of AA,⁴ a locus at the MBD is quite significant; the only other AA branch present in mainland India, Khasian, shows clear affiliation with the Palaungic language of Myanmar (Sidwell 2011) such that we can confidently suppose a pre-Khasian migration through Upper Burma to the Bhrmaputra Valley on the way to Meghalaya. Munda seems to have appeared on the east coast of India—tentatively correlating with the *Eastern Wetland Tradition*⁵ dated by archaeology to around 3500 BP—without signs of connections to any AA groups in India or immediately east of the subcontinent. In fact, the most striking lexical connections are between Munda and AA language of Indo-China, especially among the pronouns and numerals of Vietic and Bahnaric, suggesting a direct connection to the Vietnam coast.

⁴ This is itself a question worth considering, and I encourage it be tested thoroughly.

⁵ A Neolithic farming culture that predated the arrival of Indo-Aryans in Odisha.

Figure 3: Rau & Sidwell (2019:40, figure 2) proposed original dispersal region of Munda Languages.

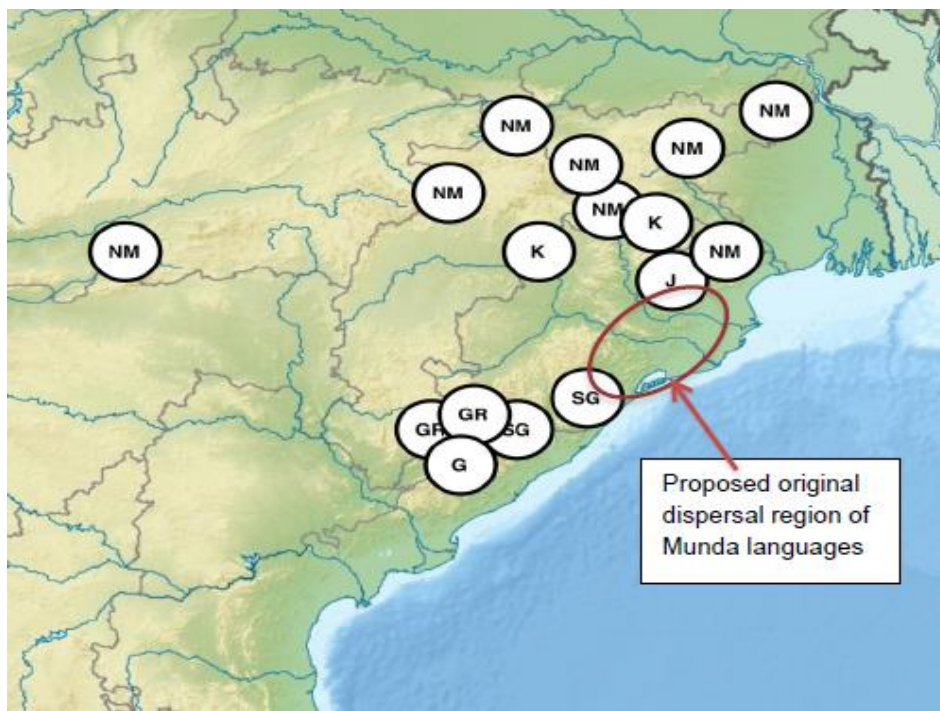
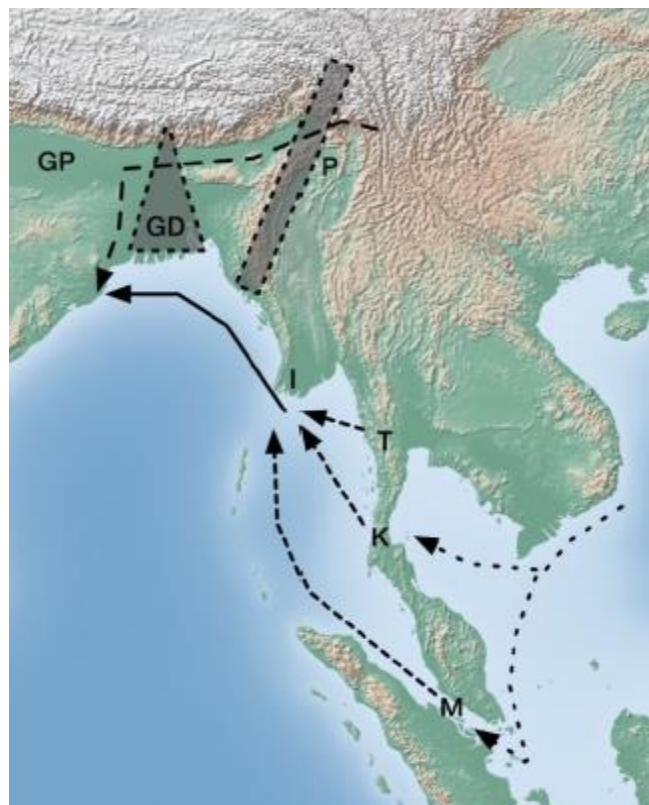


Figure 4: Rau & Sidwell (2019:45, fragment of figure 4) contrasting proposed migration routes through the Patkai Range (P) and Ganges Delta (GD) versus hypothetical maritime routes.



The case for at least a partial maritime component to the pre-Munda migration is strongly indicated by geography. The Ganges-Brahmaputra delta would have presented a significant barrier to migration in Neolithic times (see figure 4) and navigation past the mangroves and shallow river mouths has to be seriously considered. Furthermore, in such a journey across the northern Bay of Bengal, one finds the MBD is the first hospitable landing place once the mouth of the Ganges-Brahmaputra is passed, with the adjacent Chilka Lake functioning as a sheltered harbor and desirable estuarine fishery for millennia. It is as if the pre-Mundas appeared out of nowhere on the Odisha coast in Neolithic times in just the right place for landfall coming from the east. And there can be no doubt that pre-Mundas around 3500 BP had the capacity to skirt the Bengal coastline; the Nicobarese settlement involved crossing a greater stretch of open water than required to skirt the Bengal coastline, so there is simply no basis for ruling out coastal navigation as a mode of migration in the Neolithic.

Given these speculations regarding Neolithic AA navigation west of the Kra isthmus, I began to ask if coastal navigation had been even more widespread, and could it provide a wider basis to explain the distribution of AA branches? Could we examine if coastal/maritime migration provides as good or better explanation for the location of more AA branches than discussed in the MMH? There are hints of such speculations in an observation by Shorto four decades ago:

The Northern Mon-Khmers and Khasis are likely to have followed what became a Chinese trade route to India, as the Mundas may well have done before them. But there seems no overriding reason to trace routes for the Mons and Khmers, and other groups who occupied the river-plains, down the rivers from the hinterland rather than up them from the coast.
(Shorto 1979:278)

Note the phrase “up them [the rivers] from the coast”. It has been a staple of paleo-linguistics of East and Southeast Asia that language families dispersed substantially by moving down rivers, particularly down those with headwaters in the eastern Himalayas. But what happens if we reverse the logic and imagine that the early AA groups were principally coastal and estuarine dwellers whose inland migrations began upriver from the coast? Arguably several AA branches of Mainland Southeast Asia have their historical loci by the coast in estuarine settings.

Archaeology has established that the earliest Mon-Dvaravati urban settlements closely approximate the highest water line of the Chao Phraya River during the mid-Holocene sea-level maxima (Mudar 1999), suggesting that they developed from settlements founded right on the Bay of Bangkok coastline in Neolithic times.

Belonging to a later period, the earliest archaeological indications of Khmer urbanism (the Funan period) are found at locations such as Oc Eo in the Mekong Delta; from the first century CE this was a trading port on the “Maritime Silk Road” (Stark 2006, Higham 2014). Stark notes:

At least ninety “Oc Eo” period complexes have been recorded throughout southern Vietnam’s Mekong delta (Vo Si Khai 2003); contemporary sites have been reported along the coasts of peninsular Siam with similar material culture.
(Stark 2006:149)

The earliest identifiable Khmer settlements are coastal, around the gulf of Thailand, while settlements further inland or up-river are historically later, as the development of land and water management methods allowed for expansion into areas with less reliance on seasonal flooding and other forms of natural irrigation. Furthermore, similar sites are found on the Siam peninsular coast, consistent with the narratives of Coedes (1968), Wheatley (1961) and others who saw 1st century Funan as extending from the Mekong Delta, around the Gulf of Thailand, and dominating the isthmus including its west coast. While this period is nearly two millennia later than the putative Neolithic AA dispersal, it starkly makes the point that our contemporary sense of Khmers as oriented to the inland is an artifact of later history—the shift of the Khmer political center to Angkor after 800 CE, and subsequent reorientation towards an inland empire.

Over many centuries, various Austronesians came to dominate much of the Mainland Southeast Asian coastlines; Chams along the central and southern Vietnamese coasts, Malays around the peninsular, Mokens and other Sea Nomads, among others. Studies such as Thurgood (1999) have

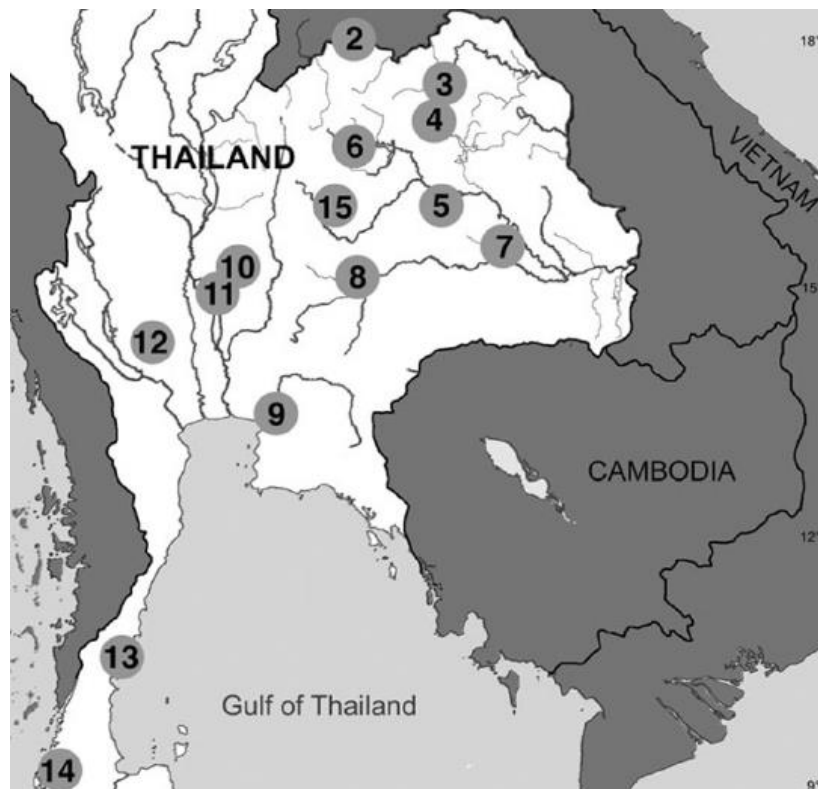
clearly established that there is an old AA substrate in the mainland Austronesian languages, and these were obviously acquired in contexts of coastal and/or insular interactions. The case seems compelling that a significant proportion of ancient AAs were oriented in their culture and economy towards the seacoasts, especially deltas and estuaries, leaving an imprint in Cham, Malay, and other Austronesian languages. Frankly, to my mind it stretches credibility that ancient AAs, as both farmers and foragers, would not take advantage to the opportunities the mainland estuarine environments present. As Higham points out:

The coasts of Southeast Asia, and particularly river estuaries, offer one of the world’s richest habitats in terms of natural bounty. On a world scene, hunter-gathers living in such marine habitats can secure their territory through permanent occupation and become very rich in social, technological and economic terms.
(Higham 2004:42)

Deltas and river flats are prime natural locations for rice farming, and deltas such the RRD have long been prime rice production regions. While upland dry rice cultivation is old in Mainland Southeast Asia and clearly important for AA (cf. Fuller & Castillo 2021 for overview), and upland cultivators are somewhat mobile, they typically do not move far when establishing new swiddens and often cycle back over the same ground after an appropriate period of fallowing. Also, upland cultivation is rather low in yield and does not support high populations or high population growth relative to lowland rice farming, which in favorable conditions can support multiple cropping within a given year. In this context, it makes sense to associate lowland rice cultivation with population growth and the hypothesized search for new estuarine environments with the kind of long-distance movements needed to account for a relatively rapid Neolithic dispersal of AA groups. Thus, we should not be surprised when Castillo (2011), citing Thompson (1996), writes:

The first evidence of domesticated rice in Thailand using macroremains dates to 2000–1500 BCE from the Neolithic period in the coastal site of KPD [Khok Phanom Di]
(Castillo 2011:115)

Figure 5: Fragment of Castillo (2011:115, figure 1) map showing sites with evidence of rice (Thailand).



Khok Phanom Di is (9) on Castillo's map (figure 5), shown right on the edge of the Bangkok plain today but four millennia ago located by the shore of the Bay of Bangkok.

There are potentially earlier sites in central and northern Thailand based on phytolith remains but apparently their evaluation is remains problematic. The evidence for early millet at these inland sites is much better. As Castillo remarks:

Higham (2002) originally proposed that rice agricultural expansion followed major riverine routes and would be archaeologically visible in interior sites, an idea previously put forth for Austroasiatic language expansion by Blust (1996). However, Ban Tha Kae and Ban Chiang are the earliest interior sites dating to the Neolithic and are reported to have rice cultivation, but the evidence is based on rice-tempered pottery, so it may be open to doubt.
(Castillo 2011:115)

We lack clear archaeological evidence for the expansion of rice cultivators along riverine routes as predicted by the traditional models of AA dispersal. Instead, there are some ambiguous indications of rice-husk impressed pottery at some early inland sites (i.e., these may be traded items). On the other hand, unambiguously, we have rice being grown by the Bangkok plain in Neolithic times. Consequently, I propose that the predecessors of the Mons and the Khmers were AAs sailing the coast looking for rich estuarine environments to settle (and perhaps bringing rice with them?). As Shorto mused, "...there seems no overriding reason to trace routes for the Mons and Khmers, ..., down the rivers from the hinterland...". Additionally, there are at least two other AA branches that we can examine from the same perspective.

Ferlus (2011) reconstructs the Pearic homeland in the vicinity of Thailand's Trat Province and Cambodia's Koh Kong Province. While the coasts of these provinces are nowadays dominated by ethnic Thais and Khmers, Pearic peoples live only tens of kilometres inland today. At the intersection of these provinces is the mouth of the Meteuk River and the modern Port of Koh Kong, which suggests a likely landing place for pre-Pearic settlers, providing both sheltered harbor and ready access to the inland.

Another potentially coastal AA branch is the AA substratum in Chamic, a putative lost AA branch, also discussed by Blench (2009). Thurgood's (1999) historical reconstruction of proto-Chamic demonstrated nearly 600 words of non-Austronesian origins in the Chamic historical lexicon, with roughly half of these identified as being AA in origin. While Thurgood assumed that this borrowing was largely from Bahnaric and Katuic into Chamic in ancient times, the direction of borrowing must have been largely from Chamic (Sidwell 2007, 2008). Furthermore, much of the proto-Chamic lexicon of AA origin cannot be identified with any specific AA branch, strongly suggesting that an unknown AA branch was spoken on the Vietnam central coast and absorbed into Chamic.

While admittedly circumstantial, we have now identified some seven AA branches whose locations can be accounted for by estuarine settlement and potential maritime arrival/departure in competition to traditional assumptions of inland origins and movement. However, there do remain AA branches that clearly fall outside the possibility of any maritime hypothesis, plus some ambiguous cases.

4 Inland Austroasiatic

The Khasian, Palaungic, Mang, Pakanic (and perhaps Khmuic) branches can be loosely grouped into a northern AA clade on the basis of sharing a form for the 1st person personal pronoun that reconstructs as *ʔə:ʔ (see Sidwell 2014 for overview of classification). Forms within Khmuic are split between reflexes of *ʔə:ʔ (Khmu, Mlabri) and *ʔap (all other Khmuic), the latter being widely (although not universally) reflected in the rest of AA (including Munda, e.g., Juang, Ho, Mundari /aj/ 'I'). Frankly, the status of Khmuic as a unitary branch is somewhat ambiguous, but for the present purposes it clearly falls into the category of geographically northern branches. It seems intuitively obvious that the distributions of all the Northern branches are readily explained by migrations upstream along the Red River and tributaries, with groups branching off as they went, eventually hitting the higher reaches of the Mekong and Salween, and ultimately descending to the Brahmaputra (see map at figure 6).

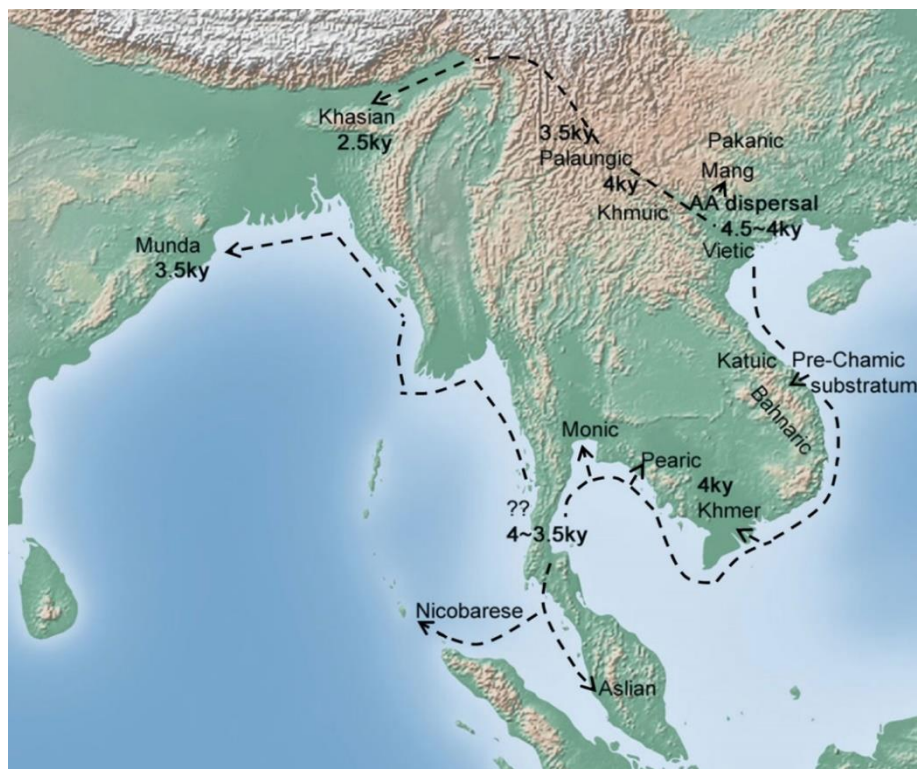
In addition to the Northern AA groups, Bahnaric and Katuic are ambiguous as their centers of diversity are in the hills of the Annamite range, with no apparent history of coastal presence. It has always been assumed more or less that these groups ascended the Annamite Range from the Mekong

side, although it is not at all clear that this is the case. We do see hints of directionality in the expansion of Kui people into Thailand and Cambodia, and of South Bahnaric speakers into Cambodia; these look to be relatively recent (Common Era) expansions given their internal homogeneity, perhaps related to the Khmer Empire providing motivations for lowland settlement. At a deeper historical level, we simply do not know who was living on the Indo-Chinese coast before the arrival of the ancient Chams, and it may well be that the ancestors of the Katuic and Bahnaric speakers were oriented to the coast in ancient times. In any case, there is nothing in particular about the distribution of Katuic or Bahnaric that contradicts an initial coastal dispersal, versus the inland alternative.

5 Conclusion

Our discussions have brought us to an interesting conclusion: one can account for the distribution of all AA branches on the bases of just two primary vectors: (1) inland up the Red River and beyond and (2) a coastal dispersal perhaps in one or more pulses, clockwise around the Indo-Chinese coast and around or across the isthmus into the Andaman Sea and beyond. We can speculate that AA migrants initially took with them rice and millet, and the wooden mortar and pestle for husking grains (with the Nicobarese, for example, eschewing these for a primarily fishery and tuber-based subsistence). Secondary expansions/migrations after the speculated coastal movements were primarily inland and up-river. We can propose that demographically, the first migrations were rather small and were focused on finding estuarine sites for settlement, rather than seeking to further ocean fishery or sea nomadism. Figure 6 presents a possible scenario for these migrations with approximate chronology.

Figure 6: Sidwell (2020:26, map 3). Speculative model of AA dispersal with maximal maritime component, including possible chronology of movements.



For more than a century, we have ignored the richest ecosystems, and the fastest modes of travel in ancient times, when considering the problem of AA dispersal and homeland. Our thinking has been conditioned by images of gradual overland and downriver movements of hunters, dry rice farmers, and vegeticulturalists. I propose that to some extent, we were projecting into the past our observations of the present and especially of those AA groups that today have under-developed economy and relatively little mobility, suggesting an unrepresentative image of historical character. Yet we are now beginning

to recognize that early rice farmers were sometimes capable coastal navigators. Higham (2021:26) discussing the introduction of rice to Indo-China states bluntly, ‘It is beyond reasonable doubt that there was a rapid coastal colonization by rice farmers.’ Should we assume a sudden abandonment of such a strategy once our early colonists had lit upon the fertile land of the RRD? I suggest that it is more likely that our colonists remained oriented to the coast for all sorts of practical reasons, but it just happens to be a serious challenge for archaeology to investigate ancient coastlines, such that sites further inland provide more stable conditions likely to favor preservation and discovery. It is also an issue that archaeology favors later eras with their greater production and diversity of artifacts. As Hung et. al. have expressed so clearly:

In many ways, the conspicuous archaeological record of the Iron Age has distracted our attention away from the likelihood of older cultural links across the South China Sea. In fact, the Iron Age connections very likely followed much older sealanes and trade-routes, ...
(Hung et al. 2013:400)

References

- Alves, Mark J. 2021. The Đông Sơn speech community: Evidence for Vietic. *Crossroads* 19 (2021):1-41.
- Anderson, Gregory. 2020. Proto-Munda Prosody, Morphotactics and Morphosyntax in South Asian and Austroasiatic Contexts. In Mathias Jenny, Paul Sidwell, Mark Alves (eds.) *Austroasiatic Syntax in Areal and Diachronic Perspective*. Leiden/Boston: Brill.
- Bates, Jennifer, Cameron Petrie, Ravindra Nath Singh. 2017. Approaching rice domestication in South Asia: New evidence from Indus settlements in northern India. *Journal of Archaeological Science* 78: 193-201. doi.org/10.1016/j.jas.2016.04.018.
- Bellwood, Peter. 2015. Vietnam and the Prehistory of Eastern Asia- a Multidisciplinary Perspective on the Neolithic. Andreas Reinecke (ed.) *Perspectives on the Archaeology of Vietnam*. Bonn, German Archaeological Institute - Berlin/ Bonn. 47-80 (in English and Vietnamese).
- Bellwood, Peter. 2021. Homelands and dispersal histories of Mainland Southeast Asian language families: a multidisciplinary perspective. In Paul Sidwell and Mathias Jenny (eds.) *The Languages and Linguistics of Mainland Southeast Asia: a comprehensive guide*. Berlin/Boston; De Gruyter Mouton. 33-44.
- Blench, Roger M. 2009. *Are there four additional unrecognised branches of Austroasiatic?* Presentation at ICAAL-4. Bangkok, October 29-30, 2009.
- Blench, Roger. 2018. Waterworld: lexical evidence for aquatic subsistence strategies in Austroasiatic. In Hiram ring and Felix Rau (eds.) *Papers from the Seventh International Conference on Austroasiatic Linguistics* (JSEALS Special Publication No. 3). Manoa: University of Hawaii Press. 174-193.
- Blust, Robert A. 1996. Beyond the Austronesian homeland: the Austric hypothesis and its implications for archaeology. *Transactions of the American Philosophical Society, New Series* 86.5: 117-158.
- Blust, Robert A. 2019. The Austronesian Homeland and Dispersal. *Annual Review of Linguistics* 5.1:417-434.
- Castillo, Cristina. 2011. Rice in Thailand: The Archaeobotanical Contribution. *Rice* 4:114-120. DOI 10.1007/s12284-011-9070-2.
- Coedes, Georges. 1968. *The Indianized States of Southeast Asia*. Translated by S.B. Cowing. Edited by W.F. Vella. Honolulu: East West Press. (translation of 1964 French edition, Paris, Editions E. de Boccard).
- Diffloth, Gérard. 2005. The contribution of linguistic palaeontology to the homeland of Austroasiatic. In: Sagart, Laurent, Roger Blench and Alicia Sanchez-Mazas (eds.). *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*. London & New York: Routledge/Curzon. 79-82.

- Diffloth, Gérard. 2009. More on Dvaravati Old Mon. Paper presented at the Fourth International Conference on Austroasiatic Linguistics. Mahidol University, Salaya. (Austroasiatic classification reproduced in: George van Driem. 2012. The ethnolinguistic identity of the domesticators of Asian rice. *Palevol* 11.2-3:117-132.
- Donegan, Patricia and David Stampe 1983. Rhythm and holistic organization of language structure. In: J. Richardson, M. Marks and A. Chukerman (eds) *Papers from the Parasession on the Interplay of Phonology, Morphology and Syntax*. Chicago: Chicago Linguistic Society. 337-353.
- Donegan, Patricia 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*. 111-120.
- Donegan, Patricia and David Stampe. 2004. ‘Rhythm and the synthetic drift of Munda.’ In R. Singh, (ed.), *The Yearbook of South Asian Languages and Linguistics*. Berlin/New York, Mouton de Gruyter: 3-36.
- Ferlus, Michel. 2009. A layer of Dongsonian vocabulary in Vietnamese. *Journal of the Southeast Asian Linguistics Society* 1:95-108.
- Ferlus, Michel. 2011. ‘Toward Proto-Pearic: Problems and Historical Implications.’ In Sophana Srichampa, Paul Sidwell and Kenneth Gregerson (eds.) *Austroasiatic Studies: papers from ICAAL4. Mon-Khmer Studies Journal Special Issue No. 3, part 1*. Dallas, SIL International; Salaya, Mahidol University; Canberra, Pacific Linguistics. 38-51.
- Fuller, Dorian Q. and Cristina Cobo Castillo. 2021. The origins and spread of cereal agriculture in Mainland Southeast Asia. In Paul Sidwell and Mathias Jenny (eds.) *The Languages and Linguistics of Mainland Southeast Asia: a comprehensive guide*. Berlin/Boston; De Gruyter Mouton. 45-59.
- Gleason, Henry Allen. 1955. *Workbook in Descriptive Linguistics*. Orlando: The Hartford Seminary foundation.
- Hanihara, Tsunehiko, Hirofumi Matsumura, Yoshinori Kawakubo, Nguyen Lan Cung, Nguyen Kim Thuy, Marc F. Oxenham, Yukio Dodo. 2012. Population history of northern Vietnamese inferred from nonmetric cranial trait variation. *Anthropological Science* 120. 2:157-165.
- Higham, Charles 2002. *Early cultures of mainland Southeast Asia*. Bangkok: River Books.
- Higham, Charles 2004. Mainland Southeast Asia from the Neolithic to the Iron age. In: Ian Glover & Peter Bellwood (eds.) *Southeast Asia: from prehistory to history*. 41-67. Abingdon: RoutledgeCurzon.
- Higham, Charles. 2013. ‘Hunter-Gatherers in Southeast Asia: From Prehistory to the Present. *Human Biology* 85.1-3:21-43.
- Higham, Charles. 2021. The Neolithic occupation of Southeast Asia. In Paul Sidwell and Mathias Jenny (eds.) *The Languages and Linguistics of Mainland Southeast Asia: a comprehensive guide*. Berlin/Boston; De Gruyter Mouton. 19-31.
- Hsiao-chun Hung, Kim Dung Nguyen, Peter Bellwood & Mike T. Carson (2013) Coastal Connectivity: Long-Term Trading Networks Across the South China Sea, *The Journal of Island and Coastal Archaeology* 8.3:384-404, DOI: 10.1080/15564894.2013.781085
- Kuiper, F. B. J. 1948. *Proto-Munda words in Sanskrit*. Amsterdam, Noord-Hollandsche Maatchappij.
- Kumar, Vikrant, Arimanda NS Reddy, Jagedeesh P. Babu, Tipiriseti N. Rao, Banrida T. Langstieh, Kumarasamy Thangaraj, Alla G. Reddy, Lalji Singh, and Battini M. Reddy. 2007. “Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations.” *BMC Evolutionary Biology* 7: 47. <https://doi.org/10.1186/1471-2148-7-47>.
- Lipson, Mark, Olivia Cheronet, Swapan Mallick, Nadin Rohland, Marc Oxenham, Michael Pietrusewsky, Thomas Oliver Pryce, Anna Willis, Hirofumi Matsumura, Hallie Buckley, Kate

- Domett, Nguyen Giang Hai, Trinh Hoang Hiep, Aung Aung Kyaw, Tin Tin Win, Baptiste Pradier, Nasreen Broomandkhoshbacht, Francesca Candilio, Piya Changmai, Daniel Fernandes, Matthew Ferry, Beatriz Gamarra, Eadaoin Harney, Jatupol Kampuansai, Wibhu Kutanan, Megan Michel, Mario Novak, Jonas Oppenheimer, Kendra Sirak, Kristin Stewardson, Zhao Zhang, Pavel Flegontov, Ron Pinhasi, David Reich. 2018. Ancient Genomes Document Multiple Waves of Migration in Southeast Asian Prehistory, *Science* 361.6397:92-95.
- McCull, Hugh, Fernando Racimo, Lasse Vinner, Fabrice Demeter, Takashi Gakuhari, J. Víctor Moreno-Mayar, George van Driem, Uffe Gram Wilken, Andaine Seguin-Orlando, Constanza de la Fuente Castro, Sally Wasef, Rasmi Shoocongdej, Viengkeo Souksavatdy, Thongsasayavongkhamdy, Mohd Mokhtar Saidin, Morten E. Allentoft, Takehiro Sato, Anna-Sapfo Malaspinas, Farhang A. Aghakhanian, Thorfinn Korneliussen, Ana Prohaska, Ashot Margaryan, Peter de Barros Damgaard, Supanee Kaewsutthi, Patcharee Lertrit, Thi Mai Huong Nguyen, Hsiao-chun Hung, Thi Minh Tran, Huu Nghia Truong, Giang Hai Nguyen, Shaiful Shahidan, Ketut Wiradnyana, Hiromi Matsumae, Nobuo Shigehara, Minoru Yoneda, Hajime Ishida, Tadayuki Masuyama, Yasuhiro Yamada, Atsushi Tajima, Hiroki Shibata, Atsushi Toyoda, Tsunehiko Hanihara, Shigeki Nakagome, Thibaut Deviese, Anne-Marie Bacon, Philippe Durringer, Jean-Luc Ponche, Laura Shackelford, Elise Patole-Edoumba, Anh Tuan Nguyen, Bérénice Bellina-Pryce, Jean-Christophe Galipaud, Rebecca Kinaston, Hallie Buckley, Christophe Pottier, Simon Rasmussen, Tom Higham, Robert A. Foley, Marta Mirazón Lahr, Ludovic Orlando, Martin Sikora, Maude E. Phipps, Hiroki Oota, Charles Higham, David M. Lambert, Eske Willerslev. 2018. The Prehistoric Peopling of Southeast Asia. *Science* 361:88-92.
- Mudar KM. 1999. How many Dvaravati kingdoms? Locational analysis of first millennium a.d. moated settlements in central Thailand. *Journal of Anthropological Archaeology* 18:1-28.
- Norman, Jerry and Tsu-lin Mei. 1976. The Austroasiatics in ancient south china: some lexical evidence. *Monumenta Serica* 32:274-301
- O'Harrow S. 1979. From Co-Loa to the Trung sisters' revolt: Viet-Nam as the Chinese found it. *Asian Perspectives* 22(2):140-163.
- Peiros, Iliia. 1998. *Comparative Linguistics in Southeast Asia*. Canberra, Pacific Linguistics (C-142).
- Peiros, Iliia. 2004. *Genetičeskaja klassifikacija avstroaziatskix jazykov*. Moskva: Rossijskij gosudarstvennyj gumanitarnyj universitet (doktorskaja dissertacija).
- Peiros, Iliia. 2011. Some thoughts on the problem of the Austro-Asiatic homeland. *Journal of Language Relationship*. 6:101-113
- Peterson, John. 2017. 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.
- Petrie, C., Bates, J., Higham, T., & Singh, R. 2016. Feeding ancient cities in South Asia: Dating the adoption of rice, millet and tropical pulses in the Indus civilisation. *Antiquity* 90(354):1489-1504. doi:10.15184/aqy.2016.210
- Petrie, Cameron, Jennifer Bates, Tom Higham, Ravindra Nath Singh. Feeding ancient cities in South Asia: dating the adoption of rice, millet and tropical pulses in the Indus civilisation. *Antiquity* 90(354):1489-1504; doi: 10.15184/aqy.2016.210
- Pinnow, Heinz-Jürgen. 1963. The position of the Muṇḍā languages within the Austroasiatic language family. in: H.L. Shorto (ed.). *Linguistic Comparison in Southeast Asia and the Pacific*. London: SOAS. 140-52.
- Pinnow, Heinz-Jürgen. 1966. A Comparative Study of the Verb in the Munda Languages. In: Norman H. Zide (ed.): *Studies in Comparative Austroasiatic Linguistics*. The Hague, Mouton (Indo-Iranian Monographs, V). 96-193.
- Rau, Felix and Paul Sidwell. 2019. The Munda Maritime Hypothesis. *Journal of the Southeast Asian Linguistics Society* 12.2:35-57.

- Renfrew, Colin. 1987. *Archaeology and language, the puzzle of Indo-European origins*. London: Jonathan Cape.
- Sagart, Laurent. 2003. The vocabulary of cereal cultivation and the phylogeny of East Asian languages. *Bulletin of the Indo-Pacific Prehistory Association* 23 (Taipei Papers, Volume 1): 127-136.
- Sagart, Laurent 2008. The expansion of *Setaria* farmers in East Asia: a linguistic and archaeological model. In: Alicia Sanchez-Mazas, Blench, R.M., Ross, M.D., I. Peiros & Marie Lin (eds.) *Human migrations in continental East Asia and Taiwan. Matching archaeology, linguistics and genetics*. 133-158. London: Routledge.
- Sagart, Laurent. 2011. The Austroasiatics: East to West or West to East. In Nicholas Enfield (ed.) *Dynamics of human diversity*. Canberra, Pacific Linguistics. 345-359.
- Schmidt, Wilhelm. 1906. Die Mon-Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens. *Archiv für Anthropologie, Braunschweig*, 5:59-109.
- Schuessler, Axel. 2006. *ABC Etymological Dictionary of Old Chinese*. Hawaii; University of Hawaii Press.
- Shorto, Harry L. 1979. The linguistic proto-history of mainland South East Asia. in: R. B. Smith & W. Watson (eds.), *Early South East Asia*. New York, Kuala Lumpur, Oxford University Press. 273-278.
- Sidwell, Paul and Felix Rau. 2014. *Austroasiatic Comparative-Historical Reconstruction: an overview*. In Mathias Jenny & Paul Sidwell (eds.) *The handbook of Austroasiatic languages*. Leiden, Boston: Brill. 221-362.
- Sidwell, Paul and Roger Blench. 2011. The Austroasiatic Urheimat: the Southeastern Riverine Hypothesis. In Nicholas Enfield (ed.) *Dynamics of human diversity*. Canberra, Pacific Linguistics. 315-344.
- Sidwell, Paul. 2007. The Mon-Khmer Substrate in Chamic: Chamic, Bahnaric and Katuic Contact. In *SEALS XII Papers from the 12th Annual Meeting of the Southeast Asian Linguistics Society* (2002), edited by Ratree Wayland et al. Canberra: Pacific Linguistics, The Australian National University. 113-128.
- Sidwell, Paul. 2008. On the sources of loans in the Proto-Chamic lexicon. In *SEALS XIV (Volume 1) Papers from the 14th Annual Meeting of the Southeast Asian Linguistics Society* (2004), edited by Wilaiwan Khanittanan. & Paul Sidwell, Canberra: Pacific Linguistics, The Australian National University. 261-267.
- Sidwell, Paul. 2009. *The Austroasiatic Central Riverine Hypothesis*. Keynote address read to the *19th Annual Meeting of the Southeast Asian Linguistics Society*, University of Social Sciences and Humanities, Vietnam National University-Ho Chi Minh City, 28 May 2009.
- Sidwell, Paul. 2010. The Austroasiatic central riverine hypothesis. *Вопросы языкового родства/Journal of Language Relationship* 4:117-134.
- Sidwell, Paul. 2011. Proto Khasian and Khasi-Palaungic. *Journal of the Southeast Asian Linguistics Society* 4.2:144-168.
- Sidwell, Paul. 2015. Austroasiatic dataset for phylogenetic analysis: 2015 version. *Mon-Khmer Studies* 44.lxviii-ccclvii.
- Sidwell, Paul. 2020. Austroasiatic Dispersal: The Indo-China Maritime Hypothesis. *Indian Linguistics* 81.12:15-32.
- Stark, Miriam. 2006. From Funan to Angkor Collapse and Regeneration in Ancient Cambodia. In Glenn M. Schwartz and John J. Nichols (eds.) *After Collapse: The Regeneration of Complex Societies*. Tucson: University of Arizona Press. 144-167.
- Starosta, Stanley 2005. Proto-East Asian and the origin and dispersal of the languages of East and Southeast Asia and the Pacific. In: Laurent Sagart, Roger Blench, Alicia Sanchez-Mazas (eds.)

- The Peopling of East Asia: putting together archaeology, linguistics and genetics*. London & New York: Routledge Curzon. 182-197.
- Stevens, C.J., Shelach-Lavi, G., Zhang, H. et al. 2021. A model for the domestication of *Panicum miliaceum* (common, proso or broomcorn millet) in China. *Veget Hist Archaeobot* 30:21-33. <https://doi.org/10.1007/s00334-020-00804-z>
- Taylor, Keith W. 2013. *A History of the Vietnamese*. Cambridge University Press.
- Thomas David. 1973. A Note on the Branches of Mon-Khmer. *Mon-Khmer Studies* 4:139-40.
- Thompson, G.B. The excavation of Khok Phanom Di, a prehistoric site in central Thailand, vol IV: subsistence and environment: the botanical evidence (the biological remains, part II). 1996. London: The Society of Antiquaries of London.
- Thurgood, Graham. 1999. *From Ancient Cham to Modern Dialects: two thousand years of language contact and change*. Oceanic Linguistics special Publications No. 28. Honolulu, University of Hawaii Press.
- Trần, Hữu Thung and Thái Kim Đình. 1997. *Từ điển tiếng Nghệ* (A dictionary of Nghe An Vietnamese). Vinh, Vietnam: Nhà Xuất Bản Nghệ An.
- van Driem, George. 2001. Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region: Containing an Introduction to the Symbiotic Theory of Language. Leiden: Brill.
- van Driem, George. 2007. Austroasiatic phylogeny and the Austroasiatic homeland in light of recent population genetic studies. *Mon-Khmer studies: a journal of Southeast Asian languages and cultures* 37: 1-14.
- Vo Si Khai. 2003. *The kingdom of Fu Nan and the Culture of Oc Eo*. In J.C.M. Khoo (ed.) Art and Archaeology of Fu Nan. Bangkok, Orchid Press, 35-86.
- Wheatley, Paul 1961. *The Golden Khersonese: Studies in the Historical Geography of the Malay Peninsula before A.D. 1500*. Kuala Lumpur: University of Malaya Press.
- Witzel, Michael. 1999. Substrate Languages in Old Indo-Aryan (Rgvedic, Middle and Late Vedic). *Electronic Journal of Vedic Studies*. 5-1:1-67.

NEW EVIDENCE FOR AUSTRO-TAI AND OBSERVATIONS ON VOWEL CORRESPONDENCES

Alexander D. SMITH

The Chinese University of Hong Kong

smith.alexander.david@gmail.com

Abstract

The Austro-Tai hypothesis proposes that Kra-Dai and Austronesian are descended from a common ancestor, Proto-Austro-Tai, which existed somewhere in Southern China several thousand years ago. In this paper, a newly created list of Kra-Dai and Austronesian shared lexical items is presented, along with already-proposed shared lexical items from earlier works resulting in a list of some 71 shared lexical items which may be ultimately inherited from a putative Proto-Austro-Tai language. The list is also used to analyze correspondences between Kra-Dai and Austronesian vowels in final syllables. It is shown that several regular correspondences are shared by the two families and that these regular correspondences most likely date back to a shared common ancestor.

Keywords: Austro-Tai, Austronesian, Kra-Dai, Historical linguistics

ISO 639-3 codes: ami, bhp, bnn, bny, bth, btx, bzg, ceb, ckv, doc, enc, iba, ilo, ind, ivv, kmc, kys, kzi, kzp, laq, lbc, lic, mak, mkg, msa, nij, nut, onb, pcc, pni, sas, skb, smr, sne, ssf, tao, tay, tgl, tha, tou, trv, tys, tyz, uun, xnb, yha, yln, yrn, yzg, zbc, zha, zyn

1 Introduction¹

The Austro-Tai hypothesis is a macro-family proposal which posits that Austronesian and Kra-Dai², two major language families of Southeast Asia and the Pacific, are descended from a common ancestor, henceforth referred to as Proto-Austro-Tai (hereafter, PAT)³. The proposal itself is not new — the first major publication on the topic, that of Benedict (1942), is nearly 80 years old, and the first ever mention of AN-KD relations (Schmidt 1906) is even older. In addition to linguistic similarities, recent genetic studies also suggest a link between AN and KD populations (Li et al. 2008). The AT proposal itself, however, has not gained wider acceptance due to issues both in the comparisons found in Benedict's

¹ Data Sources and Abbreviations are as follows: Proto-Austronesian (PAN) and Proto-Malayo-Polynesian (PMP), are from Blust and Trussel (ongoing), Proto-Kra (PK), Buyang, Qabiao, Pubiao are from Ostapirat (2000), Paha is from Li and Lou (2010). Note that Li and Lou's 2010 publication is a grammar of "Buyang", which is used as a cover term that includes several dialects. In this case, Li and Lou's "Buyang" is the Paha variety. Proto-Ong Be (POB) and Ong Be are from Chen (2018). Proto-Hlai (PH) and Hlai are from Ostapirat (2004). Lakkja is from Fan (2019). Southern and Northern Kam (S. Kam, N. Kam) are from Long et al. (1998). Proto-Tai (PT), and all lexical data of Tai languages, including Siamese, Sapa, Bao Yen, Cao Bang, Lungchow, Sangsi, Yay, and Saek are from Pittayaporn (2009). Additional abbreviations include AT (Austro-Tai), AN (Austronesian), and MP (Malayo-Polynesian).

² Kra-Dai is often referred to as Tai-Kadai.

³ A modification to the original AT proposal is given in Sagart (2004, 2005), where he states that Kra-Dai is a daughter, rather than a sister, of AN. Sagart's proposal places Kra-Dai as a sister to Malayo-Polynesian with a Formosan ancestor which he dubs FATK, or "Formosan Ancestor to Tai-Kadai". For the purposes of this paper, it is assumed that any relationship between KD and AN is at most a sister relationship, with an ancestor, PAT, giving way to two daughters, PKD and PAN.

original publication as well as sometimes serious methodological issues in his follow-up manuscripts, namely, Benedict (1975 and 1990).⁴

More recent research in AT, especially that in Ostapirat 2005, 2013, and 2018, has given new life to the AT hypothesis with methodologically strict and consistent research on a core set of shared vocabulary between AN and KD while keeping a safe distance from Benedict’s less methodologically sound post-1942 publications. Therefore, the present research looks to Benedict’s original 1942 publication and Ostapirat’s more recent research when evaluating past AT proposals and does not include comparisons from Benedict’s later works on the subject.

The present research is also meant to expand the core set of AT lexical and phonological comparisons with newly available datasets from KD. Through analysis of new datasets and application of the comparative method, this research is able to expand the number of comparisons to over 70, with much of the evidence coming from the Tai branch. The remainder of this paper is organized as follows: Section 2 gives background information on the AT Hypothesis itself and organizes a list of possible shared lexemes from various publications. Section 3 offers a list of new shared lexemes which are then added to the existing list. Section 4 discusses vowel correspondences between KD and AN. Section 5 concludes. Something to keep in mind when reading the paper is that KD and AN have differing orthographic conventions. AN *y* is equivalent to KD *j*, both indicating IPA [j]. AN *ñ* is equivalent to KD *ɲ*. AN **z* is equivalent to IPA [dʒ], and AN **j* is typically considered to be equivalent to IPA [gʲ]. In this study the different orthographical practices sometimes result in comparisons between AN -y and KD -j.

2 The Austro-Tai hypothesis

As already stated, the AT hypothesis is not new, but AT itself is not typically considered a “mainstream” language family. Before getting into the specifics of the proposal, some background on AN and KD is necessary. AN and KD occupy mostly nonoverlapping territories in Southeast Asia. AN languages are spoken in Taiwan, The Philippines, Malaysia, Indonesia, New Guinea, Madagascar, and throughout the Oceanic region. KD languages are spoken in Southern China, Laos, Thailand, parts of Myanmar, Vietnam, and Cambodia. The divide between AN and KD roughly follows the boundary between Mainland and Island Southeast Asia, although this division is fuzzy and exceptions are found.

Member languages in the two families tend to be typological opposites in many respects. In terms of phonological typology, conservative AN languages tend to have (C)V(C) syllables, four to six vowels, disyllabic canonical words, and are nontonal. As for syntactic features, conservative AN languages exhibit complex verbal morphology, ubiquitously AN voice systems, and verb-initial word orders. KD languages, in terms of phonological typology, have more complex syllable shapes such as (C)CV(:)(C), have larger vowel inventories including a higher prevalence of diphthongs, have canonically monosyllabic or sesquisyllabic words, and are tonal. Regarding morphosyntax, KD languages have little affixational morphology and a relatively strict SVO word order. A potentially interesting outcome of pursuing a research program with the AT hypothesis is learning both the nature of the proto-language itself as well as the mechanics of syntactic and phonological change that result in such opposite typologies.

⁴ In these later studies, Benedict attempts to strengthen the AT hypothesis via crude means. For example, he expands the family to include Miao-Yao in 1975, and further to include Japonic in 1990. Both of these inclusions allow for almost unlimited sources for his search for similar words, but the inclusion of these two languages is not widely supported. Another issue is the direct comparison of KD lexemes with both individual languages in AN and with late-stage proto-forms that contain innovated phonemes not present in PAN. An example of the latter is his direct comparison of Proto-Oceanic labio-velars with KD lexemes even though such phonemes were Proto-Oceanic innovations (Reid 1984-1985). Finally, Benedict expanded his comparison list by reconstructing complex and sometimes unnatural sequences of consonants to PAT with liberal use of parentheses and square brackets. Blust (2014) uses the term “proto-form stuffing” in his criticism of Benedict’s 1975 comparison between AN ‘forehead’ and KD ‘face’. To quote Blust (2014:310), “In order to relate what [Benedict] called “Indonesian” *[dɔ]aʔay=[dɔ]aʔay ‘forehead’ and related forms in Formosan languages such as Thao shaqish, Bunun daqis ‘face’ to Proto-Tai *hna ‘face’, he posited Proto-Austro-Tai *(q/)(n)dza[q]ai[s] ‘face, forehead’.”

Regarding the nature of the relationship between AN and KD, the proposals below assume that shared vocabulary is the product of inheritance from a distant common ancestor. Although it has been argued that shared vocabulary is not inherited (Thurgood 1994), this study considers the evidence sufficient for a genetic relationship hypothesis. Evidence in the form of shared vocabulary contains two important features, namely, the tendency for shared vocabulary to be “basic” and the regularity of sound correspondences between suspected cognates. The vocabulary listed in this and previous studies on the AT hypothesis overwhelmingly comes from basic vocabulary. Ostapirat (2013) makes a similar observation and notes that Chinese vocabulary in KD tends to be non-basic vocabulary whereas AN vocabulary tends to be basic. Regarding sound correspondences, Ostapirat (2005) laid a solid foundation of regular sound correspondences between AN and KD in both consonants and tone. Regularity of this kind is typically viewed as evidence for inheritance. Additionally, as will be discussed more later in Section 4, there is regularity in vowel correspondences, which adds additional evidence in support of the hypothesis that these lexemes are inherited from a common ancestor.

The current research is mostly dedicated to identifying and examining potential cognates between the two families. The current inventory of AN/KD potential cognates are from several sources stretching over multiple decades. Because of this, an attempt is made to organize a more comprehensive list before moving on to the original portion of this paper. The review begins with Benedict (1942), who gives a large list of potential cognates and later includes data sets from Ostapirat (2005, 2013).

2.1 Benedict’s comparisons

A list of comparisons from Benedict 1942 is given in Table 1, with the PAN reconstructions according to the original sources and modified wherever necessary to reflect current PAN orthographical practices. Benedict’s reconstructions were based on Dempwolff (1937), and therefore need much modification to make them compatible with modern PAN orthography. Benedict also compared some KD words directly with Indonesian, not with a reconstructed AN proto-language. In these cases, the Indonesian words are listed in italics.

Not all of the words on this list hold up to close scrutiny. Some, particularly those which are found only in Indonesian and not in any AN proto-language, should be excluded from any list of potential AT vocabulary. In this study, the acceptability of cognates are judged on three conditions: (1) Can the comparisons be reconstructed to at least one primary-level protolanguage in both AN and KD?; (2) Based on our current understanding of sound correspondences, are the comparisons regular?; and (3) Are the proper syllables being compared between the two groups (specifically, do the KD monosyllables correspond to AN final-syllables)? Based on these criteria, the words for the numerals ‘one’ through ‘ten’, ‘water’, ‘cry/weep’, ‘star/sun’, ‘eat’, ‘raw’, ‘this’, ‘bird’, ‘nose’, ‘grandfather’, ‘fire’, ‘head’, ‘tooth’, ‘die’, ‘rain’, ‘eye’, ‘black’, ‘moon’, ‘fart’, and ‘sour’ appear to pass scrutiny. The others do not, and an explanation for why they are ruled out as valid comparisons are given in the remainder of section 2.1.

Table 1: Potential AT etyma from Benedict 1942 with glosses, Proto-An, and Malay words

Gloss	An lexeme	Gloss	An lexeme
bird	*manuk	hair	<i>rambut</i>
black	*dəmdəm	head	*qulu
blind	*buCa	lungs	*pusuq
blood	*daRaQ	man	*Cau
boat	<i>perahu</i>	moon	*bulaN
body	<i>daging</i>	nest	<i>sarang</i>
bone	<i>tanduk</i>	night	*Rabiqi
breast	*susu	nine	*Siwa
cry/weep	*Caŋis	nose	*ijuŋ
day/sun	*waRi	one	*isa
die	*m-atay	rain	*Rabun/*ambun
door	<i>pintu</i>	raw	*qudip
ear	*Caliŋa	rice field	*bəna
eat	*kaən	seven	*pitu
eight	*walu	six	*ənəm
eye	*maCa	small	*qitik
fart	*qətut	sour	*qasəm
fat/oil	<i>minyak</i>	star/sun	*qajaw
father	<i>bapak</i>	ten	*sa-puluq
fire	*Sapuy	this	*ni
five	*lima	three	*təlu
flower	*buŋa	tooth	*[ŋ/n/l]ipan
foot	*qaqay	two	*duSa
four	*Səpat	water	*danum
grandfather	*əmpu (PMP)	yellow	<i>kuning</i>

2.1.1 Kra-Dai words compared with Austronesian penultimate syllables

A major difference between AN and KD is canonical word shape. AN languages are mostly disyllabic, a canonical word shape that can be reconstructed to PAN. KD languages are mostly monosyllabic or sesquisyllabic, although Ostapirat (2018) argues that PKD can be reconstructed with fully disyllabic words with internal evidence despite a total lack of disyllabic canonical words in modern KD languages. The shift to monosyllabicity in KD appears to be the natural consequence of ultimate-syllable stress, which results in the reduction and eventual deletion of the unstressed penultimate syllable but no reduction or deletion in final syllables. The result of this reduction is that monosyllabic words in KD correspond to the ultima in PAN, never the penult. Benedict, however, tended to compare KD monosyllables to both penultimate and ultimate syllables in AN at his own convenience. This practice allowed him to make many additional comparisons, but since the nature of penultimate syllable reduction as a consequence of stress placement on the ultima is well understood, the validity of comparisons between KD monosyllables and AN penultimate syllables is questionable. Therefore, only comparisons that match KD monosyllables with AN final syllables are considered valid. The following PAN reconstructions in example 1, which Benedict connects to KD vocabulary, are considered invalid because Benedict compared their penultimate syllable to his suspected KD cognates.

- | | | |
|---|--|---|
| 1 | *Caliŋa ‘ear’
*buŋa ‘flower’
*buCa ‘blind’
*pusuq ‘lung’ ⁵ | *waRi ‘day/sun’
*qaqay ‘foot’
*Rabiqi ‘night’ |
|---|--|---|

2.1.2 Words that are restricted to Indonesian and closely related languages

Other words can be discarded because they involve comparisons of KD words directly with Indonesian. This set of words is listed in example 2. In these cases, the Indonesian words cannot be reconstructed to a higher-level proto-language (PMP or PAN):

- | | | |
|---|--|---|
| 2 | <i>bapak</i> ‘father’
<i>perahu</i> ‘boat’
<i>minyak</i> ‘fat/oil’
<i>sarang</i> ‘nest’ | <i>kuning</i> ‘yellow’
<i>rambut</i> ‘hair’
<i>pintu</i> ‘door’
<i>tanduk</i> ‘horn’ (listed as bone in Benedict 1942) |
|---|--|---|

2.1.3 Words with additional issues in correspondences, semantics, and attestation

Three additional words, *daRaq ‘blood’, *qitik ‘small’, and *bəna ‘rice field’, are problematic due to their correspondences, semantics, and attestations. Details are shown in example 3.

- 3 ***daRaq** ‘blood’ - This comparison assumes a change of *q > t in KD, for example, PT *luət and or Lakkja *lie:t*¹¹, which is irregular (*q typically merges with *k in this position).
***qitik** ‘small’ - Benedict’s comparison is based only on Lati, but the Lati word could not be verified. Even so, comparison with a single language violates the reconstructability requirement.
***bəna** ‘rice field’ - This word compares KD with AN words meaning ‘river mouth’ or ‘lower part of river/tidal bore’. The semantics are considered too different to reconcile.

2.2 Ostapirat (2005 and 2013)

Other than Benedict, Ostapirat has contributed much to the field of AT studies and has introduced many additional suspected cognates. A combined list from Ostapirat 2005 and 2013 is given in this section, with the PAN reconstructions according to the original sources and modified wherever necessary to reflect current PAN orthographical practices. This is a simple list of all words that Ostapirat includes in his studies, and therefore several overlaps exist between this list and the list in Table 2.

⁵ This word also compares the word ‘lung’ to PAN *pusuq ‘heart’ with dubious semantics.

Table 2: Suspected AT etyma from Ostapirat 2005 and 2013 with gloss and PAN reconstruction

Gloss	PAn	Gloss	PAn
bear	*Cumay	moon	*bulaN
bird	*manuk	navel	*pudəR/puja
bitter/pungent	*paqiC	net	*aray
black	*tidəm	nine	*Siwa
boat	*aluja	nose	*ijun
borrow/lend	*Səzəm	one	*isa
centipede	*qali-Sipan	otter	*Sənaq
chaff/bran	*qəpah	plant	*mula
child	*aNak	rain	*quzan
clam/snail	*(kuSul)	raw/alive	*qudip
cry/weep	*Caŋis	saliva	*ŋajay
die	*m-aCay	sesame	*ləŋa (PMP)
eat	*kaən	shoulder	*qabaRa
eight	*walu	shrimp	*qudaŋ
excrement	*Caqi/*Caki	skin/scale	*kuliC
eye	*maCa	sour	*qasəm
fart	*qətut	stream	*qaRus (PMP ‘current’)
fat/oil/grease	*SimaR	sun/star	*qajaw
fire	*Sapuy	taro	*biRaq
flow/current	*qaluR	ten	*sa-puluq
grandmother	*aya	this	*ni
grandparent	*əmpu (PMP)	tongue	*Səma
hand	*qalima	tooth	*[ŋ/n/l]ipən
head	*qulu	two	*duSa
I	*aku	water	*daNum
leg/thigh	*paqa ‘thigh’	you	*Simu
louse	*kuCu		

2.2.1 Questionable comparisons

Ostapirat (2005) additionally includes the following comparisons in example 4 which are found in Atayalic, an AN primary branch, but which are not reconstructed to PAN.

4	yawn	Proto-Atayal	*surab
	mouse	Proto-Atayal	*qawlid
	leaf	Proto-Atayal	*ʔabag

These three comparisons are all used to demonstrate word-final voiced stop correspondences between KD and AN. There are two issues with these comparisons. First, the words themselves cannot be reconstructed to PAN with internal evidence. Although the presence of KD cognates theoretically allows for reconstruction without additional AN evidence under the AT hypothesis, the words in question compete with much more robustly attested PAN reconstructions *Suab ‘to yawn’, *labaw ‘mouse’, and *waSaw ‘leaf’, each with evidence from multiple AN primary branches. The second issue is methodological. Many proposed KD-AN cognates and sound correspondences are evidenced by multiple witnesses. The proposals in 4, however, are not corroborated with any second witnesses. The correspondence between PAN *-b and suspected KD cognates, for example, relies on this single lexeme. Considering the confinement of AN examples of potential word-final voiced stops to Atayal, the

competition between Proto-Atayal reconstructions and more robustly attested PAN reconstructions, and the lack of second witnesses for the proposed word-final correspondences, these words are left off of the list. Note that Ostapirat (2005:120) also points out that these comparisons are tentative.

An additional questionable comparison is *aray ‘net’, which Ostapirat reconstructs based on the comparison between PAN *aray ‘net’, Tai *hæe*, Kam-Sui *re*, and Hlai *ra:j*. The source of the PAN reconstruction is unknown, however, so it’s difficult to analyze the potential relationship between these words. It is therefore left off of the list unless it can later be verified with an AN source.

2.3 Listing established comparisons

Next, the various comparisons from Benedict (1942) and Ostapirat (2005; 2013) are combined into a single list. The resulting list is a moderately sized table (shown in Table 3) of likely shared lexemes between AN and KD numbering 52. Reconstructed forms reflect their AN reconstructions.

Table 3: Combined list of KD-AN etyma from Benedict (1942) and Ostapirat (2005; 2013) with glosses and PAN reconstructions

Gloss	PAN	Gloss	paN	Gloss	PAN	Gloss	PAN
bird	*manuk	fart	*qətut	moon	*bulaN	sour	*qaləsəm
alive; raw	*qudip	fire	*Sapuy	navel	*puja	sun; star	*qajaw
bear	*Cumay	five	*lima	nine	*siwa	thigh; leg	*paqa
bitter	*paqiC	flow	*qaluR	nose	*[i/u]ju ŋ	this	*i-ni
black; dark	*dəmdə m	four	*Səpat	one	*isa	three	*təlu
boat	*aluja	grandparen t	*əmpu	otter	*Sanaq	to cry	*Caŋis
chaff	*qəpah	grease	*Sima R	saliva	*ŋajay	to die	*m-aCay
child	*aNak	hand	*qalim a	sesame	*ləŋa	to plant	*mula
current	*qaRus	head	*qulu	seven	*pitu	tongue	*Səma
eat	*kaən	I	*aku	shoulder	*qabaR a	tooth	*[ŋ/n/l]ipə n
eight	*walu	lend	*Səzam	shrimp	*qudaŋ	two	*duSa
excrement	*Caqi	louse	*kuCu	six	*ənəm	water	*daNum
eye	*maCa	mist	*Rabun	skin; scale	*kuliC	your	*kamu

2.3.1 Notes on irregular comparisons

In addition to the 52 words in Table 3, there are six additional words which are frequently cited as being unproblematic AT vocabulary, but which nevertheless have some irregularities which should be mentioned. These six additional comparisons are shown in Table 4 with a more detailed discussion of each lexeme afterward. Inclusion in this list does not imply that these comparisons should be removed, but simply that there are some irregularities which may be explained better in future work.

Table 4: Proposed AT etyma from Ostapirat (2005) with irregularities

Gloss	PAN
clam; snail	*kuSul/*(-kul)‘clam; snail’
fat; grease; oil	*SimaR ‘fat; grease; oil’
grandmother	*aya ‘grandmother’
rain	*quzan ‘rain’
taro	*biRaq ‘taro’
ten	*puluq ‘ten’

***puluq ‘ten’** - The first irregular comparison involves *puluq ‘ten’. Most basic KD and AN numerals are straightforwardly connected. For example, there is good evidence that the numerals ‘one’ to ‘nine’ are shared by KD and AN. In Buyang, for example, the numerals ‘two’ to ‘nine’ are clearly related to their PAN counterparts, as shown in example 5.

5	Buyang	PAN
	ea ^{A1} ‘two’	*duSa
	tu ^{A1} ‘three’	*təlu
	pa ^{A1} ‘four’	*Səpat
	ma ^{A2} ‘five’	*lima
	nam ^{C1} ‘six’	*ənəm
	tu ^{A2} ‘seven’	*pitu
	ðu ^{A2} ‘eight’	*walu
	va ^{A1} ‘nine’	*Siwa

The word for the numeral ‘ten’, however, differs in the presence of a seemingly irregular reflex of *q. Comparisons of ‘ten’ include the following in example 6.

6	PAN *puluq	PH *apu: ^C	Paha vat ^{D1} / pwat ^{D1}	Buyang put ^{D1}
---	------------	-----------------------	---	--------------------------

The assumed history of this word is as follows: First, the penultimate vowel deletes, causing an intermediary *pl- cluster in PKD daughter languages. This simplified to *p as in Paha vat^{D1} and Buyang put^{D1}. However, there are still issues with this comparison. PAN *-q corresponds to -k in KD: Paha naak^{I1} ‘otter’ : PAN *Sanaq ‘river otter’ and Paha taak³³ ‘vomit’ : PAN *utaq ‘vomit’. Ostapirat (2005) postulates that the word-final sequence *-uq undergoes a fronting process: *-uq > *-uiq > *-uq (-ut in Kra). This condition is unexpected, since word-final [q] tends to have a lowering and backing effect on preceding vowels, not a fronting effect. Another possible comparison with *-uq, PAN *-tuq ‘to fall’ : PH *ʔtuk ‘to fall’, Paha tək⁵⁵, Buyang tuk⁵⁴, S. Kam tok⁵⁵ may indicate that *q from *-uq merged with *-k just as it did after other vowels. If this is true, then word-final t in this comparison diminishes its strength. Sagart (2010) proposes that -t is from a linker which is now fused on the KD root, although the Austronesian equivalent is only found in Philippine languages.

***aya ‘grandmother’** - This word is reflected as ‘grandmother’ in KD, but probably referred to one’s paternal aunt/paternal aunt’s husband in PAN. This reconstruction holds for MP, but Formosan reflexes mean either ‘mother’ (Taokas) or ‘mother; mother’s sister’ (Atayal). At any rate, it is possible that apparent comparisons between KD ‘grandmother’ and AN ‘paternal aunt/paternal aunt’s husband’ underwent semantic differentiation.

***quzan ‘rain’** - Only one comparison is shown in KD, but this comparison has an l reflex of word-final *-N, which is considered a result of the short vowel reflex in Laha dʒal. There are not many words that can be used to test if this hypothesis is true. The only other reflex of a word-final -N is from *bulaN, where it is reflected with -n and the vowel is long. Medial *z does not provide additional

insight, since there is no overlap between reflexes of *səzam ‘to borrow’ and *quzan ‘rain’. Additionally, *z may be affected by the preceding *u, making direct comparison between reflexes of *z from *səzam and *z from *quzan difficult.

***biRaq ‘taro’** - Reflexes of *biRaq in KD pose issues because of the irregular devoicing of word-initial *b in PT *pruək and Paha *pwaak*¹¹. It is possible, however, that devoicing occurs due to the intermediate *br- cluster from PKD. There is not enough data to tell if this is regular, however.

***kuSul ‘clam; snail’** - Ostapirat reconstructs *kuSul from PMP *kuhul with additional KD evidence: Tai *hɔɔi*, Kam-Sui *khuj*, Hlai *tshei*, Kra *ci*. He points to the typical *S > *h sound change which took place between PAN and PMP, suggesting that although there is no Formosan evidence that the PAN reconstruction can be validated with KD evidence. This comparison may be correct, although there is no other corroborating evidence for the l : -j correspondence between AN and KD. If *-l merged with *-R, then the change may have followed suit.

***SimaR ‘grease; fat; oil’** - Two issues are found with this comparison. First, Tai reflexes of words with high vowels in the penultimate syllable and the low vowel in the final syllable typically have high-vowel reflexes (*a > *u): PAN *bulaN : PT *ɬluən^A, PAN *tubah : PT *C̣.buə^A, PAN *puja : PT *ɬwuː^A. The low-vowel reflex, Tai *man*, is therefore unexpected. Second, there is no other corroborating evidence to suggest that *-R regularly became KD -n. The only other comparison with word-final *R is from *qaluR, where *-R became -j in KD (PT *ɬlwaj^A).

3 New comparisons

In the following section, additional suspected cognates are presented which are the result of recent research into the relationship between KD and AN. The goal of listing these comparisons is to expand the list of KD-AN cognates, which should allow for more accurate comparisons and descriptions of sound correspondences. Since there are relatively few suspected cognates at the current stage of AT research, the most important task at this stage is to try and expand the list as much as possible while maintaining a strict adherence to the comparative method.

The suspected cognates are listed in table 5 as individual AT etyma with the current PAN reconstructed forms and updated definitions based on both AN and KD. After Table 5, each suspected AT lexeme is presented with examples from both AN and KD and further discussions and explanations where necessary.

Table 5: New Austro-Tai etyma

Gloss	PAN	Gloss	PAN	Gloss	PAN
afraid; timid; fear	*talaw	uncertainty marker	*nu	to come; arrive	*datəŋ
derris root; fish poison	*tubah	rattan	*quay	to fall	*-tuq
fish hook	*kabit	shadow	*qaNɪŋu	to sell	*saliw ‘to
hold in the fist	*kəmkəm	sick	*sakit	to transplant	*Canəm
leech	*məCak	spotted	*bəlaŋ		

***talaw ‘afraid, timid, fearful’**

In AN, there are two reconstructions with overlapping meaning: *takut ‘fear’ and *talaw ‘timid; fearful; coward’. The first tends to refer to the feeling of fear, whereas the second is a description of a person’s actions or characteristics. Reflexes in KD tend to simply refer to the state of being afraid. There may not have been a strict distinction between ‘afraid’ and ‘coward’ in AT with both being attributed to the word *talaw.

AN examples: PAN *talaw ‘timid; fearful; coward’, Amis *talaw* ‘to be afraid’, Itbayaten *taxaw* ‘cowardliness’, Cebuano *talaw* ‘to back off; be afraid to do s.t.’, Singai *təru* ‘to be afraid’, Aoheng *tao* ‘to be afraid’.

KD examples: PT *^hla:w^A ‘afraid’, Bao Yen *la:w*^{A1} ‘afraid’, Qabiao *la:w*²⁴, Mak *lu*¹, S. Kam *jaw*¹³

***quay ‘rattan’**

Reflexes of *quay are restricted to Tai in KD, but this may be a product of sampling, since many resources do not include ‘rattan’ in basic vocabulary.

AN examples: Pazeh *ɣuay*, Amis *ɣoway*, Cebuano *uwáy*, Singai *ui*, Aoheng *ui*

KD examples: PT *^ç.wa:j^A ‘rattan’, Siamese *wa:j*^{A1}, Sapa *va:j*^{A1}, Lungchow *va:j*^{A1}, Yay *va:j*^{A1}

***Canəm ‘to transplant a crop; plant; burry’**

KD evidence allows for a distinction between *mula ‘to plant’ and *Canəm ‘to transplant’. In AN, the additional meaning of ‘to bury’ is applied to *Canəm, but this appears to be a semantic extension not present in KD.

AN examples: Amis *tadəm* ‘grave; tomb’, Cebuano *tanúm* ‘to plant’, Malay *tanam*, Chamorro *tanom*

KD examples: PT *^t.nam^A ‘to transplant’, Siamese *dam*^{A1}, Lungchow *dam*^{A1}, Saek *tram*^{A1}, Paha *tam*^{C1} ‘to plant’, Buyang *ɣdam*²⁴.

***tubah ‘derris root; fish poison made from the derris root’**

This comparison involves semantics which may not be obviously linked. The derris root is commonly ground up into a substance which is used to poison fish in a contained body of water, traditionally used to gather large quantities of fish with little effort. The use of the term as ‘fish poison’ in Tai, but as the name for the plant in PAN, is not unexpected. Like *quay, reflexes of *tubah are restricted to Tai. This is also most likely a product of sampling, since most resources do not include entries for ‘derris root’ or ‘fish poison’.

AN examples: Saisiyat *ta-toba?* ‘fish poison’, Pazeh *ta-tuba* ‘derris poison’, Itbayaten *tova* ‘plant used for fish poison’, Cebuano *túba* ‘kind of croton plant’, Malay *tuba* ‘derris root’

KD examples: PT *^ç.buə^A ‘fish poison’, Siamese *buə*^{A1}, Sapa *bu*^{A1}, Lungchow *bu*^{A1}, Saek *viə*^{A1}

***kəmkəm ‘fist; hold in the fist’**

AN examples: Itbayaten *kəmkəm* ‘handful’, Cebuano *kumkúm* ‘hold something in the hand’, Simular *xənkəm* ‘a closed handful’, Kaidipang *koŋgomo* ‘to hold in the fist’

KD examples: PT *kam^A, Siamese *kam*^{A1}, Sapa *kam*^{A1}, Lakkja *kam*⁵¹, Qubiao *kam*²¹³

***datəŋ ‘to arrive; reach a place’**

AN examples: Itbayaten *ratəŋ* ‘arrival’, Ilokano *datəŋ* ‘arrival’, Malay *datəŋ* ‘to come’, Sasak *datəŋ* ‘to come; arrive’

KD examples: PT *^ç.təŋ^A ‘arrive’, Siamese *t^huŋ*^{A1}, Sapa *t^huŋ*^{A1}, Lungchow *t^hrŋ*^{A1}, Saek *t^həŋ*^{A2}, POB *^dəŋ^{A1}, Lakkja *təŋ*²³¹ ‘come’, S. Kam *təŋ*⁵⁵ ‘come’

***qaNiŋu ‘shadow’**

AN examples: Kavalan *niŋu*, Bunun *qaniŋu*, Itbayaten *anino*, Cebuano *anínu*, Bimanese *ninu*

KD examples: PT *^ŋaw^A, Siamese *ŋaw*^{A2}, Bao Yen *ŋɤw*^{A2}, Yay *ŋaw*^{A2}, PH *aŋa:u*^C, Paha *ŋau*⁴⁵

***kabit ‘fishing hook’**

AN examples: Pazeh *kabit*, Amis, *kafit*, Tagalog *kabit*, Iban *kabit*

KD examples: PT *^βet^D ‘fish hook’, Siamese *bet*^{DS1}, Cao Bang *bet*^{DS1}, Sapa *bit*^{DS1}, Bao Yen *brt*^{DS1}, Lungchow *bit*^{DS1}

***məCaŋ ‘paddy leech’**

AN examples: Amis *ɣaɣintaŋ*, Kakanabu *niməca?*, Ilokano *alintá*, Malay *lintah*, Singhi *rimotah*

KD examples: PT *da:k^D ‘leech’, Siamese t^ha:k^{DL2}, Bao Yen ta:k^{DL2}, Yay ta:k^{DL2}, POB *d̪a:k^{D1}, PH *ʔta:k, Paha ŋa³¹taak³³, Lakkja la:k^{DL2} (compare *maCay → Lakkja *plej*)

***saliw ‘to sell’**

AN examples: Amis *caliw*, Ivatan *mapasaliw*, Ilokano *saliw*, Proto-Sangiric *saliu*

KD examples: PH *aRi:u^C, PK *s-ywi

***nu ‘marker of uncertainty’, *i-nu ‘where’, *a-nu ‘who’**

AN examples: Seediq *ma-nu* ‘what; which’, Thao *mi-ni* ‘why’, Itbayaten *di-no-h* ‘where’, Kayan *hi-noʔ* ‘where’

KD examples: Paha *nau* ‘who’, Mulam *nau*² ‘who’, N. Kam *nəu*²² ‘who’ / ʔo²²*nəu*²² ‘where’, Qabiao *njau*⁴⁵ ‘who’

***-tuq ‘fall’ (a monosyllabic root)**

This is the only comparison here which includes a monosyllabic root. In AN studies, monosyllabic roots are typically -CVC segments that appear at the end of a word and are similar to phonaesthemes in that they are not lexical but seem to have recurring meanings where they occur (Brandstetter 1916; Blust 1988). The AN examples therefore come from different lexemes which utilize this root but do not necessarily descend from a single PAN source. The existence of monosyllabic roots in AN with comparisons in KD may indicate the presence of monosyllabic words in PAT which were modified in AN to conform to a disyllabic requirement, creating irregularities in pre-final syllables but a common monosyllabic word-final root.

AN examples: Bintulu *gatuʔ*, Malay *rentoh*, Kelabit *tutuʔ*, Berawan *sito*

KD examples: PH *ʔtuk ‘fall’, PK *tok^D ‘fall’, Buyang *tuk*^{D1} ‘fall’, S. Kam *tok*⁵⁵ ‘fall’

***sakit ‘sick; to hurt; be in pain’**

Both *sakit and *bəlaŋ are restricted to MP in AN, which poses an issue for these comparisons. However, unlike Proto-Atayal *surab ‘yawn’, *qawlid ‘mouse’, *ʔabag ‘leaf’, which compete with more robustly attested PAN reconstructions, *sakit ‘sick’ does not have a PAN word with which it competes. This may be due to the loss of reflexes of a putative PAN *sakit in Formosan languages but its retention in MP and KD. *bəlaŋ, however, competes with PAN *paCak ‘spotted’ and is therefore more problematic, although the sound correspondences between PT *bla:ŋ and PMP *bəlaŋ are regular.

AN examples: Ilokano *sakit*, Tagalog *sakit*, Malay *sakit*, Karo Batak *sakit*

KD examples: PT *ke:t^D ‘to hurt’, Shangsi *ket*^{DS1}, Yay *cet*^{DS1}, Saek *ke:t*^{DS1}, Lakkja *tɛ:t*²⁴ ‘to hurt, ache’, N. Kam *kit*³³

***bəlaŋ ‘spotted’**

AN examples: Ngaju Dayak *bəlaŋ*, Malay *bəlaŋ*, Sasak *bəlaŋ*, Makassarese *bəllaŋ*

KD examples: PT *bla:ŋ^B ‘spotted’, Siamese *da:ŋ*^{B1}, Sapa *ba:ŋ*^{B1}, Bao Yen *bja:n*^{B1}, Cao Bang *da:ŋ*^{B1}

With the additional comparisons listed above, a list of 71 comparisons may be presented in Table 6. The list will undoubtedly change as additional comparative research is conducted on the two families. However, the number of apparently valid PAT lexemes with reflexes in both KD and AN continues to increase and the number is likely to go up, rather than down.

Table 6: Combined list of AT etyma from Benedict, Ostapirat, and the present research with gloss and PAN reconstruction

Gloss	PAN	Gloss	PAN
afraid	*talaw	one	*isa
aku	*aku	otter	*Sanaq
alive; raw	*qudip	rain	*quzan
bear	*Cumay	rattan	*quay
bird	*manuk	saliva	*ŋajay
bitter	*paqiC	sesame	*lɔŋa
black; dark	*dəmdəm	seven	*pitu
boat	*aluja	shadow	*qaNiŋu
chaff	*qəpah	shoulder	*qabaRa
child	*aNak	shrimp	*qudaŋ
clam; snail	*kuSul	sick	*sakit
current	*qaRus	six	*ənəm
derris root	*tubah	skin; scale	*kuliC
eight	*walu	sour]	*qaləsəm
excrement	*Caqi	spotted	*bəlaŋ
eye	*maCa	sun; star	*qajaw
fire	*Sapuy	taro	*biRaŋ
fish hook	*kabit	ten	*puluq
five	*lima	thigh	*paqa
flow	*qaluR	this	*i-ni
four	*Səpat	three	*təlu
fresh water	*daNum	to come; arrive	*datəŋ
grandmother	*aya	to cry	*Caŋis
grandparent	*əmpu	to die	*m-aCay
greese; fat	*SimaR	to eat	*kaən
hand	*qalima	to fall	*-tuq
head	*qulu	to fart	*qətut
hold in fist	*kəmkəm	to plant	*mula
leech	*məCak	to sell	*saliw
lend	*Səzam	to transplant	*Canəm
louse	*kuCu	tongue	*Səma
mist; cloud	*Rabun	tooth	*[ŋ/n/l]ipən
moon	*bulaN	two	*duSa
navel	*puja	uncertainty marker	*nu
nine	*siwa	you	*kamu
nose	*[i/u]juŋ		

4 Notes on the vowel correspondences

With the establishment of a discrete list of AT comparisons, the discussion may now turn to the issue of correspondences between AN and KD. Since Ostapirat (2005) has done much work on the correspondence between consonants, the following section will mostly focus on correspondences

between the vowels. Importantly, it is shown that there are regular reflexes of final-syllable vowels which appear to be conditioned by both the quality of the preceding vowel (assuming that PAT was disyllabic, see Ostapirat 2018), and by the coda.

4.1 The low-vowel *a

Although Ostapirat (2005) focused on consonant correspondences between AN and KD, he does make a specific statement on vowel correspondences in Ostapirat (2013). In that study, it was suggested that PKD distinguished between *a: and *ɤ:, and that these vowels both correspond to PAN *a. Under this hypothesis *a: and *ɤ: merged in PAN but remained distinct in KD. In Tai, however, differential reflexes of *a appear to be conditioned by the height of the preceding penultimate vowel. Where *a is preceded by a high vowel, it has a high-vowel reflex in Tai, either *u or *uə. Where *a is preceded by a non-high vowel, it has a low-vowel reflex in Tai, *a:. Examples are organized in Tables 7 and 8. Table 7 shows high-vowel reflexes in Tai, and Table 8 shows low-vowel reflexes.

*Table 7: High-vowel reflexes of *a in Tai*

	moon	derris root	vomit	boat	navel	bear	hand
PAN	*bulaN	*tubah	*utaq	*aluja (paddle)	*puja	*Cumay	*qalima
PT	*ɓluən ^A	*ɕ.buə ^A	*rɯuək ^D	*ɕ.rɯuə ^A (boat)	*ɗwɯ:ː ^A	* ^h mwɯj ^A	*mwɯ:ː ^A
Siamese	duən ^{A1}	buə ^{A1}	ra:k ^{DL2}	ruə ^{A2}	du:ː ^{A1}	mi:ː ^{A1}	mu:ː ^{A2}
Sapa	buən ^{A1}	bu ^{A1}	haʔ ^{DL2}	hu ^{A2}	du ^{A1}	mi ^{A1}	mu ^{A2}
Bao Yen	buən ^{A1}	buə ^{A1}	ra:k ^{DL2}	luə ^{A2} -i	-	mi:ː ^{A1}	mu:ː ^{A2}
Cao Bang	buən ^{A1}	buə ^{A1}	ra:k ^{DL2}	luə ^{A2}	duə ^{A1}	mi ^{A1}	mu ^{A2}
Lungchow	bɯ:n ^{A1}	bu:ː ^A	la:k ^{DL2}	lu:ː ^{A2}	-	mi:ː ^{A1}	mu:ː ^{A2}
Shangsi	bun ^{A1}	-	luk ^{DL2}	lu ^{A2}	-	muj ^{A1}	moj ^{A2}
Yay	duən ^{A1}	-	ruək ^{DL2}	ruə ^{A2}	duə ^{A1}	muəj ^{A1}	fuɯj ^{A2}
Saek	bliən ^{A1}	viə ^{A1}	ruək ^{DL2}	ruə ^{A1}	duə ^{A1}	muəj ^{A1}	mu:ː ^{A2}

*Table 8: Low-vowel reflexes of *a in Tai*

	otter	eye	die	leg	afraid
PAN	*Sanaq	*maCa	*m-aCay	*paqa	*talaw
PT	*na:k ^D	*p.ta:ː ^A	*p.ta:j ^A	*p.qa:ː ^A	* ^h la:w ^A
Siamese	na:k ^{DL2}	ta:ː ^{A1}	ta:j ^{A1}	k ^h a:ː ^{A1}	-
Sapa	naʔ ^{DL2}	ta:ː ^{A1}	ta:j ^{A1}	xa:ː ^{A1}	-
Bao Yen	na:k ^{DL2}	p ^h ja:ː ^{A1}	p ^h a:j ^{A1}	k ^h a:ː ^{A1}	la:w ^{A1}
Cao Bang	na:k ^{DL2}	t ^h a:ː ^{A1}	t ^h a:j ^{A1}	k ^h a:ː ^{A1}	la:w ^{A1}
Shangsi	na:k ^{DL2}	t ^h a:ː ^{A1}	t ^h a:j ^{A1}	ha:ː ^{A1}	-
Yay	na:k ^{DL2}	ta ^{A1}	ta:j ^{A1}	ka ^{A1}	la:w ^{A1}
Saek	na:k ^{DL2}	pra:ː ^{A1}	pra:j ^{A1}	kwa:ː ^{A1}	la:w ^{A1}

It also appears that a palatal consonant in the onset of a final syllable may trigger high-vowel reflexes in Tai, for example, PAN *Səzam [sə'dʒam] → PT *ʒu:m^A ‘to borrow’. There is at least one exception to this, PAN *aNak ‘child’ which corresponds with PT *lu:k. Other than this single exception, the above conditions play out in numerous comparisons.

Outside of Tai, reflexes of *a are less consistently conditioned by the preceding vowel. Some words which have regularly conditioned high-vowel reflexes of *a in Tai, such as reflexes of *lima ‘five’ and *aluja ‘paddle; boat’ have low-vowel reflexes in Hlai, *ma:* and *ra:* respectively. However, *a is consistently reflected with a long vowel in subgroups outside of Tai, even where Tai has a high-vowel reflex: PAN *Sanaq : PH *əna:k : Paha *na:k*¹¹.

Following Ostapirat (2018), who reconstructs fully disyllabic PKD, I view the split of *a to *a: and *ɤ: as a change that arose in parallel developments after the breakup of PKD which occurs through interactions between penultimate and ultimate syllable vowels. In Tai, there was a regular raising of *a after a penultimate high vowel. In other subgroups, raising is less regular, but this attests to the parallel nature of these developments.

Another piece of evidence that the split of *a in KD subgroups is a parallel innovation is that outside of Tai, it is common for the penultimate vowel features to spread to the onset of the final syllable, if the penultimate vowel is high (Ostapirat 2018). For example, POB *zuaŋ : PAN *qudaŋ ‘shrimp’, POB *duak : PAN *utaq ‘vomit’. In these two examples, the high-back vowel *u has spread its features onto the final-syllable vowel and later deleted. Penultimate vowel feature transfer is apparent in some Tai reflexes, for example, *rwuək^D ‘to vomit’, but not others, for example, *Ç.6uə^A ‘fish poison’. High vowels therefore had an effect on final-syllable vowel reflexes, but these are not uniform across subgroups, suggesting parallel development.

To summarize, the development of *a in final syllables is complex. In neutral environments where the penultimate vowel was non-high, *a consistently lengthens to *a:. Where there is a high vowel in the penultimate syllable, however, Tai has a regular raising of the final vowel, but other subgroups have a mixture of lengthening (*a > *a:) and raising/feature transfer (iCa > Cia/Ci and *uCa > Cua/Cu). It is therefore likely that PAT had a single low vowel *a, which first underwent lengthening in PKD *a > *a: and then later underwent a series of parallel splits typically conditioned by the height of preceding penultimate syllable vowels which co-occurred with the eventual loss of the penultimate syllable in most KD branches.

4.2 Reflexes of *ə

The central vowel schwa also undergoes a split, but reflexes of schwa do not show a clear condition. In both KD and AN, a mid/low central vowel shares irregular reflexes with the high back vowel, *u, both apparently from PAT *ə. To begin, PAN *ə typically corresponds to PT, PH, POB, Lakkja /a/ and PK and Kam /ə/ (/ɐ/ in S. Kam, /ə/ in N. Kam). These are shown in Table 9 (only reflexes of the vowels are listed, while the words themselves are in the Appendix).

*Table 9: Typical correspondence sets involving PAT *ə*

PAn	PT	PH	POB	PK	Lakkja	S Kam	
ə	a	-	-	ə	-	-	*tanəm ‘plant’
ə	a	-	-	-	a	-	*kəmkəm ‘hold in fist’
ə	a	a	a	ə	a	ɐ	*dəmdəm ‘black’
ə	a	a	-	ə	a	ɐ	*ipən ‘tooth’

It is therefore assumed PAT *ə became PAN *ə, PKD *ə, PT *a, PH *a, POB *a, PK *ə, Lakkja a, S Kam ɐ. There are several examples, however, where PAN *ə corresponds unexpectedly to a back or central vowel. The unexpected reflexes are highlighted in Table 10.

*Table 10: Irregular correspondence sets involving pat *ə and KD reflexes*

PAn	PT	PH	POB	PK	Lakkja	S Kam	
ə	-	u	-	ə	-	-	*ənəm ‘six’
ə	ɤ	-	-	-	u	ɐ	*qasəm ‘sour’
ə	ɤ	-	ə	u	a	ɐ	*datəŋ ‘arrive’

These irregularities cross over into AN as well, where there are several cases of PAN *u irregularly corresponding to what are typically reflexes of *ə in KD. Two examples are organized in Table 11, and additional reflexes of the final vowel in *manuk ‘bird’ are included to demonstrate regular reflexes of *u in final syllables. Once again, unexpected reflexes are highlighted, assuming that these all reflect a PAT central vowel.

Table 11: Irregular correspondence sets involving PAT *ə and AN reflexes

PAn	PT	PH	POB	PK	Lakkja	S Kam	
u	a	a	a	-	u	ɐ	*daNum ‘water’
u	ɯ	-	-	-	ə	ə	*Rabun ‘sky; cloud’
u	o	-	u	o	o	o	*manuk ‘bird’

One solution to this problem is to posit two separate phonemes: *ə (= regular reflexes) and *ɯ (= irregular reflexes). This solution remains problematic, however. If there really were two phonemes, *ə and *ɯ, and *ɯ had a tendency to irregularly merge with *ə, then it is possible that there are cases where *ɯ has merged with *ə in all extant languages, making accurate reconstruction impossible. Another possibility is that irregularities in reflexes of *ə all arise from a single phoneme, which typically became *ə but in some cases irregularly became *u. In this scenario, it is more beneficial to reconstruct schwa as a high-vowel like [i] or [ɯ], which may more naturally split to both *ə and *u. Irregular developments of high central vowels are not uncommon. The same tendency for central and high-back vowels to irregularly interact is apparent in Land Dayak languages. Reflexes of PMP *ə are reconstructed as a high vowel in Proto-Land Dayak, [i] (Smith 2019), but reflexes are often irregular, resulting in reflexes of *a*, *ə*, *u*, and *∅* in Land Dayak language Bistaang. Some examples from Bistaang are organized in example 7.

7	Bistaang (Land Dayak, Rensch et al. 2012)		
	PMP	Bistaang	Change
	*bətias > *bətis >	bates ‘calf of the leg’	*ə > a
	*zəlaq >	jara? ‘tongue’	*ə > a
	*ləsuŋ >	rsoŋ ‘mortar’	*ə > ∅
	*təluR >	turoh ‘egg’	*ə > u
	*bulu >	bluh ‘body hair’	*u > ∅
	*silu >	srəh ‘finger nail’	*u > ə

With regard to PAT *ə, because both *ə and *u reflexes are found in AN, a pre-PAN stage where *ə had not yet undergone this split is required. Merger with *u in some words may therefore arise after pre-PAN but before PAN, since the irregularities with *ə and *u are only visible in AN through comparison with KD. That is to say, within AN itself, there is no evidence that some instances of the vowel *u may ultimately be from a more ancient central vowel. A schematic is given in example 8 which shows the various stages of schwa development.

8	PAT *ə ([i] or [ɯ]) → pre-PAN *ə ([i ~ ɯ]) → PAN *ə / *u
---	--

In KD, the same irregular changes happen, but only after PKD began to diversify, since the irregularities are apparent in KD-internal comparisons. The fluidity between more schwa-like realizations of *ə and more u-like realizations persisted into PKD.

The PAT central vowels are therefore of two types. The low vowel *a, which underwent lengthening in PKD and further developed splits in reflexes conditioned by penultimate vowels in many KD daughter languages, and the high-vowel, which is for now written as schwa *ə, which stabilized in AN but again underwent splits in KD which are explained as arising from the instability of high-central vowels and their tendency to undergo unconditioned changes.

4.3 Diphthongs

In most KD branches, AN word-final diphthongs *aw and *ay correspond to identical diphthongs in KD with additional vowel lengthening, *aw : *a:w and *ay : *a:j, but the presence of a high-vowel in the penult had similar effects in diphthongs as elsewhere, resulting in regular high-vowel reflexes in Tai and a tendency towards high-vowel reflexes in other KD branches. In example set 9, reflexes of *talaw,

*qajaw, and *m-aCay reflect lengthening of *a, whereas reflexes of *Cumay have high-vowel reflexes, or have transferred features of the penult directly onto the final syllable vowel.

- 9 *talaw ‘afraid’, PT *^hla:w^A, Qabiao *laau*²⁴
 *qajaw ‘star’, PT *t.na:w^A, PH *ara:u^A, Mak *ʔda:u*¹
 *m-aCay ‘to die’, PT *p.ta:j^A, POB *dā:j
 *Cumay ‘bear’, PT *^hmwuuj^A, PH *mui^A, S. Kam *me*⁵⁵, Lakkja *kũ:i*

4.4 High-vowels

In final syllables, KD high vowels often reflect a set of conditioned splits which depend on the features of the now-lost penultimate syllable, with an additional condition triggered by the presence or absence of a final consonant. In many subgroups, high vowels break into diphthongs with a lowered nucleus and high front or back off-glides. In Tai, for example, this development is regular. Tai maintains distinctions between reflexes of *aw and *u, and between *ay and *i, however, in the length of the syllable nucleus. For example, although *u becomes *aw*, it does not merge with *aw, which becomes *a:w*. Reflexes of final high vowels in Tai are shown in Table 12.

Table 12: Tai reflexes of high vowels in open final syllables

	‘I’	‘louse’	‘this’
PAN	*aku	*kuCu	*ni
PT	*kaw ^A	*traw ^A	*naj ^C
Bao Yen	kɤw ^{A1}	h ɤw ^{A1}	naj ^{C2}
Cao Bang	kɤw ^{A1}	t ^h ɤw ^{A1}	nɤj ^{C2}
Lungchow	kaw ^{A1}	haw ^{A1}	naj ^{B2}
Shangsi	kaw ^{A1}	t ^h aw ^{A1}	noj ^{C2}
Yay	ku ^{A1}	raw ^{A1}	ni ^{C2}
Saek	ku: ^{A1}	raw ^{A1}	ni: ^{C2}

In Hlai and Ong Be, diphthongization is also attested, but in at least PH and PK, reconstructions maintain the monophthong, and diphthongization does not occur in all languages. Some examples are listed in 10.

- 10 *aku ‘I; me’ - PH *aku > Hlai *hou*¹, Ong Be *hau*², PK *ku > Paha *ku*³²², Pubiao *kau*^{A1}, S. Kam *jau*²¹²
 *ni ‘this’ - PH *ni > Hlai *nei*², Ong Be *nia*², PK *ni > Paha *ni*⁵⁵, Pubiao *nai*^{C2}, S. Kam *nai*³³, Lakkja *ni*²³¹

In closed final syllables, a lowering of high vowels can also be observed in Tai (*i > e, *u > o). There are not as many examples of high vowels in closed final syllables as elsewhere, so these observations are tentative, but they do repeat themselves in a number of comparisons. All examples of high-vowel lowering in closed final syllables are associated with voiceless stop codas. It is not clear if these generalizations can be applied to words with voiced codas. Tai reflexes are organized in Table 13.

Table 13: Tai reflexes of high vowels in closed final syllables

	‘bird’	‘painful’	‘hook’
PAN	*manuk	*sakit (sick)	*kabit
PT	*ᶑ.nok ^D	*ke:t ^D (hurt)	*ᶑet ^D
Siamese	nok ^{DS2}	-	bet ^{DS1}
Bao Yen	nok ^{DS2}	-	bɔt ^{DS1}
Cao Bang	nok ^{DS2}	-	bet ^{DS1}
Lungchow	nuk ^{DS2}	-	bit ⁵
Shangsi	-	ket ^{DL1}	-
Yay	rok ^{DS2}	-	-
Saek	nɔk ^{DS2}	ke:t ^{DL1}	-

Outside of Tai lowering in closed final syllables can be observed in some languages, especially in Kra, but elsewhere high-vowel lowering is generally not a regular process. Examples are listed in 11.

- 11 ***-tuq ‘to fall’ (root)** - PH *ʔtuk, PK *tok^D, S. Kam tok⁵⁵
 ***qətut ‘to fart’** - PH *(?)tu:t, POB *düt,
 ***sakit ‘sick, in pain’** - Lakkja tse:t²⁴, S. Kam ʔit³²³

Despite regular lowering in Tai, high vowels remain high in final syllables if conditioned by the presence of a high vowel in the reconstructed penult. For example, reflexes of *qudip ‘life; alive; raw’, are high in most groups, including Tai, although Ostapirat 2000 reconstructs PK *(k)ᶑep^D, again, indicating that the affects that high-vowels in penultimate syllables had on final-syllable vowels are not uniform across the family. Examples are organized in 12.

- 12 ***qudip ‘alive; raw’** - PT *ᶑ.dip^D, PH *uri:p > Hlai ri:p, POB *zip, PK *(k)ᶑep^D

It is important to make these correspondences between various subgroups and possible AT etyma clear, since the major task in AT studies is still determining the number of supported comparisons between the two families. A better understanding of regular sound correspondences allows those working in AT to rule out comparisons which may be only superficially similar, strengthening the core set of comparisons. Also, the more that is understood about correspondences, the better we will be able to spot potential comparisons that might have previously gone overlooked. The vowel correspondences discussed above are summarized below:

- PAT *a underwent lengthening in KD, becoming *a:. Further, in most branches the quality of the penultimate vowel influences reflexes of *a:. In Tai, high-vowel penults result in a raising of reflexes of *a: to *u and in other subgroups the features of the penultimate vowel are often transferred to the final-syllable vowel.
- PAT *ə underwent an unconditioned split which manifests itself in a mismatch between AN and KD words. Typically, PAN *ə corresponds to *a in most KD branches, and *ə in PK. Irregular correspondences have PAN *ə and *u corresponding to KD u, u, ə, and a. It is hypothesized that PAT *ə was a high vowel [i], which then underwent a split.
- Diphthongs, *aw and *ay, developed as *a. They lengthened but were also affected by the presence of a high vowel in the reconstructed penult.
- The high vowels *i and *u did not change in final syllables between PAT, PAN, and PKD, although in several KD branches, word-final high vowels underwent regular diphthongization, *i > aj and *u > aw, and lowered in closed final syllables to mid vowels in several branches.

5 Conclusion

As counted in this paper, there are just over 70 reasonably well-attested and supported comparisons between AN and KD which may descend from a common ancestor PAT. Comparisons meet the conditions that (1) they are reconstructable to at least one primary branch in each family, (2) they are

regular in regards to sound correspondences, and (3) KD monosyllables correspond to the final syllable in AN, assuming that KD monosyllabicity was driven primarily by the reduction and eventual deletion of an unstressed penultimate syllable. The numbers derive largely from previous works, but with 14 additional new comparisons.

With an increase in comparisons between AN and KD comes an increase in our understanding of the sound correspondences which exist between the two families. Although irregularity certainly exists, there is a high amount of regularity as well. The vowel correspondences, of particular interest in the present research, show typical developments. There was likely a PAT low-vowel *a, which remained unchanged in PAN but underwent lengthening in PKD, yielding *a:, and was affected by the presence of high vowels in PAT penultimate syllables. The central vowel, PAT *ə, yields unpredictable reflexes both in KD and AN. It is evident that PAT *ə was a high-central vowel, which became either *ə or *u in PAN, and *a in most KD branches but is also reflected by high and mid back-vowels with irregularity.

Lengthening of *a also occurs in the diphthongs, where *aw and *ay become *a:w and *a:y respectively in KD. In many subgroups, this lengthening leads to a merger avoidance, since many languages have high-vowel breaking in open final syllables (Tai, for example, in which high-vowel breaking is regular). High vowels also undergo lowering to mid vowels where they appear in closed syllables in Tai, as well as some other subgroups, like Kra. This lowering, however, is interrupted by the presence of a high vowel in the reconstructed penult and is only attested in words that end in a voiceless stop. More comparisons are undoubtedly needed to make more concrete descriptions on high-vowel development. When compared to AN, at least at the PAN and PMP levels, the vowels change much less frequently. Vowel breaking, high vowel lowering, and the effects of high-vowel penultimate syllables on reflexes of final-syllable vowels are present in AN, but not at first-order proto-languages.

The AT Hypothesis remains a tentative hypothesis, although the evidence in its favor continues to grow. The evidence for a special relationship between AN and KD is both of a higher quality and quantity now than any time in the past, and it is hoped that more research in the area will help us understand the precise nature of this relationship.

References

- Benedict, Paul K. 1942. Thai, Kadai, and Indonesian: A new alignment in South-Eastern Asia. *American Anthropologist* 44:576-601.
- Benedict, Paul K. 1975. *Austro-Thai language and culture, with a glossary of roots*. New Haven: HRAF Press.
- Benedict, Paul K. 1990. *Japanese/Austro-Tai*. Ann Arbor: Karoma.
- Blust, Robert. 1988. *Austronesian root theory: an essay on the limits of morphology*. Amsterdam, John Benjamins.
- Blust, Robert. 2014. Some recent proposals concerning the classification of the Austronesian languages. *Oceanic Linguistics* 53.2: 300-391.
- Blust, Robert and Stephen Trussel. ongoing. *The Austronesian Comparative Dictionary*. Online: www.trussel2.com/acd/
- Brandstetter, Renward. 1916. *An introduction to Indonesian linguistics*. Trans. by C.O. Blagden. Monograph XV. London: The Royal Asiatic Society.
- Chen, Yen-ling. 2018. *Proto-Ong-Be*. Doctoral Dissertation, Department of Linguistics, University of Hawai'i at Mānoa.
- Dempwolff, Otto Chr. 1937. *Vergleichende Lautlehre des austronesischen Wortschatzes. Deduktive Anwendung des Urindonesischen auf austronesische Einzelsprachen*, Vol 2. Berlin: Reimer.
- Fan, Wenjia. 2019. *A Grammar of Lakkja, South China*. Doctoral Thesis, The University of Melbourne, Asia Institute.
- Li, Hui, Bo Wen, Shu-Juo Chen, Bing Su, Patcharin Pramoonjago, Yangfan Liu, Shangling Pan, Zhendong Qin, Wenhong Liu, Xu Cheng, Ningning Yang, Xin Li, Dinhbinh Tran, Daru Lu, Mu-Tsu Hsu, Ranjan Deka, Sangkot Marzuki, Chia-Chen Tan, and Li Jin. 2008. Paternal

- genetic affinity between western Austronesians and Daic populations. *BMC Evolutionary Biology* 8:146. <https://doi.org/10.1186/1471-2148-8-146>
- Li, Jinfang and Yongxian Lou. 2010. The Buyang language of South China: grammatical notes, glossary, texts and translations. Canberra: Pacific Linguistics.
- Long, Y., Zheng, G., and Geary, D. N. 1998. *The Dong language in Guizhou Province, China*. Summer Institute of Linguistics and the University of Texas at Arlington publications in linguistics, publication 126. Dallas, TX: Summer Institute of Linguistics.
- Ostapirat, Weera. 2000. Proto-Kra. *Linguistics of the Tibeto-Burman Area* 23.1:1-251.
- Ostapirat, Weera. 2004. Proto-Hlai Sound System and Lexicons. In *Studies on Sino-Tibetan Languages: Papers in Honor of Professor Hwang-cherng Gong on His Seventieth Birthday*. Edited by Ying-chin Lin, Fang-min Hsu, Chun-chih Lee, Jackson T.-S. Sun, Hsiu-fang Yang, and Dah-an Ho. Institute of Linguistics. Academia Sinica, Taipei, Taiwan: 121-175.
- Ostapirat, Weera. 2005. Kra-Dai and Austronesian: notes on phonological correspondences and vocabulary distribution. In *The peopling of East Asia: putting together archaeology, linguistics and genetics* ed. by, Laurent Sagart, Roger Blench, and Alicia Sanchez-Mazas. 107-131. London: Routledge Curzon.
- Ostapirat, Weera. 2013. Austro-Tai revisited. Plenary presentation at the 23rd annual meeting of the Southeast Asian Linguistics Society (SEALS), May 29-31, Chulalongkorn University.
- Ostapirat, Weera. 2018. Reconstruction disyllabic Kra-Dai. Plenary presentation at the 28th meeting of the Southeast Asian Linguistics Society (SEALS 28), 17–19 May, 2018, Wen-zao Ursuline University of Languages, Kaohsiung, Taiwan.
- Pittayaporn, Pittayawat. 2009. *The phonology of Proto-Tai*. Doctoral Dissertation, Department of Linguistics, Cornell University.
- Reid, Lawrence. 1984-1985. Benedict's Austro-Tai hypothesis — an evaluation. *Asian Perspectives* 26.1: 19-34.
- Rensch, Calvin. R., Carolyn M. Rensch, Jonas Noeb, and Robert S. Ridu. 2012. *The Bidayuh language: yesterday, today, and tomorrow (revised and expanded)*. Number 33 in SIL E-Books. Kuching: SIL International.
- Sagart, Laurent. 2004. The higher phylogeny of Austronesian and the position of Tai-Kadai. *Oceanic Linguistics* 43.2:411–444.
- Sagart, Laurent. 2005. Tai-Kadai as a subgroup of Austronesian. In *The peopling of East Asia: putting together archaeology, linguistics and genetics* ed. by, Laurent Sagart, Roger Blench, and Alicia Sanchez-Mazas. 177-181. London: Routledge Curzon.
- Sagart, Laurent. 2010. Why does the Tai-Kadai word for 'ten' end in -t? Paper presented at 22nd Paris Meeting on East Asian Linguistics (JLAO), Paris, 1-2 July 2010.
- Schmidt, Wilhelm. 1906. *Die Mon-Khmer Völker, ein Bindeglied zwischen Völkern Zentralasiens und Australasiens*. Braunschweig: Friedrich Vieweg und Sohn.
- Smith, Alexander D. 2019. A second look at Proto-Land Dayak vowels. *Oceanic Linguistics* 58.1:110–142.
- Thurgood, Graham. 1994. Tai-Kadai and Austronesian: the nature of the historical relationship. *Oceanic Linguistics* 33(2): 345–368.

Appendix 1: data sets from Kra-Dai with proto-Austronesian comparisons

afraid: PAN *talaw, PT *^hla:w^A, Bao Yen la:w^{A1}, Cao Bang la:w^{A1}, Lungchow la:w^{A1}, Yay la:w^{A1}, Saek la:w^{A1}, Qabiao laau²⁴, Buyang i¹¹vau²⁴, S. Kam jau¹³.

arrive PAN datəŋ PT *C.təŋA Siamese t^huŋA1 Sapa t^huŋA Bao Yen t^huŋA1 Cao Bang t^huŋA1 Lungchow t^huŋA1 Shangsi Yay taŋA2 Saek t^haŋA2 PH Hlai POB dəŋ A1 Ong Be Gelao Paha khau³³ Pubiao Qabiao Buyang Lakkja taŋ231 (come) Mak S. Kam təŋ55 (come) N. Kam təu25

- bear:** PAN *Cumay, PT *^hm^wu^jA, Siamese mi:^{A1}, Sapa mi^{A1}, Bao Yen mi:^{A1}, Cao Bang mi^{A1}, Lungchow mi:^{A1}, Shangsi m^yA¹, Yay m^wu^jA¹, Saek m^wu^jA¹, PH *m^{ui}A, Paha mi³²², Pubiao m^hj^eA², Lakkja ku^β:i, S. Kam me⁵⁵.
- bird:** PAN manuk-manuk, PT *^ç.nok^D, Siamese nok^{DS2}, Sapa nuk^{DS2}, Bao Yen nok^{DS2}, Cao Bang nok^{DS2}, Lungchow nuk^{DS2}, Yay rok^{DS2}, Saek n^{ok}DS², PH *sa^C, Hlai tatt⁷, POB nuk^{D2} Ong Be nok⁸ Gelao ma³³no³⁵, Paha nok¹¹, Pubiao nok^ηA², Qabiao nuk⁴⁵/niuk⁴⁵, Buyang ma⁰nuk¹¹, Lakkja mlok⁷, Mak nok⁸, S. Kam mok²¹, N. Kam no^ʔ13.
- black:** PAN *d^əm^də^m, PT *^ç.dam^A, Siamese dam^{A2}, Sapa dam^{A2}, Bao Yen dam^{A2}, Cao Bang dam^{A2}, Lungchow dam^{A2}, Shangsi nam^{A1}, Saek ram^{A1}, PH *([?])dam^C, Hlai dom³, POB *zam^{A1}, Ong Be lam¹, Paha lam³²²/dam³¹, Pubiao [?]dam^{A1}, Qabiao dam⁵³, Buyang [?]dam³¹, Lakkja lam⁵¹, Mak nam¹, S. Kam n^{əm}55, N. Kam n^{əm}35.
- blow:** PAN *Səyup, Buyang hip⁵⁴, Lakkja jəp²⁴, S. Kam səp²¹, N. Kam səp¹¹.
- boat:** PAN *aluja ‘to paddle’, PT *^ç.r^wu^əA ‘boat’, Siamese ru^əA², Sapa hu^rA², Bao Yen lu^əA²⁻¹, Cao Bang lu^əA², Lungchow lu:^{A2}, Shangsi lu^{A2}, Yay ru^əA², Saek ru^əA¹, PH *ura^A ‘boat’, Paha da³³, S. Kam lo⁵⁵, N. Kam la³⁵.
- chaff; bran:** PAN *qəpəh, POB v^ə:BC², Gelao pau^{B1}, Buyang faa^{B1}, Lakkja faa^{B1}, N. Kam pa³³.
- child:** PAN *aNak, PT *lu:^kD, Siamese lu:^kDL², Sapa lu^{ɔ̄}DL², Bao Yen luk^{DL2}, Cao Bang luk^{DL2}, Lungchow luk^{DL}, Shangsi lak^{DL1}, Yay luuk^{DS2}, Saek luuk^{DL1}, PH *ali:k, Hlai luuk⁷lau², POB l^ə:k^{D2}, Ong Be lək⁸, Gelao la³⁵lie⁵⁵, Pah laak¹¹, Buyang la:k¹¹, Mak lak⁸, S. Kam lak³¹ʔun³²³, N. Kam la^ʔ31ʔun³³.
- cry:** PAN *Cajis, PT *t.haj^C, Siamese ha:j^{C1}, Bao Yen haj^{C1}, Cao Bang haj^{C1}, Lungchow haj^{C1}, Shangsi haj^{C1}, Yay taj^{C1}, Hlai ηai³, POB *ηaj^{BC1}, Ong Be ηai³, Pah ηit¹¹, Buyang ηiet^{D1}, Mak ηe³, S. Kam ne³²³, N. Kam ηe³³.
- derris root:** PAN *tubah, PT *^ç.bu^əA, Siamese bu^əA¹, Sapa bu^əA¹, Bao Yen bu^əA¹ Cao Bang bu^əA¹, Lungchow bu:^A, Saek vi^əA¹.
- die:** PAN *m-aCay, PT *p.ta:j^A, Siamese ta:j^{A1}, Sapa ta:j^{A1}, Bao Yen p^ha:j^{A1}, Cao Bang t^ha:j^{A1}, Lungchow ha:j^{A1}, Shangsi t^ha:j^{A1}, Yay ta:j^{A1}, Saek pra:j^{A1}, POB *d^a:j^{A1}, Ong Be dai¹, Qabiao tie⁵³, Buyang ma⁰t^e54 ‘kill’, Lakkja plei⁵¹, S. Kam t^{əi}55, N. Kam t^{əi}35.
- eat:** PAN *kaən, PT *ku^ŋA, Siamese kin^A, Sapa kin^{A1}, Bao Yen kin^{A1}, Cao Bang kin^{A1}, Lungchow kin^{A1}, Shangsi k^ŋA¹, Yay k^un^{A1}, Saek kin^{A1}, POB k^{ən}A¹, Ong Be k^{ən}1, Gelao ka³¹, Paha kaan³²², Pubiao k^{ən}A¹, Qabiao k(ɰ)ən⁵³, Buyang ka:n⁵⁴, Lakkja tsen⁵¹, S. Kam tan⁵⁵.
- eight:** PAN *walu, PH *aRu^A, Paha mu³¹, Pubiao r^hu^uA², Qabiao m^ə0zu³³, Buyang ^{ɔ̄}uu^{A2}.
- excrement:** PAN *Caqi, PT *^ç.quⁱj^C, Siamese k^hi:^{C1}, Bao Yen k^hi:^{C1}, Cao Bang k^hi:^{C1}, Lungchow k^hi:^{C1}, Shangsi k^hoy^{C1}, Yay haj^{C2}, Saek ^ɣaj^{C2}, PH *aka:i^C, POB ka:^jBC², Paha q^ε33, S. Kam ^ʔe³¹, N. Kam ^ʔe³¹.
- eye:** PAN *maCa, PT *p.ta:^A, Siamese ta:^{A1}, Sapa ta:^{A1}, Bao Yen p^ha:^{A1}, Cao Bang t^ha:^{A1}, Lungchow ha:^{A1}, Shangsi t^ha:^{A1}, Yay ta:^{A1}, Saek pra:^{A1}, PH *ata, Hlai tsha¹, POB d^a:^{A1}, Ong Be da¹, Gelao mu³³tu³¹, Paha ma⁵⁵da³²², Pubiao tee^{A1}, Qabiao te⁵³, Buyang ma⁰ta⁵⁴, Lakkja pla⁵¹, Mak da¹, S. Kam ta⁵⁵, N. Kam ta³⁵.
- father:** PAN *ama(x)/*aba, PT *bo:^B, Siamese p^hɔ:^B, Sapa po^B, Bao Yen p^ɔ:B², Cao Bang ɔ:^{B2}, Lungchow po:^{B2}, Shangsi po^{C2}-t, Yay po^{B2}, Saek p^hɔ:^{B2}, PH *pa^C, Hlai pha³, Ong Be be²lau⁴, Gelao a⁵⁵ba³³, Pah pa³³, Pubiao pe²¹³, Buyang pa¹¹, Lakkja pe⁵⁵, Mak p^{əu}4, S. Kam pu³¹.
- fire:** PAN Sapuy, PT *w^xj^A, Siamese faj^{A2}, Bao Yen p^hxj^{A2}, Cao Bang v^xj^{A2}, Lungchow faj^{A2}, Shangsi foy^{A2}, Yay fi^{A2}, Saek vi:^{A2}, PH *api^A, Hlai fei¹, POB *v^ə:j^{A2}, Ong Be v^{əi}2, Gelao pia³³, Paha pwi³²² Pubiao pei^{A1}, Qabiao p^{əi}53, Buyang pui⁵⁴, Lakkja pu:i¹, Mak v^{əi}1, S. Kam pui⁵⁵, N. Kam wi³⁵.

five: PAN *lima, PH *ama^A, Hlai pa¹, Gelao mluw³¹, Paha ma³³, Pubiao mfaaa^{A2}, Qabiao ma³³, Buyang ma³¹².

flow; current: PAN *qaluR, PT *hlwaj^A, Siamese laj^{A1}, Sapa laj^{A1}, Bao Yen lwɔj^{A1}, Cao Bang lwɔj^{A1}, Lungchow laj^{A1}, Shangsī laj^{A1}, Yay laj^{A1}, POB lə:j^{A1}, Ong Be ləi¹, Gelao hlei³³, Paha qwi³²², Pubiao ləi^{A1}, Qabiao ləi⁵³, Buyang lui^{A2}, Mak lu:i¹, S. Kam ʔui³⁵, N. Kam tui¹¹.

grandfather: PMP *əmpu, PT *puw^B, Siamese pu:^{B1}, Sapa pu^{B1}, Bao Yen pu:^{B1}, Cao Bang pu^{B1}, PH *pau^C.

hand: PAN *qalima, PT *mwu:^A, Siamese mu:^{A2}, Sapa mu:^{A2}, Bao Yen mu:^{A2}, Cao Bang mu:^{A2}, Lungchow mu:^{A2}, Shangsī moy^{A2}, Saek mu:^{A2}, PH *mi^A, Hlai meuw¹, POB mə:^{A2}, Ong Be mɔ:², Gelao pa³¹mi³¹, Paha ma³³qe⁴⁵⁽³²²⁾, Pubiao hmii^{B1}, Qabiao qa⁰hmi²¹³, Lakkja mie²³¹, S. Kam mja²¹², N. Kam mja²².

head: PAN *qulu, PT kraw^C/truə^A, Siamese /huə^{A1}, Sapa /hu¹, Bao Yen huə^{A1}, Cao Bang /t^huə^{A1}, Lungchow /hu:^{A1}, Shangsī law^{C1}/, Yay caw^{C1}/, Saek t^hraw^{C1}/, PH *uRəu^C, Hlai gwou³, POB ha:w^{BC1}, Ong Be hau³, Gelao tɛ:³¹lui³⁵, Buyang qa⁰ɔu¹¹, Lakkja kjeu⁵¹, Mak tau³, S. Kam kau³²³, N. Kam kau³³.

hold in fist: PAN *kəmkəm, PT *kam^A, Siamese kam^{A1}, Sapa kam^{A1}, Bao Yen kam^{A1}, Cao Bang kam^{A1}, Lungchow kam^{A1}, Shangsī kam^{A1}, Yay kam^{A1}, Saek kam^{A1}, Lakkja kam⁵

hook: PAN *kabit, PT *bet^D, Siamese bet^{DS1}, Sapa bit^{DS1}, Bao Yen bɛt^{DS1}, Cao Bang bet^{DS1}, Lungchow bit⁵.

I: PAN *aku, PT *ku:^A/*kaw^A, Siamese ku:^{A1}, Sapa ku^{A1}, Bao Yen /kɔw^{A1}, Cao Bang /kɔw^{A1}, Lungchow /kaw^{A1}, Shangsī /kaw^{A1}, Yay ku^{A1}, Saek ku:^{A1}, PH *aku^A, Hlai hou¹, Ong Be hau², Paha ku³²², Pubiao kau^{A1}, Qabiao kau⁵³, Buyang ku⁵⁴, Lakkja tsi⁵¹, S. Kam jau²¹², N. Kam jau²².

leech: PAN *məCak, PT *da:k^D, Siamese t^ha:k^{DL2}, Sapa ta^ɔ^{DL2}, Bao Yen ta:k^{DL2}, Cao Bang ɖa:k^{DL2}, Lungchow tek^{DL2-v}, Yay ta:k^{DL1}, Saek t^ha:k^{DL2}, PH *ʔta:k, POB ɖa:k^{D1}, Paha ɲa³¹taak³³.

leg; thigh: PAN *paqa ‘thigh’ PT *p.qa:^A ‘leg’, Siamese k^ha:^{A1}, Sapa xa:^{A1}, Bao Yen k^ha:^{A1}, Cao Bang k^ha:^{A1}, Lungchow k^ha:^{A1}, Shangsī ha:^{A1}, Yay ka^{A1}, Saek kwa:^{A1}, Hlai ha¹, Ong Be va², Paha ɣa¹¹, Buyang ʔaa^{A1}, S. Kam pa⁵⁵, N. Kam pa³⁵lau³¹.

lend/borrow: PAN *Səzam, PT *ʔju:m^A, Siamese ju:m^{A1}, Cao Bang jum^{A1}, Lungchow jim^{A1}, Shangsī jom^{A1}, Lakkja lam⁵¹, S. Kam jam⁵⁵, N. Kam jam³⁵.

louse: PAN *kuCu(x), PT *traw^A, Siamese haw^{A1}, Sapa haw^{A1}, Bao Yen hɔw^{A1}, Cao Bang t^hɔw^{A1}, Lungchow haw^{A1}, Shangsī t^haw^{A1}, Yay raw^{A1}, Saek raw^{A1}, PH *utu^A, Hlai fou¹, Paha ɔu³³, Qabiao qa⁰tau⁵³, Buyang qa⁰tu⁵⁴, Lakkja ta:u, Mak tɔ²təu¹, S. Kam tau⁵⁵, N. Kam tau³⁵.

moon: PAN *bulaN, PT *ɓluwən^A, Siamese duwən^{A1}, Sapa buwən^{A1}, Bao Yen buwən^{A1}, Cao Bang buwən^{A1}, Lungchow bɔ:n^{A1}, Shangsī bun^{A1}, Yay duwən^{A1}, Saek bliwən^{A1}, PH *ɲa:n^A, Hlai nyaan¹, Paha naan³³⁽³²²⁾, Pubiao nin^{A1}/taan^{A1}, Qabiao taan⁵³, Buyang luŋ¹¹ten¹¹, Lakkja man¹¹lie:ŋ²¹⁴, Mak ni:n², S. Kam kwan⁵⁵ŋan⁵⁵, N. Kam mjan³⁵.

navel: PAN *puja, PT *dɔwu:^A, Siamese du:^{A1}, Sapa du^{A1}, Cao Bang duə^{A1}, Shangsī boy^{C1-t}, Yay duə^{A1}, Saek duə^{A1}, PH *uri^A, POB *dɔ:^{A2}, Buyang ʔduə^{A1}, S. Kam pjo⁵⁵ljo⁵⁵.

nine PAN *Siwa, Paha dfa³³, Pubiao eja^{B1}, Qabiao mə⁰æia²¹³, Buyang vaa^{B1}.

nose: PAN *ujun/ijun, PT *ɖaŋ^A, Siamese ʔdaŋ^{C1-t}, Sapa daŋ^{A1}, Bao Yen daŋ^{A1}, Cao Bang daŋ^{A1}, Lungchow daŋ^{A1}, Shangsī daŋa^{A1}, Yay daŋ^{A1}, Saek daŋ^{A1}, PH *(ʔ)daŋ^A, POB zəŋ^{A1}, Ong Be lɔŋ¹, Pubiao taŋ^{C1}, Qabiao qa³³taŋ³³, Buyang qa⁰tiŋ²¹³, Lakkja naŋ⁵¹, S. Kam nəŋ⁵⁵, N. Kam naŋ³⁵.

one: PAN *isa, PH *ci^C, Hlai tshew³/tsuw², Ong Be hɔ³, Gelao tsi⁵⁵, Pah ti⁵⁵, Pubiao tɛja^{C1}, Qabiao tæia³³.

otter: PAN *Sanaq, PT *na:k^D, Siamese na:k^{DL2}, Sapa na^ɔ^{DL2}, Bao Yen na:k^{DL2}, Cao Bang na:k^{DL2}, Shangsī na:k^{DL2}, Yay na:k^{DL2}, Saek na:k^{DL2}, PH *(ə)na:k, Paha naak¹¹.

- rattan** PAN *quay, PT *Ç.wa:j^A, Siamese wa:j^{A1}, Sapa va:j^{A1}, Bao Yen wa:j^{A1}, Cao Bang wa:j^{A1}, Lungchow va:j^{A1}, Shangsi wa:j^{A1}, Yay va:j^{A1}, Saek va:j^{A1}.
- raw; unripe; alive:** PAN *qudip ‘alive’, PT *Ç.dip^D ‘raw; uncooked’, Siamese dip^{DS1}, Sapa dip^{DS1}, Bao Yen dip^{DS1}, Cao Bang dip^{DS1}, Lungchow dip^{DS1}, Shangsi dip^{DS1}, Yay dip^{DS1}, Saek rip^{DS1}, PH *uri:p, Hlai ri:p/vi:p, POB *zip^{D1}, Paha ðap³³, Pubiao ?dap^{D1}, Buyang ?a?dip.
- sell:** PAN *baliw/*saliw, PH *aRi:u^C, PK *s-ywi.
- seven:** PAN *pitu, PH *?tu^A, Paha ðu³³, Pubiao tuu^{A1}, Qabiao mə⁰tu⁵³, Buyang tuu^{A2}.
- shadow:** PAN *qaNiŋu, PT *ŋaw^A, Siamese ŋaw^{A2}, Bao Yen ŋw^{A2}, Cao Bang ŋw^{A2}, Lungchow ŋaw^{A2}, Shangsi ŋaw^{A2}, Yay ŋaw^{A2}, Saek ŋaw^{A2}, PH *aŋa:u^C, POB ŋu:j^{A1}.
- shoulder:** PAN *qabaRa, PT *Ç.ba:^B, Siamese ba:^{B1}, Sapa ba:^{B1}, Bao Yen ba:^{B1}, Cao Bang ba:^{B1}, Lungchow ba:^{B1}, Shangsi ba:^{B1}, Yay ba^{B1}, Saek va:^{B1}, PH *ava^B, Hlai tsu²va², POB via^{BC1}, Ong Be bik⁷via³, Paha ka⁰γ⁵⁵ma⁵⁵, Pubiao hmaa^{B1}, Buyang qa⁰?ba¹¹, Mak ha¹, N. Kam pja¹¹.
- shrimp:** PAN *qudaŋ, PH *ura:ŋ^A, POB *zuaŋ^{A2}, Lakkja tsoŋ²⁴, S. Kam toŋ³³.
- sick; pain; to hurt:** PMP *sakit, PT *ke:t^D ‘hurt’, Shangsi ket^{DL1}, Yay cet^{DL1}, Saek ke:t^{DL1}, Paha ði¹¹, Lakkja tse:t²⁴, S. Kam ?it³²³, N. Kam kit³³.
- six:** PAN *ənəm, PH *(ə)num^A, Hlai tom¹, Gelao nam^{A1}, Paha nam³¹, Pubiao hnam^{A1}, Qabiao mə⁰hnam⁵³, Buyang nam²⁴.
- sky/cloud:** PAN *Rabun ‘cloud’, PT *6un^A ‘sky’, Siamese bon^{A1}, Bao Yen bon^{A1}, Cao Bang bɔn^{A1}, Shangsi bon^{A1}, Yay bun^{A1}, Saek bun^{A1}, Buyang ?bun⁵⁴, Lakkja bən⁵¹, Mak ?bən¹, S. Kam mən⁵⁵, N. Kam mən³⁵.
- spotted:** PMP *bəlaŋ, PT *6la:ŋ^B, Siamese da:ŋ^{B1}, Sapa ba:ŋ^{B1}, Bao Yen bja:ŋ^{B1}, Cao Bang da:ŋ^{B1}.
- star:** PAN *qajaw ‘day’, PT *t.na:w^A, Siamese da:w^{A1}, Sapa da:w^{A1}, Bao Yen da:w^{A1}, Cao Bang da:w^{A1}, Lungchow da:w^{A1}, Shangsi da:w^{A1}, Yay da:w^{A1}, Saek tra:w^{A1}, PH *ara:u^A, Hlai raau¹, Lakkja tau²⁴, Mak ?da:u¹?dai⁵.
- sour:** PMP *qəsəm, PT *sɔm^C, Siamese som^{C1}, Bao Yen thom^{C1}, Cao Bang lom^{C1}, Lungchow lum^{C1}, Shangsi lom^{C1}, Yay èam^{C1}, Saek sam^{C1}, Lakkja khjum²⁴, S. Kam səm¹³, N. Kam səm¹³.
- taro:** PAN *biRaq, PT *pruək^D, Siamese p^huək^{DL1}, Sapa p^hu?DL1, Bao Yen p^huək^{DL}, Cao Bang p^huək^{DL1}, Lungchow p^hr:k^{DL1}, Shangsi p^hyk^{DL1}, Yay puək^{DL1}, Paha pwaak¹¹, Buyang ðaak^{D2}, Lakkja ja:k²⁴, S. Kam jak³²³.
- this:** PAN *-ni, PT *naj^C, Siamese ni:^{C2-v}, Sapa ni^{C2}, Bao Yen naj^{C2}, Cao Bang nxj^{C2}, Lungchow naj^{B2}, Shangsi noy^{C2}, Yay ni^{C2-v}, Saek ni:^{C2-v}, PH *ni^B, Hlai nei², Ong Be nia², Gelao nyi³⁵, Paha ni⁵⁵, Pubiao nai^{C2}, Qabiao nai⁴⁵, Buyang ni¹¹, Lakkja ni²³¹, S. Kam nai³³, N. Kam nai⁴⁴.
- three:** PAN *təlu, Paha tu³²², Pubiao tau^{A1}, Qabiao tau⁵³, Buyang tuu^{A1}.
- to fall:** PAN *-tuq, PH *?tuk, Gelao to³⁵, Paha tək⁵⁵, Pubiao tək^{D1}, Buyang tuk^{D1}, S. Kam tok⁵⁵, N. Kam to⁵⁵.
- to fart:** PAN *qətut, PT *k.tɔt^D, Siamese tot^{DS1}, Sapa tut^{DS1}, Bao Yen tɔt^{DS1}, Cao Bang tɔt^{DS1}, Shangsi t^hot^{DS1}, Yay rat^{DS1}, Saek ret^{DS1}, PH *(?)tu:t, POB *dūt^{D1}, Paha ðat⁵⁵, Pubiao tat^{D1}, Buyang tut^{D1}, Lakkja kjə:t³, S. Kam tət⁵⁵.
- to plant:** PAN *mula, PH *uRa^A/*uga^A, Hlai gwa¹, S. Kam mja²¹², N. Kam mja²².
- to transplant:** PAN *tanəm, PT *t.nam^A, Siamese dam^{A1}, Bao Yen dam^{A1}, Cao Bang dam^{A1}, Lungchow dam^{A1}, Shangsi dam^{A1}, Yay dam^{A1}, Saek tram^{A1}, PH *?dap, Paha dam³³wa³²².
- tongue:** PAN *Səma, Gelao du³⁵mau³¹, Paha ma³¹, Pubiao m^hje^{A2}, Qabiao mie³³, Buyang mee^{A2}, Lakkja waP, Mak ma², S. Kam ma²¹², N. Kam ma²².

tooth: PAN *[ŋ/n/l]ipən, PT *wan^A, Siamese fan^{A2}, Bao Yen p^{han}A², Cao Bang van^{A2}, Lungchow fan^{A2}, Yay fan^{A2}, PH *ipan^A, Hlai fan¹, Lakkja wan², S. Kam pjen⁵⁵, N. Kam pjen³⁵.

two: PAN *duSa, Gelao su³¹, Pah èa³²², Pubiao ɛee^{A1}, Qabiao æe⁵³, Buyang ɛa⁵⁴, S. Kam ja²¹², N. Kam ja²².

uncertainty marker: PAN *-nu, Ong Be ləu²/nə²na³, Paha pa³³nau³³, Pubiao njau, Qabiao niau⁴⁵, Buyang nə^{A2}, S. Kam nəu²¹², N. Kam nəu²².

vomit: PAN *utaq, PT *rwuək^D, Siamese ra:k^{DL2}, Sapa ha^{DL2}, Bao Yen ra:k^{DL2}, Cao Bang ra:k^{DL2}, Lungchow ɬa:k^{DL2}, Shangsi luk^{DL2}, Yay ruək^{DL2}, Saek ruək^{DL2}, PH *apa:k, Hlai feek⁷, POB *duak^{D2}, Ong Be duak, Gelao qə³³ta³⁵, Paha taak³³, Buyang ta:u³¹², Lakkja ta:k, Mak du:k.

water: PAN *danum, PT *Ç.nam^C, Siamese na:m^{C2}, Sapa nam^{C2}, Bao Yen nam^{C2}, Cao Bang nam^{C2}, Lungchow nam^{C2}, Shangsi nam^{C2}, Yay ram^{C2}, Saek nam^{C2}, PH *nam^C, Hlai nom³, POB nam^{BC2}, Ong Be nam⁴, Lakkja num¹¹, S. Kam nəm³¹, N. Kam nəm³¹.

you: PAN *kaSu/Simu, PT *muŋ^A/*mau^A, Siamese muŋ^{A2}, Sapa muŋ^{A2}, Bao Yen /mɯŋ^{A2}, Lungchow /mau^{A2}, Shangsi maŋ^{A1-t}, Yay muŋ^{A2}, Saek muŋ^{A2}, PH *mi^A, Hlai meu¹, Ong Be mə², Gelao mu³¹, Paha mə³¹, Pubiao maa^{A2}, Qabiao mi³³, Buyang maa^{A2}, Lakkja ma²³¹.

BASIC CLAUSES IN ITAWIT

Chirbet C. AYUNON Shirley N. DITA
Cagayan State University De La Salle University-Manila
chirbet_ayunon@dlsu.edu.ph shirley.dita@dlsu.edu.ph

Abstract

This paper presents a description of the two main types of clauses observed in a language spoken in Northern Luzon, the Itawit language. It examines the clauses of Itawit using gathered spoken data, including utterances in naturally occurring conversations, spoken-like data and narrative data. Non-verb constituent usually heads a nonverbal clause. This study gives attention to the five types of nonverbal clauses: nominal predicate, adjectival predicate, existential predicate, prepositional predicate, and locative predicate. Verbal clauses, on the other hand, are usually headed by verbs that occupy the initial position in the clause. This analysis gives affirmation to Reid and Liao's (2004) argument that Philippine languages, including Itawit now, typically follow right-branching clause structure. That is, heads of constructions usually appear in the initial position in the constructions. This paper also distinguishes intransitive constructions from transitive constructions.

Keywords: verbal clause, non-verbal clause, transitive construction, intransitive construction

ISO 639-3 codes: itv

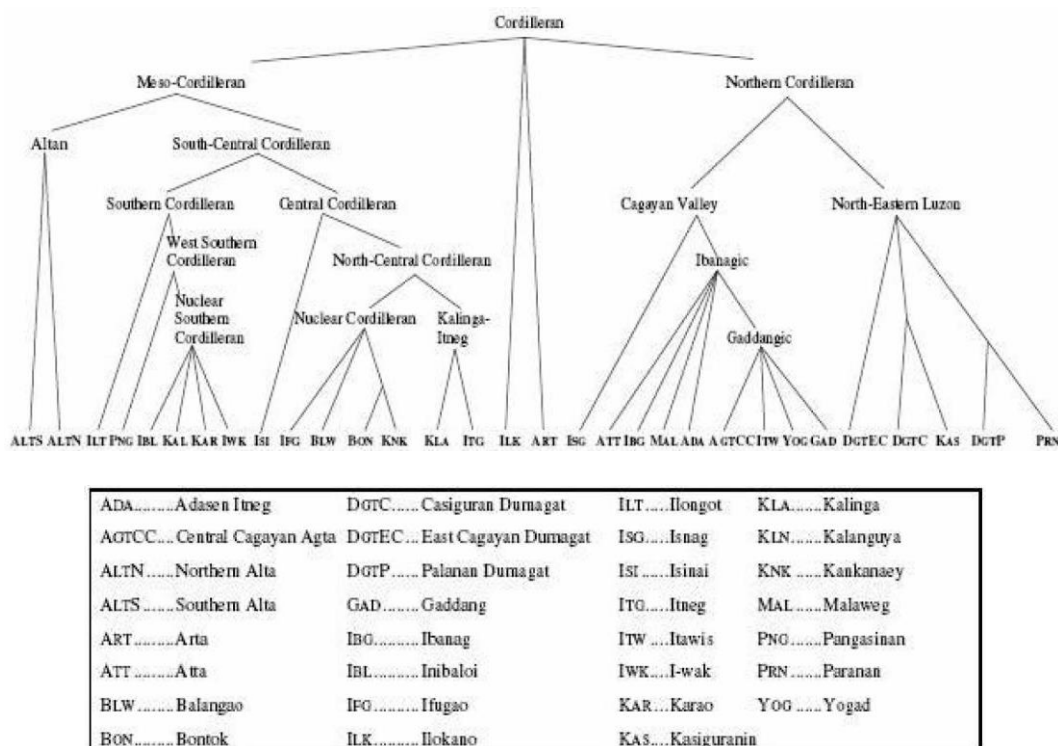
1 Introduction

Itawit is one of the many ethnolinguistic groups in Cagayan, together with Agta, Paranan, Kalinga, Gaddang, Yogad, Bugkalot, Ilocano, Ifugao, Kalinga, Tinguian, and Ibanag. "Itawit" comes from the prefix *i*, meaning "people of" and the word *tawid*, or "across the river," and thus it means "the people from across the river." Simons and Fennig (2018) maintain that the Itawit language is given alternate names Itawes and Itawis. This probably explains the preponderance of the use of the term Itawes, or Itawis in existing literature. However, in the latest edition of *Peoples of the Philippines*, Zafra and Lucero (2017) offered an explanation by saying that their language was spelled "Itawes" since the Spanish colonial period, but they call themselves Itawit because they pronounce letter [s] as [t] when it is the last letter of the word. Aside from that, the Itawit speakers themselves insist that they and their language, be called *Itawit*; hence, this article consistently uses the term Itawit.

Itawit belongs to the Cordilleran subgroup of Malayo-Polynesian branch of the Austronesian language family (Reid (1974, 2006). Cordilleran is comprised of the Central group, the Southern Cordilleran subgroup, and the Northern Cordilleran subgroup. Northern Cordilleran consists of at least the following languages: Ibanag, Gaddang, Yogad, Isnag, Malaweg, Itawis (also called Itawit), Ilokano, and the languages of the various Negrito groups of Cagayan, Isabela, and Aurora provinces, labeled variously as Agta, Atta, and Dumagat.

Itawit is classified as a member of the Cagayan Valley sub-group of the Northern Cordilleran group of Northern Luzon, Philippines, as shown in Figure 1. Ibanag, Ga'dang, Northern Cagayan Agta, Attam, Yogad, and Isnag are also members of the Cagayan Valley sub-grouping.

Figure 1: The revised sub-grouping of Cordilleran languages (Reid 2006:4)



In terms of the status of endangerment or development, Ethnologue labels Itawit as a developing language, following Lewis and Simons (2010) Expanded Graded Intergenerational Disruption Scale or EGIDS, a tool that is used to measure the status of a language in terms of endangerment or development. Labeled as such, a developing language is a ‘language that is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.’

Spoken by nearly 189,000 speakers, as seen in Ethnologue, Itawit as a language, is inadequately described as there are very few studies on the description of the grammar of the language. What may be the earliest work on Itawit is the publication of Jalojot (1937) entitled *Diskripsyon ng Klos na Verbal ng Wikang Itawit*. Next was probably Natividad and Solomon’s (1970) list of phrases and clauses in Itawis then an Itawit wordlist consisting of 372 lexical items as unpublished language data collected by SIL International in the Philippines in 1976. This was followed by Dita (2013), who wrote an Itawit-Filipino Dictionary. A few authors have delved into comparing some linguistic features of Itawit with other languages. One was the work of Bollas and Hernandez (2013), which presented the phonology of Tagalog, Cebuano and Itawis by introducing the languages’ phonemic inventories, contrastive pairs and their phonotactics. In the same year, after having realized that existence of modest records on pluralization especially in Ayta Mag-antsi and Itawit, Bollas and Supnet (2013) delved into a comparative analysis of pluralization in four Philippine languages: Ayta Mag-antsi, Itawit, Bikol and Tagalog. The results show that Itawit and Ayta Mag-antsi have unique pluralizing morphemes, especially in forming proper and common nouns, as compared to Bicol and Tagalog. In 2014, Bollas and Hernandez focused on anaphors and anaphoric relations of these anaphors expressed in Tagalog, Hiligaynon and Itawit using Principle A of the Government and Binding theory. MacKenzie (n.d.) provided an overview of reduplication in Itawit and argued that several reduplicative templates are available in Itawit. His work demonstrated that reduplication in Itawit shows a resistance to destroying syllable contact. Another linguistic work that focused on morphological processes of Ibanag and Itawit was written by Elli and Isidro in 2013. Their study showed that both languages seem to be quite closely related because the morphological processes are almost the same. Itawis, as argued, has fewer morphological processes in nouns and verbs than Ibanag because it does not have full reduplication in nouns, stress placement and added morphemes in verbs. It was in 2016 when the Department of Education of Region 02, in close coordination with the *Komisyon sa Wikang Filipino*, drafted and

released an Itawit orthography, which was written by teachers who are speakers of the language and has been validated by a group of elders who served as consultants.

Considering the above-mentioned works, very few linguistic features of Itawit have been studied, which leads us to the observation that Itawit is relatively less studied among the Philippine languages. Hence, this paper attempts to present a descriptive analysis of the basic clauses in Itawit. The main aim of this paper is to present an analytic description of the clauses in Itawit as seen in the gathered spoken data and as attested by the chosen language informants.

2 Methodology

The data samples on which the descriptions and analyses are based are composed of transcribed recordings of naturally occurring conversations involving twelve male speakers and eighteen female speakers. All conversations were recorded by the researchers from November 2019 to February 2020. These conversations fell in various contexts such as school grounds (teachers with students or other teachers), public plazas (friends and neighbors), and home (couples and children). The participants were selected on the following basis: (a) They are native speakers of Itawit; (b) They are knowledgeable about the Itawit culture; and (c) They can read and write in Itawit. Aside from the naturally occurring conversations, the paper also made use of colloquial-written data which came from two active Facebook pages that primarily encourage the speakers of Itawit to use the language in the community. Appropriate consent was secured from the administrators of the Facebook pages, and utterances taken from the pages were posts from 10 August 2019 to 4 September 2020.

To add to the natural conversations and spoken-like data, the researchers also gathered narrative data from Itawit speakers using the Pear Film, a six-minute-long video which is designed to elicit reactions and narrations from the participants using their language. The informants for Pear stories are Itawit teachers and students from an Itawit-speaking barangay, which is the smallest administrative division in the Philippines. They were chosen based on the same criteria set previously mentioned. They were requested to watch the six-minute film, and after, they were asked to narrate the story using Itawit. They were reminded to just describe what they had viewed and that there were no right or wrong answers in the task. All their responses were recorded and eventually transcribed for linguistic analysis. A corpus of written data was also gathered, but these were used for cross-referencing purposes only as the objective was to reflect the features of present-day Itawit.

It must be noted though that the clauses are taken from varied data (written, spoken and spoken-like data and narrative data); hence, they may be compared for similarities, and in cases where there are discrepancies and irregularities, the researchers consulted Itawit native speakers/informants before making any final grammatical judgment. Hence, working closely with the language informants, the authors also employed some linguistic elicitation techniques.

3 Itawit as a VSO Language

It is generally argued that the Philippine languages are basically VSO languages. By VSO, we mean that sentences produced in Philippine languages follow the structure which begins with the predicate. Thus, such clauses are constructed with verbs occurring in the initial position, while pronominals, modifiers, and objects are positioned after the verb. Itawit speakers generally begin their clauses with a verb, especially in spoken discourse. However, some utterances may also start with a non-verb item. This section discusses the two types of clauses in Itawit: nonverbal clauses and verbal clauses.

3.1 Non-Verbal Clauses

Nonverbal clauses in Itawit are headed by a constituent which does not belong to the category of verbs. Five types of nonverbal clauses are presented in this subsection: nominal predicates (3.1.1), adjectival predicates (3.1.2), existential predicates (3.1.3), prepositional predicates (3.1.4), and locative predicates (3.1.5).

3.1.1 Nominal predicate clauses

As the term ‘nominal predicate clause’ implies, the predicate is a nominal and thus takes the initial position and is followed by the nominative complement which can be a full NP or a pronominal. Four types of nominal predicate clauses are discussed here: (a) classificational, (b) identificational, (c) quantificational, and (d) possessive.

a. Classificational nominal clause

A classificational nominal clause is defined by Reid and Liao (2004:436) as “those that classify the entity expressed in the nominative phrase of the clause”. The authors maintain that the predicate noun is typically a bare noun without a specifying determiner, and since it is a predicate, it is interpreted as the head of the predication. As observed, the lexical items in bold phase that begin the utterances are both nouns that classify the entities: *Marge* and *atawa*.

- (1) ***Empleyadu*** *i* *Marge* *kang* *munsipyo*.
 employee PERS Marge OBL municipal office
 ‘Marge is an employee at the municipal office.’
- (2) ***Seaman*** *ya* *atawa* *na*.
 seaman DET husband GEN
 ‘Her husband is a seaman.’

Example (1) begins with the noun *empleyadu*, which is followed by the nominal marker *i*, which signals the nominal *Marge*. The nominative complement in the said clause is, therefore, a full NP: *i Marge*. The peripheral argument *kang munsipyo* is case-marked as OBL. Example (2) also begins with the noun *seaman*, which does not have direct counterpart in Itawit. It must be noted that the noun *atawa* is introduced by a nominal marker *ya*. *Ya atawa na* then serves as the nominative complement. Notice that the pronominal *na* is case-marked as genitive because it is used to imply an idea of possession.

In some cases, the nominal complement may be expressed by a pronoun, either a free pronoun or a bound pronoun which may or may not encliticize with the nominal predicate.

- (3) ***Seaman*** *iggina*.
 seaman ABS.3s
 ‘He/she is a seaman.’
- (4) ***Memestra*** *ira*.
 teachers ABS.3p
 ‘They are teachers.’
- (5) ***Empleyadu*** *nak* *kang* *munisipyo*.
 employee ABS.1s OBL municipal.office
 ‘I am an employee at the municipal office.’

b. Identificational nominal clause

Reid and Liao (2004) define identificational clauses as those in which the predicate provides specific identification for the entity expressed in the nominative noun phrase. They further mentioned that, while classificational predicates are typically bare nouns, identificational predicates are either a definite common noun, or a personal noun, or a personal or demonstrative pronoun.

- (6) ***I*** ***Luisa*** *ya* *agatadag*.
 PERS Luisa DET standing
 ‘Luisa is the one standing.’

- (7) ***I*** ***Dr. Carreon*** *ya* *wahi=k.*
 PERS Dr. Carreon DET sibling=GEN.1s
 ‘Dr. Carreon is my sibling.’

The examples above have specific names (6) Luisa and (7) Dr. Carreon, which are introduced the nominal marker *i*. The nominal marker primarily is used to identify the entity expressed in the nominative noun phrase. In other cases, an independent personal pronoun can also be a predicate in an identificational nominal clause as in example (8).

- (8) ***Yakan*** *ya* *nangiddan* *kan* *bahat.*
 ABS.1s DET PERF-give OBL banana
 ‘I was the one who gave banana.’

c. Quantificational nominal clause

A quantificational clause begins with a quantifier, usually a numeral which quantifies the entity expressed by the nominative complement, as in examples (9) and (10):

- (9) ***Lima kilo*** *nga* *karne* *yo* *ginatang na* *kam=palengke.*
 five kilos LIG meat DET bought ERG.1s OBL=market
 ‘He/she bought 5 kilos of meat at the market.’

- (10) ***Tallu*** *kansyon* *kiningwa=k.*
 three song PERF-make=ERG.1s
 ‘I made/wrote three songs.’

Obviously, clause (9) begins with a quantifier specifically a unit of measurement used to refer to the noun *karne*. What seems distinct in the clause (9) is the peripheral argument *kampalengke*, which seems to involve the blending of the locative term *kang* with the noun *palengke* that resulted in a one-word locative phrase: *kampalengke*. Clause (10) begins with a quantifier referring to the noun *kansyon*. The clause exhibits encliticisation of the pronominal *ku* to the verb *kiningwa* resulting to *kiningwak*.

- (11) ***Tanga-supot*** *nga* *baggat* *yo* *iniddan=na.*
 one bag/pack LIG rice DET PERF-give=ERG.3s
 ‘A bag of rice is what she/he gave.’

In example (11), the form of measurement indicated in the utterance is given in terms of *supot* ‘one bag/one pack’, which is usually distributed as relief item. Among the Itawit speakers, the traditional measurement of rice comes in terms of *salop* and then eventually in terms of *kilos*. As seen in clause (11) *Tanga-supot* is a measurement expression in Itawit.

It should also be noted that quantificational words may also include time numerals, as in examples (12) and (13):

- (12) ***Alas-sais*** *kan* *gabi* *messimu yaw* *nga* *programa.*
 6 o’clock OBL night will.start DEM LIG program
 ‘The program will start is at 6 o’clock in the evening.’

- (13) ***Alas-dose*** *nak* *mallubet.*
 12 o’clock ABS.1s IMP-go.home
 ‘I will go home at 12 o’clock.’

d. Possessive nominal clause

Possessive nominal predicates are a subclass of identificational predicates. They may contain a possessive pronoun, a genitive or a locatively marked noun phrase interpreted as a possessor in the predicate position.

- (14) **Kwak** *yaw* *nga libro.*
 POS.1s SPA.PROX LIG book
 ‘This book is mine.’

Clause (14) involves an absolute possessive which is marked as POS first person singular in the initial sentence position, which makes it a possessive nominal clause. A similar case is presented in the clause (15) but is given a twist because the possessive word *kwa* is followed by the possessor *Mandy* signaled by the determiner *i*. The sentence also uses a demonstrative locative item marked as LOC.PROX.

- (15) **Kwa** *i* **Mandy** *yaw* *nga* *asassanat.*
 POS GEN Mandy SPA.PROX. LIG doll
 The doll belongs to Mandy.’

3.1.2 Adjectival Clauses

Dita (2007) maintains that Ibanag has a lexical category ‘adjectives’ which are also observed in Itawit. Seen as bare and derived adjectives, they head adjectival clauses. Below we take a close look at the two kinds of adjectival predicates: qualificational and comparative adjectivals.

a. Qualificational adjectival clauses

These items provide a description to the nominal subject. As adjectives, their main role is to qualify the NPs. Clauses (16) and (17) are examples of bare or unaffixed adjectival predicates, while clauses (18) and (19) are examples of affixed or derived adjectives.

- (16) **Gwapa** *ne* *babay* *kanne* *kanto.*
 beautiful DET girl OBL cornerstreet
 ‘The girl at the corner of the street is beautiful.’

- (17) **Dakal** *ya* *vulan* *sangaw.*
 big DET moon now
 ‘The moon is big now.’

- (18) **Nakasta** *ya* *sinnun* *na.*
 nice DET dress GEN
 ‘Her dress is nice.’

- (19) **Sissingngat** *ya* *gulay.*
 very.delicious DET vegetable.
 ‘The vegetable is very delicious.’

It must be noted that the described adjectives are introduced by nominal markers as in (16) *ne babay*, (17) *ya vulan*, (18) *ya sinnun*, and (19) *ya gulay*.

b. Comparative adjectival clauses

When a comparative adjectival clause is used, it usually describes two or more entities. Examples (20) and (21) show the comparative degree and superlative degree of the adjective.

- (20) *Mas* *narenu* *ya* *bale=m* *annet* *cha bale=k*.
 COMP clean DET house=GEN.2s than LOC house=GEN.1s
 ‘My house is cleaner/tidier than your house.’

Clause (20) uses the comparative degree of the adjective *mas narenu* in the sentence-initial position. What is distinct in the utterance is the encliticisation of the second-person pronoun *mu* to the host word *balay* (*balay+mo= balem*) and the first-person singular pronoun *ku* to the host word *balay* (*balay+ku=balek*). Both words seemed to have undergone a phonological change called monophthongization of the original diphthong *ay*; hence, they become *balek* and *balem*. Recall that in Ibanag, the GEN.1s is converted, which is represented only by a glottal stop, but in Itawit, it would seem to appear that the GEN.1s is overt, as seen in clause (20).

- (21) *I* *Saling* *ya* *kapianan* *kan* *ira* *ngammin*
 PERS Saling DET most.intelligent OBL PLU all

nga *maggurwahi*.
 LIG siblings
 ‘Saling is the most intelligent among the siblings.’

Utilizing the superlative degree of the adjective, clause (21) involves the item *kapianan*, which makes the clause a comparative adjectival clause.

3.1.3 Existential Predicate Clause

As the term suggests, existential clauses may express the existence of something or may express possession of something. These constructions also have their negative counterparts. Clauses (22), (23) and (24) show existence, while clauses (25), (26) and (27) show their negative counterparts.

- (22) *Hinian* *dua* *saku* *nga* *baggat* *kang* *kusina*.
 EXI two sacks LIG rice OBL kitchen
 ‘There are two sacks of rice in the kitchen.’
- (23) *Hinian* *furaw* *nga* *bahuy* *cha* *lawan*.
 EXI white LIG pig LOC outside
 ‘There’s a white pig outside.’
- (24) *Nian* *relief* *goods/rasyon* *kattu kabi*.
 EXI relief goods yesterday
 ‘There were relief goods that arrived yesterday.’

The first two examples, (22) and (23), use the existential word *hinian*, while the last example (24) uses *nian*, which seems to be a shortened version of the existential word *hinian*. Examples (25), (26), and (27) express a negative existential meaning.

- (25) *Awan* *tallung* *nu* *Lunit*.
 NEG.EXI class FUT Monday
 ‘There will be no classes on Monday.’
- (26) *Awan* *kwartu=k*.
 NEG.EXI money=GEN.1s
 ‘I don’t have money.’

- (27) **Awan** *kami* *kan* *unag* *balay*.
 NEG.EXI ABS.1pe LOC inside house
 ‘We are not inside the house.’

3.1.4 Prepositional Clause

Reid and Liao (2004) stated that prepositional phrases may be heads of clausal predicates, hence, the term prepositional predicate constructions. They maintain that prepositions are found in languages throughout the Philippines as prepositional heads of clausal predicates. The preposition *para* is used to signal the benefactive role, as in examples (28) and (29):

- (28) **Para** *cha* *anak=ku* *yaw* *nga* *sassanat*.
 for LOC child=GEN.1s DEM LIG doll
 ‘This doll is for my child.’
- (29) **Para** *kannikayu* *yaw* *nga* *kansyon*.
 for OBL.2pe DEM LIG song
 ‘This song is for you.’

3.1.5 Locative clauses

The locative phrase in the following sentences is introduced by the locative particle *kang*. Specifically, locative phrases, as Dita (2007) explains, can be a specific name of a place, or a spatial location as seen in examples (30), (31) and (32).

- (30) **Kang** *Disneyland* *da* *nga* *nabbakasyon*.
 OBL Disneyland ABS.3p REL went.on.vacation
 ‘They went on a vacation to Disneyland.’
- (31) **Kang** *Linao* *nak* *mattrabahu*.
 OBL Linao ABS.1s work
 ‘I work at Linao.’
- (32) **Kang** *eskwela na* *nga* *natafulan* *ya* *sakkalang=na*.
 OBL school GEN.3s LIG found DET ring=GEN.3s
 ‘She found her ring at the living room.’

These examples indicate the use of the locative phrase at the beginning of the sentences. Specifically, the locative phrases express a specific name of a place (*Disneyland*, *Linao*) and a spatial location (*eskwela*).

In some cases, deictic pronouns maybe used as locatives, as in examples (33), (34) and (35).

- (33) **Kanyaw** *ta* *massimmu*.
 here ABS.1pi will.meet
 ‘We will meet here.’

(Note: Speakers from Amulung use *kanyo* instead of *kanyaw*.)

- (34) **Kanyo** *ta* *massimmu*.
 here ABS.1pi will.meet
 ‘We will meet here.’

- (35) **Kannay** *da* *naggungut*.
 there ABS.3p fought
 ‘They fought there.’

It should be noted that speakers from Amulung, an Itawit-speaking municipality, may say the same statement this way. In clause (36), *kanne* is probably the shortened version of *kannay*, while *ra* is likely the shortened version of *ira*, the third-person plural absolutive.

- (36) **Kanne** *ra* *naggungut*.
 there ABS.3p fought
 ‘They fought there.’

When locative phrases are introduced by a deictic pronoun, the locative particle may be dropped, as in examples (37) and (38).

- (37) **Kanyaw** **balay** *nak* *nahanak*.
 here house ABS.1s gave.birth
 ‘I gave birth at home.’

- (38) **Kanyo** **bale=nak** *nanak*.
 here house=ABS.1s gave.birth
 ‘I gave birth at home.’

Note that clause (38) is an utterance from an Amulung speaker who used *kanyo* as a shortened version of *kanyaw* and *nanak* a shortened version of *nahanak*.

- (39) **Kattuna** **Manila** *nak* *maggatang* *kang* *regalu=k*.
 there Manila ABS.1s will.buy OBL gift=GEN.1s
 ‘I will buy my gift there in Manila.’

The above discussion matches Reid and Liao’s (2004) argument that Philippine languages follow right-branching clause structure. That is, clausal constructions usually begin with the predicate, while the nominal complements, adjuncts and other modifiers follow after the predicate.

3.2 Verbal Clauses

Since Philippine languages are generally VSO languages, the normal construction in any Philippine language would be to typically start the sentence with a verb functioning as the predicate, followed by the nominal and verbal complements. These constructions are called verbal clauses, named as such because these clauses are usually headed by verbs that occupy the initial position in clauses. Two types of verbal constructions are presented here: intransitive and transitive constructions.

Transitivity in Philippine languages is determined by the type of complements given to the verbs. As stressed by Reid and Liao (2004), it is the type of complements that a verb takes that determines its transitivity, not the number. Since this study analyzes a Philippine language spoken in northern Luzon, transitivity will be determined by the types of complements given to the verb. On the other hand, valency refers to the number of core arguments that a clause has. When a clause has one core argument, it is called monadic or monovalent. When it takes two core arguments, it is called dyadic or bivalent; when it has three core arguments, it is referred to as triadic or trivalent (Dixon and Aikhenvald 2000). For instance, in Itawit, a monovalent construction contains only one core argument, which may be the actor or experiencer in the clause. As core arguments, the pronominal or the nominal marker is encoded by the absolutive case.

3.2.1 Intransitive Construction

Reid and Liao (2004) define an intransitive construction as having a verb with only a single nominal complement. This single complement is referred to as the core argument, an argument that is needed to complete the meaning of the sentence. Such a complement in this type of construction may carry either the actor or the undergoer. An intransitive construction may have one or two or several complements as peripheral arguments, but it only has one core argument. Additionally, there are clauses which are semantically intransitive. Dita (2007) termed these ambient clauses in Ibanag. This is also evident in Itawit utterances.

a. Ambient Clauses

Ambient clauses do not exhibit core arguments. As explained by Dita (2007), these are semantically intransitive constructions as they refer to temporal states and hence may not require any accompanying nouns. Consider clauses (40) and (41) provided by Dita (2007:49).

(40) *Magguran.*
IMP-rain
'It's raining.'

(41) *Nabbaddyu.*
PERF-storm
'It stormed.'

Similar constructions are also observed in Itawit. See examples (42), (43) and (44).

(42) *Nakillakit.*
PERF-lightning
'There was lightning.'

(43) *Nallunig.*
PERF-earthquake.
'The was an earthquake.'

(44) *Mapafuk.*
CONT-drizzle
'It's drizzling.'

b. Monovalent (monadic) intransitive

Itawit utterances also exhibit monovalent construction which contains only one core argument. Pronouns, like NPs, can also serve as core arguments. As core arguments, pronominal or the nominal markers are encoded by the absolutive case. Clauses (44) and (45) exhibit the use of NPs without adjuncts after them. All the above clauses have only one core argument which is marked as ABS. The core argument in clause (45) is a genderless pronoun in Itawit, *iggina*, which comes after the verb in the past tense *naddangot*, while clause (46) has the third-person plural pronoun as the core argument. Clauses (47) and (48) have nominal core arguments introduced by the nominal marker *yo*.

- (45) *Naddangot* *iggina.*
 PERF-cry ABS.3s
 ‘S/he cried.’
- (46) *Nakkarela* *da.*
 PERF-run ABS.3pe
 ‘They ran.’
- (47) *Naggalak* *yo* *abbing.*
 PERF-laugh ABS child
 ‘The child laughed.’
- (48) *Nanadag* *yo* *babay.*
 PERF-stand.up ABS girl
 ‘The girl stood up.’

NPs may also exhibit adjuncts of time, place and manner after them. Examples (49) and (50) have pronominal NPs with adjuncts after them. These include the pronominals *ka* and *nak*, which are placed after the verb *mazzihut* and *matrabahu*. These pronominals serve as the core argument in these clauses, so they are in the absolutive case. Notice that the adjuncts of time appear after the pronominal complements.

- (49) *Mazzihut* *ka* *akkinagalgaw.*
 take.a.bath ABS.2s everyday.
 ‘(You) take a bath everyday.’
- (50) *Matrabahu* *nak* *sonu* *umma.*
 will.work ABS.1s FUT tomorrow.
 ‘I will work tomorrow.’

Dita (2010) explains that, in Ibanag, adjuncts of time, place and manner may appear after the noun complement. In Itawit, as seen in clause (51), the adverb appears before the verb. The word *gavvat*, which is placed at the beginning of the sentence before the verb, functions as an adverb of time.

- (51) *Gavvat* *naddangot* *i* *Maria.*
 suddenly cried ABS Maria
 ‘Maria suddenly cried.’

c. Bivalent/Divalent (dyadic) intransitive

A bivalent intransitive construction has two nominal complements: actor/experiencer and theme/patient. The theme refers to an NP that expresses an entity which is a state or a location or which is undergoing a motion (Trask 1993). Such an entity is always non-human, either animate or inanimate. This is different from the patient, which functions the same way as the theme but is human and/or animate. In case-marking the argument, the actor in a bivalent intransitive construction is always marked as absolutive (ABS), and the theme is marked as oblique (OBL). The nominal complement in bivalent intransitive construction may take the form of a pronominal or a full NP.

With an absolutive pronoun

This type of construction consists of only one core argument which is encoded by the absolutive pronoun. The core argument in clause (52) is the pronominal in first person singular *nak*, while clause (53) has the third-person genderless pronoun *iggina*, both are encoded absolutive. The themes in the

following clauses, *kurtina* ‘curtains’ in (52) and *baggat* ‘rice’ in (53) are encoded by the oblique *kang*, which means they are not core arguments, but rather they are peripheral arguments.

- (52) *Nabbambal* **nak** *kang* *kurtina* *kattuna* *kayan*.
 washed **ABS.1s** OBL curtain OBL river
 ‘I washed curtains in the river.’
- (53) *Nakkokot* **iggina** *kang* *baggat*.
 stole **ABS.3s** OBL rice
 ‘She/he stole rice.’

With an absolutive full NP

A bivalent intransitive construction may also have a full NP instead of a pronominal as a nominal complement. Full NPs are typically actors in this type of clause, and they can be substituted by a pronominal. The themes are also encoded by the oblique marker *kang*. Examples (54) and (55) show a bivalent construction with an absolutive full NP *i nanang* in (54) and *i wahik* in (55). The absolutive full NP is the only core argument in the sentence. The themes, *bahuy* ‘pig’ and *Mondonggo* (an authentic Itawit dish) and *kusina* ‘kitchen’ are further encoded by the oblique *kan*.

- (54) *Naggatang* *i* *nanang* *kang* *lima* *nga* *bahuy*.
 bought **ABS** mother OBL five LIG pig
 ‘Mother bought five pigs.’
- (55) *Nangan* *i* *wahi=k* *kang* *Mondonggo* *kang* *kusina*.
 Ate **ABS** sibling=GEN1.s OBL Mondonggo OBL kitchen
 ‘My sibling ate Mondonggo in the kitchen.’

3.2.2 Transitive Constructions

An Itawit transitive construction, unlike an intransitive one, requires two core arguments: the agent and the patient. In this case, the agent is in the ergative case, and the patient, which is always a human complement, is labeled in the absolutive case. It may either be bivalent or trivalent. The following section gives a discussion of bivalent and trivalent transitive constructions.

a. Bivalent / divalent (dyadic) transitive

Bivalent transitives contain two core arguments which could be full noun phrases or plain pronominals. As previously mentioned, the agent is case-marked as ergative, and the patient is case-marked as absolutive, while the other peripheral arguments present are case-marked as oblique.

With two full NPs

In a bivalent transitive construction, the two full NPs are the agent and the patient, which are the core arguments; hence, they are marked as ergative and absolutive respectively. If the clause has other NPs, they are peripheral arguments which are case-marked as oblique.

- (56) *Inuffunan* *i* *mestra/tru* *ya* *abbing*.
 helped ERG teacher ABS child
 ‘The teacher helped the child.’
- (57) *Pinakan i* *tatang* *ne* *makilelimut*.
 fed ERG father ABS beggar
 ‘Father fed the beggar.’

With two pronominals

- (58) *Nessussuk da iggina.*
 hid ERG.3p ABS.3s
 ‘They hid him/her.’
- (59) *Kinanna=na yakan.*
 hit =ERG.3s ABS.1s
 ‘He/she hit me.’

With pronominal ergative and full-noun absolutive complements

In this construction, the core arguments may be a combination of an ergative pronominal and an absolutive full NP.

- (60) *Inallangngan=na ya anak=na kattu kabi.*
 scolded=ERG.3s ABS child =GEN.3s OBL yesterday
 ‘She scolded her child yesterday.’
- (61) *Netoli da yaw assasanat.*
 returned ERG.3pe DEM doll
 ‘They returned this doll.’
- (62) *Padayawan tayu i Afu!*
 worship ERG.1p ABS God
 ‘Let us worship God.’
- (63) *Hinaradaral=na ya kofunna*
 Destroyed=ERG.3s ABS friend=GEN.3s

kanya kiklase=na ira
 OBL classmates=GEN.3s. PLU
 ‘S/he destroyed his/her friend to his/her classmates.’

With full noun ergative and pronominal absolutive complements

This construction has an agent which is a full NP and a patient which is pronominal.

- (64) *Nassingan ne Pedro ira.*
 saw ERG Pedro ABS.3p
 ‘Pedro saw them.’
- (65) *Nessussuk ne bagitolayiggina.*
 hid ERG young.man ABS.3s
 ‘The young man hid her/him.’

b. Trivalent (triadic) transitive

This kind of construction has three core arguments: the agent and the benefactive which are humans and the theme which is non-human. Here, the agent is case-marked as ergative, the theme as absolutive, and the benefactive as oblique. Let us consider the following examples.

With three full NPs

There are three core arguments in clauses (66) and (67) which take the form of full NPs.

- (66) *Nepangwa* *i* *angkel* *Jimmy* *kan* *langkapi*
 built **ERG** uncle Jimmy **ABS** bed

ne *anak* *na*.
OBL child GEN.3a

‘Uncle Jimmy built a bed for his child.’

- (67) *Nelutu* *ne* *anti* *Suping* *kan* *zinagan*
 cooked **ERG** aunt Suping **ABS** dinuguan

para *kanne* *mestru=k*.
 for **OBL** male.teacher=GEN.1s

‘Aunt Suping cooked dinuguan for my male teacher.’

With two pronominals and a full NP

In clauses (68) and (69), the agent and the benefactive are pronominals, while the theme is a full NP.

- (68) *Netturat* *nak* *kan* *kasyon* *para* *kan* *ikayu* *ngammin*.
 wrote **ERG.1s** **ABS** song for **OBL** you all
 ‘I wrote a song for all of you.’

- (69) *Gumatang* *ka* *kan* *espeho* *para* *kaniggina*.
 buy **ERG.2s** **ABS** mirror for **OBL.3s**
 ‘Buy a mirror for her/him.’

6 Conclusion

This paper has been concerned with describing and analyzing the types of clauses in Itawit, a language spoken by nearly 189,000 speakers in Northern Luzon. Nonverbal clauses in Itawit are headed by a constituent which does not belong to the category of verbs. Verbal clauses, on the other hand, are usually headed by verbs that occupy the initial position in the clause. The analysis of this Itawit data supports Reid and Liao’s (2004) argument on the right-branching clause structure of Philippine languages. Moreover, the paper affirms the finding that transitivity in Philippine languages is determined by the type of complement given to the verb and valency refers to the number of core arguments that a clause has. The paper has clearly distinguished intransitive construction from transitive construction. An intransitive construction has a verb with only a single nominal complement. It may have several complements as peripheral arguments, but it only has one core argument. Intransitive construction then may be monovalent with one core argument as the actor, and it may be bivalent with two nominal complements, actor/experiencer and theme/patient. An Itawit transitive construction, unlike intransitive, requires two core arguments: the agent and the patient.

References

- Bollas, Abigail, & Hernandez, Viveca. 2013. Comparative analysis on the phonology of Tagalog, Cebuano and Itawis. Web.
https://s3.amazonaws.com/academia.edu.documents/31847507/Comparative_Phonology_Paper.pdf
- Bollas, Abigail, & Hernandez, Viveca. 2014. Comparative analysis on the anaphors of Tagalog, Hiligaynon and Itawis. Web.
https://www.academia.edu/8216727/Comparative_Syntax_Anaphora_in_Tagalog_Hiligaynon_and_Itawis.

- Bollas, Abigail, & Supnet, Chiara Paula. 2013. Pluralizing morphemes in four Philippine languages: Bikol, Tagalog, Itawis and Ayta Mag-Antsi. Web. https://www.academia.edu/8216871/Pluralizing_Morphemes_in_Four_Philippine_Languages_Bikol_Tagalog_Itawis_and_Ayta_Mag-Antsi.
- Dita, Shirley N. 2007. A reference grammar of Ibanag (Doctoral dissertation, De La Salle University, Philippines).
- Dita, Shirley N. 2013. *Itawit-Tagalog-English trilingual dictionary*. Manila: National Commission on Culture and the Arts.
- Dixon, R.M.W. & Aikhenvald, Alexandra Y. 2000. Introduction. In R. M. W. Dixon and A. Y. Aikhenvald (Eds.), *Changing valency: Case studies in transitivity*, 1-29. Cambridge and New York: Cambridge University Press.
- Elli, Marley Vea Clarisse & Isidro, Abigail Loren J. 2013. A contrastive analysis of Ibanag and Itawis morphological processes. Web. https://www.academia.edu/25990119/A_Contrastive_Study_of_Ibanag_and_Itawis_Morphological_Processes
- MacKenzie, Laurel. 2010. Reduplication in Itawis. Poster presented at the 2010 Annual Meeting of the Linguistic Society of America. Available in: <http://dx.doi.org/10.3765/exabs.v0i0.517>
- Natividad, Mateo & Solomon, Syble. 1970. *A guide to learning Itawis dialect*. Manila: Peace Corps.
- Reid, Lawrence A. 1974. The Central Cordilleran subgroup of Philippine languages. In Papers of the First International Conference of Comparative Austronesian Linguistics, Honolulu. *Oceanic Linguistics*. 5:511-560.
- Reid, Lawrence A. 1978. Problems in the reconstruction of Proto-Philippine construction markers. In S. A. Wurm & L. Carrington (Eds.), *Second International Conference on Austronesian Linguistics: Proceedings, Facsile I – Western Austronesian*, (pp. 33-66). Pacific Linguistics Series C, No. 61.
- Reid, Lawrence A. 2006. On Reconstructing the morphosyntax of Proto-Northern Luzon. *Philippine Journal of Linguistics*. 37:1-64.
- Reid, Lawrence A. & Liao, Hsiu-chuan. 2004. A brief syntactic typology of Philippine Languages. *Language and Linguistics*. 5.2:433-490.
- Schachter, Paul & Otanes, Fe T. 1972. *Tagalog reference grammar*. Berkeley: University of California Press.
- Simons, Gary & Fennig, Charles D. (Eds.) 2018. *Ethnologue: Languages of the world*, Twentyfirst edition. Dallas, Texas: SIL International. Web. <https://www.ethnologue.com>
- Trask, Robert Lawrence. 2013. *A Dictionary of Grammatical Terms in Linguistics*. New York: Routledge.
- Zafra, Galileo & Lucero, Rosario Cruz. 2017. Peoples of the Philippines. *Encyclopedia of Philippine Art*, 2nd ed. Manila: Cultural Center of the Philippines.

DIRECTIONAL TERMS IN IDU, A LANGUAGE OF NORTHEAST INDIA

Roger BLENCH

*McDonald Institute for Archaeological Research
University of Cambridge
rogerblench@yahoo.co.uk*

Abstract

The Idu language is spoken in the extreme northeast of Arunachal Pradesh. Its traditional terrain stretched between the Tibetan Plateau and the plains of the Brahmaputra. The Idu traditionally lived in highly dispersed settlements. As a consequence, orientation was determined more by up versus down as well as by the sides of the many rivers which cross their area. The paper describes the lexicon of Idu in describing directions. The morphosyntax of directionals is quite mixed; some can be identified as deictics with up/down semantics, while others are best described as specialized adverbs. In addition, there are verbs of motion which are mark directionality. All these forms incorporate the same semantic frame. There is probably no distinct morphosyntactic category of directional. The paper illustrates each term with example sentence contexts and represents visually directional oppositions as they appear to speakers. Historically, it seems unlikely that Idu recognized cardinal directions, North/South etc., but the existing lexicon has been interpreted in terms of modern terminology, which sometimes creates confusion for speakers.

Keywords: Idu; directionals

ISO 639-3 codes: clk, byw, bod, mlv, wno

1 Introduction: directionals

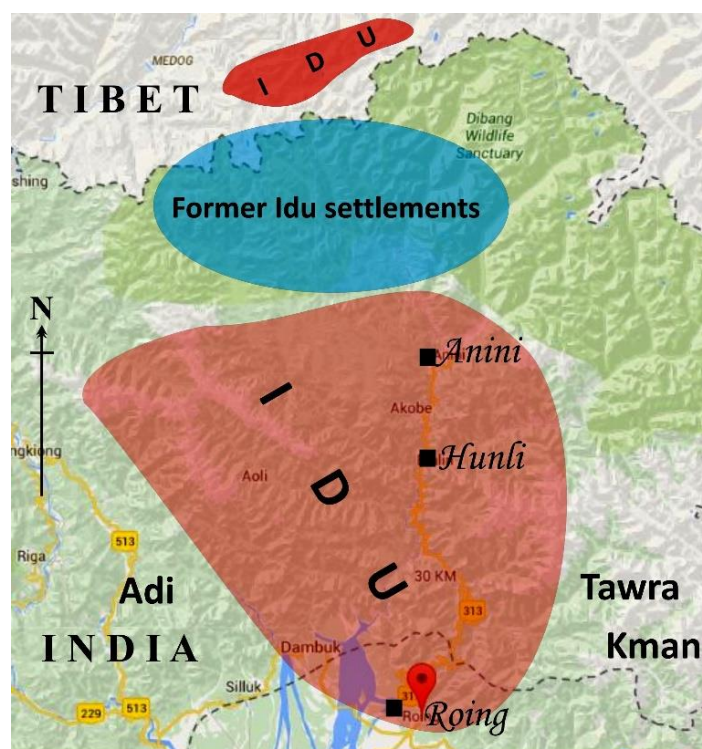
The observation that the natural environment in which people live has a strong impact on the grammatical and semantic systems they have evolved goes back at least to Sapir (1912). However, the literature on directionality or topographical deixis is quite sparse, partly because the languages where this has typically been studied are plains or maritime people. Such languages focus on spatial terminology, and there are a number of descriptions or orientation in Austronesian for example (Ozanne-Rivierre 1997; Bennardo 2002; Alexandre 2003 for Mwotlap [mlv]; Burung 2013 for Wano [wno]). Schapper (2014) deals explicitly with elevation in the Alor-Pantar languages. Although Africa has plenty of montane areas, few descriptions exist of topographical deixis. However, see Wolff (2006) which discusses some of the languages of the Mandara mountains in northeastern Nigeria. In the Himalayan region, few studies have described directionals, but see Bickel (1997) for Belhare [byw]; Caplow (2007) for Tokpe Gola Tibetan [bod]. Similarly, there have been some general considerations of spatial coordinate systems such as Dixon (2003) and Burenhult (2008).

This paper concerns the directionals and other deixis of the Idu people, who live in the northeast of Arunachal Pradesh, itself in the extreme northeast of India. The Idu language resembles Trans-Himalayan typologically, although any genetic relationship is yet to be demonstrated. The term ‘Mishmi’, often found in the literature and still current among the Idu in some contexts, is used in the travel literature as far back as the early nineteenth century to refer to three distinct peoples, the Idu, Tawrā [=Taraon] and the Geman [=Miju]. The common name ‘Digaru’, the name of a major river, is also in use. Culturally speaking, these two languages were historically grouped with Kman, as the Mishmi. The Idu live principally in Dibang Valley District with some settlements in Lohit and E Siang in Arunachal Pradesh. The ‘Upper’ Idu are known as ‘Luoba’ or ‘Khoba’ in China where there are a few villages. The 1971 census in India recorded around 7,700 individuals self-identifying as Idu Mishmi, although this is no measure of language competence.

Ironically, given there is only a small scattering of villages in Tibet, the technical literature is largely in Chinese, including at least one full-length grammar. Publications include Sun et al. (1980), Sun (1983a and b, 1999), Ouyang (1985) and Jiang (2005)¹. An extensive dictionary and draft grammar based on the form of the language spoken in India can be found on the author’s website² while Blench (2019) summarizes the phonology and orthography used in this paper.

The Idu live in vertiginous territory, extending from the plains of the Brahmaputra Valley to Tibet. However, prior to the earthquake of 1951, there were almost no settlements on the plain, and the general pattern was dispersed settlement along six main river valleys (Baruah 1988; Bhattacharjee 1983). In this area of Arunachal Pradesh the inclines are particularly steep, and almost all movement would have been up and down the mountainside. Unlike some of the neighbors of the Idu, such as the Tani, who spread around the angle of the mountain range, ‘up’ would always have been North for the Idu and ‘down’ would have been South. The river valleys which cut across Idu territory were the main axes of communication, and the rope-bridges which used to span the major rivers were vital to trade and social intercourse. As a consequence, the Idu language evolved to reflect this pattern of orientation, which is now somewhat discordant with their post-earthquake geography. Following the massive destruction, many villages and individuals moved down to the plain to live among the Assamese and Nepali resident there.

Map 1: The territory of the Idu



Key: INDIA Nation State
 Adi Ethnic group
 ■ Major Idu settlement
 - - - - - International boundary
 Idu territory

¹ These publications are accessible and indeed available for sale on the internet. I cannot read Chinese, but a combination of English glosses and example sentences means that is possible to establish general correspondences with data from the Indian side. The Chinese notation of tones seems to correspond broadly with those recorded for the present study

² [Idu resources \(rogerblench.info\)](http://rogerblench.info)

Map 1 shows the approximate extent of Idu today. It has recently been reported that the few Idu villages in China (where they are known as Lhoba) have been forcibly removed further inland, and that attempts are being made to Sinicize this small outlying population.

As a consequence of their geographical situation, Idu has developed an extensive set of directionals, adverb-like forms which incorporate deixis. These ‘topographical deictics’ are reported for Tani languages (Post 2011, unpublished) and are likely to be prominent in cultures living in steep environments, where ‘up’ and ‘down’, ‘north’ and ‘south’ are more relevant than conventional cardinal directions.

This paper³ describes the movement verbs marked for direction, the basic deictics of Idu and then goes on to exemplify the different types of directional, based on orientation towards the mountains and plains, the rivers, and the village. It concludes there is a strong relationship between this lexical richness and the physical environment of the Idu.

2 Movement verbs marked for direction

The natural environment the Idu people inhabit ensures that there is a semantic merger between climb and other more usual forms of ascension. So ‘climb’ and ‘go up’ use the same term marking elevation, but do not incorporate directionality, as shown in Sentence 1.

1. shû ‘to climb, ascend, go up’
 āyā īmú èyà shū.gà
 that man mountain climb.ing
 ‘That man is climbing the mountain.’

Similarly, and more alarmingly, ‘fall, go down, descend’ are also merged into one verb. This has particular resonance for Idu speakers, as falling off a mountain is a particular type of death for which a specific level of the underworld is reserved. As sentence 2 indicates, the verb is associated with the soul leaving the body.

2. cōpô(tó) to fall, go down, descend
 ngá àyá nē cōpōtó àbà nē īlīngā āyāngā bā
 I hill from fall PERF and soul go
 ‘I fell from the hill and my soul left my body.’

3. Deictics

Idu conventional deictics do not form part of the directional system and correspond quite well to those in English, as they are not marked for direction or orientation. The following examples of spatial deictics and demonstratives provide a brief summary of their use.

3. àlá here
 nàbā àlá jí.gāyí
 Father here sit.PRES
 ‘Father is sitting here’

³ Data for the paper was gathered in a series of field trips to Idu territory, 2015-2020. I would like to thank Mite Lingi and Hindu Meme both for help with working on understanding directionals and developing the transcription of Idu.

4. àhí there
 àhí m̄a isēyā m̄e dè.gà?
 there LOC who AFF stand.ing
 “Who is standing there?”
5. ēcā this
 ēcā yū àkā taci pùmà
 this beer be strong very really
 “This rice-beer is very strong”
6. âyâ that
 âyâ imú hābrū mbrā
 that man eat too much very
 “That man is a glutton”

4 Directionals

It is in the field of directionals that Idu is particularly rich. Directionals can be subdivided into five major semantic categories, summarized in Table 1.

Table 1: Categories of Idu directionals

Category	Comment
Verticality	up/down in relation to the position of the speaker
Cardinal directions	North/South etc. (though there is evidence these are modernised interpretations of more traditional terms)
Rivers	upstream/downstream
Villages	towards the upper/lower part of the village
Handedness	right/left

Table 2 lists the lexemes associated with Idu directional categories and these are systematically exemplified in the following sections.

Table 2: Lexemes used in Idu directionals

Category	Idu	Gloss
Verticality	àdrí	straight up (speaker is on the ground)
	àmá	straight down (e.g., speaker is in a tree)
	àyùmà	downwards
	àyùmányū	towards downwards
	ètòlō	upwards
	ètò(lō)nyū	towards upwards
Cardinals	àló, yàló	upwards, North
	(y)àlónyú	northern side
	ātú	up there North
	átúdrī	up there on top, high up
	ātúyā	there upwards there North (close)
	àmá	down South
	àmāyā	there South, downwards (close)
	àmányú	southwards
	àpí	on the south side, down there South (remote)
	àhí	over there East or West
	àhíyā	here East or West (close to speaker)
àhílā	there East or West (close to speaker)	

Category	Idu	Gloss
Rivers	àno	downstream
	àrhó	upstream
	àhínyū	on the other side
	(maci) hrēgēnyū	on the other side esp. rivers
	(maci) ēkōnyū	on this side esp. rivers
	ēlānū	on this side
	ēwānyū	on this side Hill dialect
	īlīn(y)ū	on my side
Villages	ànggōcá	towards the upper part of the village
	ànggōpò	towards the low-lying part of the village
Hand	ēcānyū	right side
	lākēnyū	left side

Two aspects of Idu syntax relevant to the directionals can be noted here. Word order in Idu is fairly free, with prominence or focus denoted by fronting locatives and qualifiers. Idu frequently omits subject pronouns where these can be inferred from context, and thus directionals are often in initial position in a clause or sentence. Idu has a large array of locative markers, both bound and unbound, which often have the effect of multiplying the marking of position or orientation of subjects and speakers.

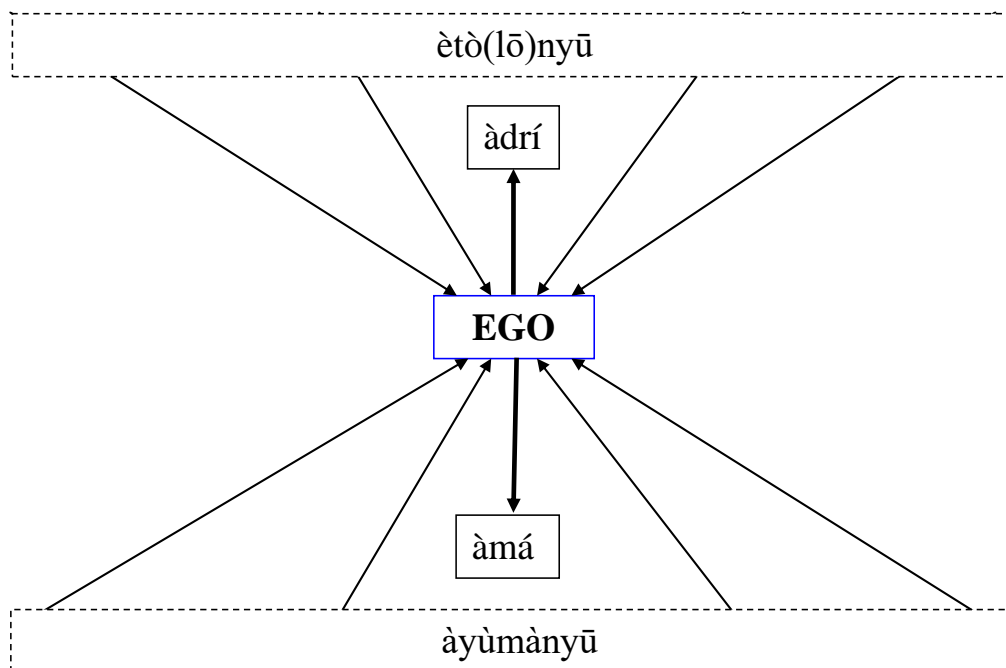
Verticality

The Idu environment is montane, so it is frequently necessary for a speaker to point up and down to a hearer to indicate the position of something. However, trees are also the dominant vegetation type, and must often be climbed for their fruit and other products, hence the vertical directionals exemplified below. Sentences 7-10 provide examples of these terms in use.

7. àdrí straight up, up there (e.g., if you are on the ground)
āyā àdrí àcápù àkhà ābā á
that up shelf keep IMP AFF
“Keep that on the shelf there”
8. àmá straight down, down from (e.g., if you are in a tree)
āsīmbó.ànè àmá àndōnggō dō.àjà chō!
tree.LOC there down jump.IMP HORT
“ [You], jump down from the tree! “
9. ètònyū upwards (up from the ground)
ètònyū shū.hí.mì.yì
upwards climb.can.NEG.PRS
“It is hard to climb upwards“
10. àyùmànyū downwards, down from there
āyā mánē àyùmànyū ébò.àjà gò chì.prā.prà.yì
there LOC downwards fall.IMP from walk.good.is.PRS
“It is pleasant to walk down from [the mountain]”

The mountain is not explicitly mentioned, but an Idu speaker would infer this from the context.

Figure 1: presents the vertical directionals as a graphic



Cardinal directions

The European system of cardinal points is quite a recent adoption in Idu and has been superimposed on a previous system which essentially marked upwards (i.e., towards Tibet) and downwards (towards the Assam plains). Although speakers translate the directionals using ‘North’ and ‘South’ today, this is a modern gloss on an initially quite distinct system of orientation. Sentences 11-17 provide examples of these terms in use.

- 11a. àlô North upwards [yàlô in Upper dialects]
ngá àlô Anini nē jā
 I north ànīnī from come down
 “I came down from Anini“
- 11b. àlô North upwards [omitting subject pronoun]
àlô mrāā nē jā?
 north hill from come?
 “Have you come down from the hill?“
- 11c. yàlô [Upper dialect]
Hūli yàlô khà.gāyí
 Hunli up there lie.PRS
 “Hunli is up there“
12. âtú up there North
ngá ālīyā ātú ànggōcá jí.gāyí
 my brother up there north side live.PRS
 “My brother is living up there in the upper part“

13. àtúdrí up there, on top
àtúdrí ìmūdù.mā prā lì.gāyí
 up there sky.LOC bird fly.PRS
 “Up in the sky, a bird is flying”
14. âtúyâ there upwards North
âtúyâ m̀òcá prà.m̀. nē h̀d̀!
 there near COP.AFF.LOC EXCL
 “It is very near up there!”
15. àmāyā there downwards, South
àmāyā g̀è á
 there go and come LOC
 “Go there and come back”
16. àpí from the south side, down there South
àpí nyú.nē ìtū á yí?
 south you.LOC come QM AFF
 “Have you come from the South?”
17. yàlónyú north (downwards)
yàlónyú nē lì.gā.à j̄ā.yì
 northern side from fly.PL.FOC come.PF
 “They flew down from the northern side”

Idu did not originally distinguish East and West, as there was only a generic term which meant ‘sideways’. If modern cardinal directions are required, they are expressed with borrowings from English. Note that degree of remoteness from the speaker is lexically specified, in contrast to the North/South distinctions. Sentences 18-21 provide examples of these terms in use.

- 18a. àhí sideways, over there East or West, close to speaker
àhí ìmú khàgà dè.gāyí
 there person one stand.PRS
 “One person is standing there ”
- 18b.
àhí mā ìsīyā.mè dè.gà?
 there LOC who standing
 “Who is standing there?”
19. àhínyû that side, the direction you are facing (East or West only)
àhínyû bā.bā hìmì.yā
 there go.IMP be able.AFF
 “Go over to that side”

The following two terms in 20 to 21 appear to be interchangeable.

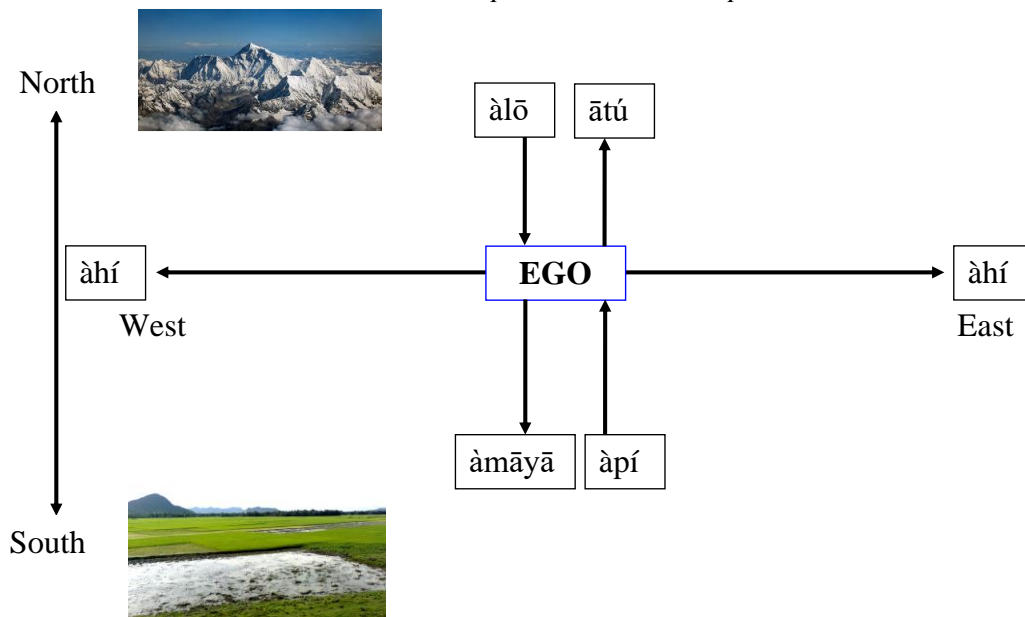
20. àhílâ sideways, there East or West, remote from speaker
 àhílâ ikū khàgà sí tēnè khà.gà mā
 there dog one die CONN lie.PRT AFF
 “There’s a dead dog over there”

- 21a. àhíyâ sideways, there East or West, remote from speaker
 àhíyā āsīmbó.ā prā.a khàgà ndō.gāyí
 there tree.LOC bird.SG one perch.PRS
 “A bird is perching on that tree over there”

- 21b.
 àhíyā isīyā.gā ó.ó á?
 there whose.LOC house.LOC QM
 “Whose house is that over there?”

Figure 2 shows the directionals exemplified above in relation to the mountains and the plains.

Figure 2: Directionals in relation to mountains/ plains and cardinal points



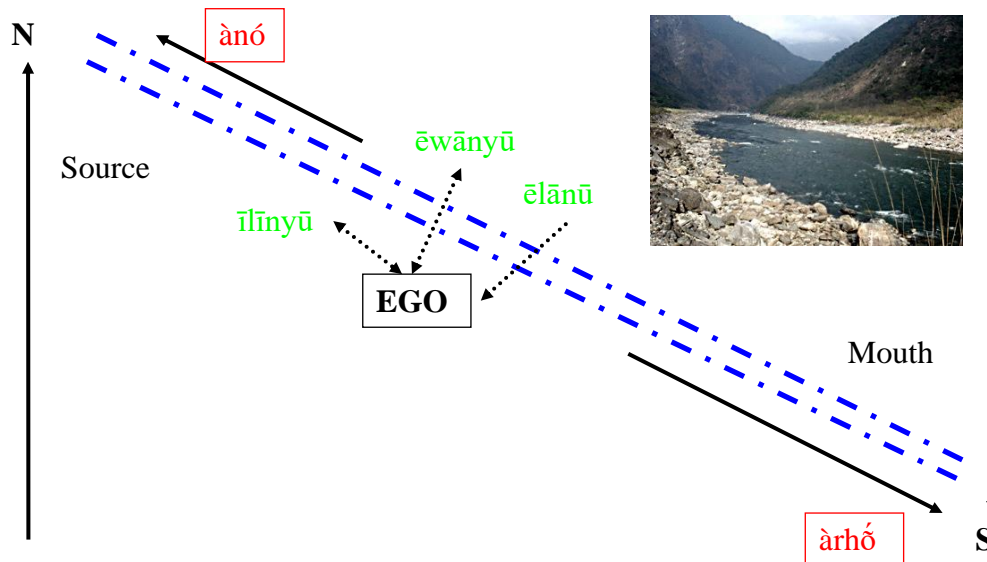
Rivers

Idu terrain is heavily dissected by rivers, which are usually dry in the later part of the year, but flooded suddenly following the snowmelt on the Tibetan Plateau. Although some of these have been bridged recently, and risky crossings using heavy iron cables have been possible for a century or more, being on the right side of a river at a given time takes on great importance in Idu life. Hence, the terminology marked both the direction of flow from the source to the mouth, as well as the side of the where the speaker is. Sentences 22-29 provide examples of these terms in use.

22. ànó downstream (towards the mouth)
màcì ànó dùyū bā chō
 water downstream side go HORT
 “Let’s go downstream”
23. àrhǒ upstream (towards the source)
màcì àrhǒ dùyū īmú āgū.gā.gà āthú.jīyà?
 water upstream side people walk.ing.PRS see.QM
 “Do you see people walking upstream?”
24. ēwānyū on the other side
màcì ēwānyū dè gānē grá.gāyí
 river other side stand then shout.PRS
 “He is standing on the other side of the river and shouting”
25. īlīn(y)ū on my side (originally of a river)
nyú īlīn(y)ū ngá mbrōmrò jí
 you sg. this me with sit
 “You sit this side with me”
26. èlānū on this side (originally of a river)
nyú èlānū ibī.lǒ
 you sg. this side come from.LOC
 “You come this side please”

These words have been extended to more general contexts in recent times.

Figure 3: Directionals from a riverbank



On both sides

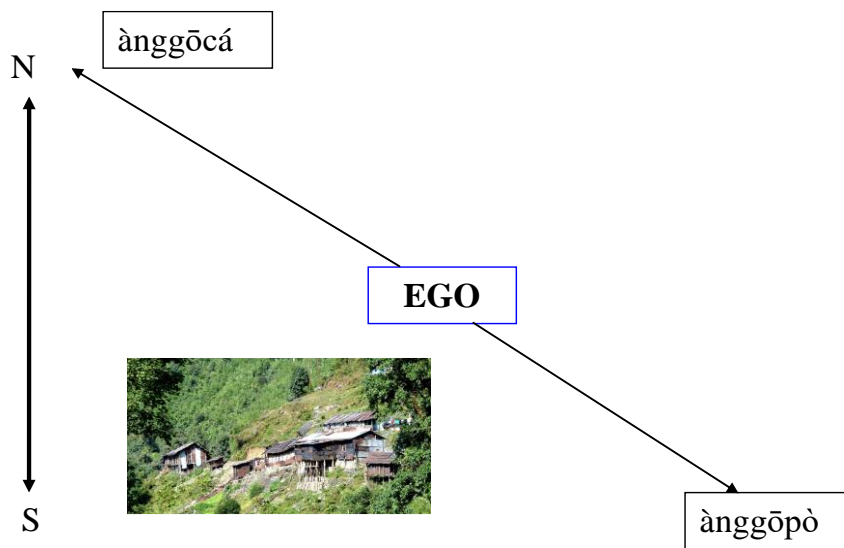
The following locative expressions have been extended from terms which applied originally only to rivers.

27. ēnōnyū hòyà both sides
 ēnōnyū hòyà nē āgū.prà.gāyì mǎ
 both sides LOC walk.possible.PST AFF
 It is possible to walk from both sides
28. ēnōnyū...dēgà describes something which is positioned both sides of the subject
 nyu ēnōnyū īmú dē.gà
 you side people stand.PRS
 “There are people standing on both sides of you”
29. ēnōnyū dùnū at both ends
 ngá.cì ó ēnōnyū dùnū màcì khà.gà
 my house both sides water lie.PRS
 “Water is lying on both sides of my house”

Villages

It is likely that prior to the 1951 earthquake, most Idu villages were very small, consisting of only a few households, scattered up and down the hillside and sometimes not even in sight of one another. Hence, it was useful to specify the direction of the ‘upper’ and ‘lower’ village. Sentences 30 and 31 provide examples of these terms in use.

30. ànggōcá towards the upper part of the village
 ngá ó hē Ējēnggò àtīkō ànggōcá dùnyū khà.gāyì
 I house LOC Ejengo village upper part side lie.PRS
 “My house is in the northern side of Ejengo village”
31. ànggōpò towards the low-lying part of the village
 ēcá ànggōpò dùnyū.nē āmá yà.gā mpū
 here downside side.LOC wind blow.PRS AFF
 “The wind blows here from the downside”

Figure 4: Directionals within the village**Handedness**

Idu also recognises left and right in respect of an individual. Unlike the other directionals, these could either be treated as nominals or adverbials. Sentences 32 and 33 provide examples of these terms in use.

32. ēcānyū right side
 Ìméhí.cì nyùkō ēcānyū khà.gāyí
 Imehi.POSS room right side lie.PRS
 “Imehi’s room is on the right side”
33. lākēnyū left side
 ēcā ngá lākēnyū dè.ga
 this my left side stand.AFF
 “This is on my left side”

5 Conclusions

Conventional deictics in Idu are quite sparse and closely resemble those in English in terms of semantic structure. However, Idu has an extremely rich system of directionals, in relation to mountains and plains, rivers and villages, which is a reflection of the steep environments the Idu traditionally inhabited. The neighbouring Tani people have a system which is evidently related conceptually although there are apparently no common lexical items. It is likely these directional systems are more widespread in the region than is apparent in the descriptive literature.

References

- Alexandre, François 2003. Of men, hills and winds: space directionals in Mwotlap. *Oceanic Linguistics* 42.2:407-437.
- Baruah, Tapan Kumar 1988. *The Idu Mishmis*. Itanagar: Government of Arunachal Pradesh. [reprint of a 1960 publication].
- Bennardo, Giovanni (ed.) 2002. *Representing space in Oceania: Culture in language and mind*. Pacific Linguistics, 523. Canberra: Australian National University.
- Bhattacharjee, Tarun 1983. *The Idus of Mathun and Dri Valley*. Shillong: Government of Arunachal Pradesh.

- Bickel, Balthasar 1997. Spatial operations in deixis, cognition, and culture: Where to orient oneself in Belhare. *Language and conceptualization*. ed. by Eric Pederson and Jan Nuyts, 46-83. Cambridge University Press.
- Blench, Roger M. 2019. Attempts to Write the Idu Mishmi Language and a Proposal for a Modern Orthography. In: *Change and continuity among tribals: the Idu Mishmis of the Eastern Himalayas*. Tarun Mene & S.K. Chaudhury eds. 115-130. Bhopal: Indira Gandhi Rashtriya Manav Sangrahalaya / Rawat Publications: Jaipur.
- Burenhult, Niclas 2008. Spatial coordinate systems in demonstrative meaning. *Linguistic Typology* 12.99-142.
- Burung, Willem 2013. Elevative deixis in Wano. *Journal of Literature and Art Studies* 5.8:634-642 doi: 10.17265/2159-5836/2015.08.008
- Caplow, Nancy J. 2007. Directionals in Tokpe Gola Tibetan discourse. *Trends in linguistics: Linguistics of the Himalayas and beyond*. ed. by Roland Bielmeier and Felix Haller, 23-46. Berlin: Mouton de Gruyter.
- Dixon, R.M.W. 2003. Directionals: a cross-linguistic typology. *Studies in Language* 27.1:61-112.
- Jiang, Di. 2005. *Yidu yu yanjiu*. Beijing: The Nationalities Press.
- Ouyang, Jueya 1985. *A brief introduction to the Luoba language*. [in Chinese]. Beijing: Mǐnzú Chūbǎnshè.
- Ozanne-Rivierre, Françoise 1997. Spatial reference in New Caledonian languages. *Referring to space. Studies in Austronesian and Papuan languages*. ed. by G. Senft 84-100. Oxford Studies in Anthropological Linguistics. Oxford: Oxford University Press.
- Post, Mark W. 2011. Topographical deixis and the Tani languages of North East India. *North East Indian Linguistics*, 3. ed. by Hyslop, Gwendolyn, Morey, Stephen, and Mark W. Post, 137-154. New Delhi, India: Cambridge University Press India.
- Post, Mark W. ined. *The environmental shaping of language: topographical deixis from the Himalayas to Far North Queensland*. Conference handout, unpublished
- Sapir, Edward 1912. Language and environment. *American Anthropologist* 14.226-242.
- Schapper, Antoinette 2014. Elevation in the spatial deictic systems of Alor-Pantar languages. *The Alor-Pantar languages: History and typology*. ed. by Marian Klamer, 247-284. Berlin: Language Science Press.
- Sun, Hongkai 1983a. *Yidu Luobayu gaiyao* [A brief introduction to Idu (Luoba) language]. [In Chinese]. *Mǐnzú Yǔwén* 6.63-79. Beijing: Mǐnzú Chūbǎnshè.
- Sun, Hongkai 1983b. *The languages of the peoples of the Six River Valley Region and their genetic classification*. [In Chinese]. *Mǐnzú Xuébào*. Kunming: Yunnan People's Publishing Company.
- Sun, Hongkai 1999. On the Himalayan languages of the eastern Himalayan area in China. *Linguistics of the Tibeto-Burman Area* 22:61-72.
- Sun, Hongkai, Shaozun Lu, Jichuan Zhang & Jueya Ouyang. 1980. *Menba, Luoba, Dengren de yuyan* [The languages of the Moba, Luoba and Deng people]. [In Chinese]. Beijing: Zhōngguó Shèhuì Kēxué Chūbǎnshè.
- Wolff, H. Ekkehard 2006. Encoding topography and direction in the verbal systems of Lamang and Hdi (Central Chadic). *Studies In African Linguistics* 35:221.

NOTES ON NOMINAL MARKING AND NOUN PHRASE ELEMENTS IN KINARAY-A

Marvin C. CASALAN
University of Antique
marvin_casalan@dlsu.edu.ph

Shirley N. DITA
De La Salle University
shirley.dita@dlsu.edu.ph

Abstract

This paper aims to discuss the nominal marking system of the Kinaray-a language. Kinaray-a is the language of the Karay-a people, who live in the provinces of Antique, Capiz, Guimaras, Iloilo, Palawan (KWF 2016) and some mountainous areas bordering Aklan and Antique in the Philippines. However, the Kinaray-a variety discussed in the present study is the one spoken in the southern part of Antique Province. The analysis is based on written and spoken corpora which include Kinaray-a online news reports, literary texts, utterances in naturally occurring conversations, and spoken-like narratives. The paper discusses a prototypical noun phrase of Kinaray-a among other Philippine languages. A Kinaray-a noun phrase may contain two categories of nominal markers: determiners and demonstratives. Furthermore, the paper supports Tanangkingsing's (2009) description that noun phrases in Philippine languages may contain case markers, plural markers, determiners, possessive or genitive pronouns, numerals, modifiers, or ligatures, as they are also evident in Kinaray-a. Notes on types of Kinaray-a nouns, their morphological formation and linguistic elements, such as bare nouns, borrowed nouns, and affixed nouns are also presented in the appendix part of paper.

Keywords: nominals, nominal marking system, Kinaray-a, Philippine language, language documentation

ISO-39: krj

1 Introduction

The Ethnologue describes Kinaray-a as an Austronesian language in the West Bisayan subgroup of the Malayo-Polynesian branch. This is the language of the people in the Province of Antique of the Philippines. It is also spoken by Karay-a people who live in mountainous areas bordering Aklan and Antique (Abadiano 1980) and in the different provinces of Capiz, Guimaras, Iloilo, and Palawan (KWF, 2016).

According to Tiongson (1994), the term Kinaray-a is derived from *iraya* meaning 'upstream,' prefix *ka* meaning 'companion' and an *in* meaning 'to have undergone something'. Alternate names are Hiniray-a, Karay-a, and sometimes Hamtikanon (Delos Santos, 2010) and Antiqueño (KWF, 2016). Simons and Fennig (2018) report that there are approximately 380,000 speakers of the language and that it has a language status of 4 (educational). Currently, the language is recognized by Department of Education (DepEd) - Philippines as one of the major languages under the Mother Tongue-Based Multilingual Education (MTBMLE curriculum). It is also commonly used in tourism signage in the Province of Antique, as well as in internet news reports and Kinaray-a writers' published literary works. Delos Santos (2003) claimed that the actual number of speakers is undetermined as the language has long been incorrectly classified as *Ilongo* (Hiligaynon). According to him, Hiligaynon was the only primary language considered in Region VI before 1980s as it was the language spoken by those who were dominant in the local government, religion, education, and culture. Kinaray-a was only recognized as a language later because of the development of Kinaray-a literature in the 1980s.

Figure 1: Distribution of Kinaray-a Language in the Philippines
(Atlas ng mga Wika ng Filipinas, 2016:100)



In this paper, the Kinaray-a nominals and the nominal marking system are described. It will emphasize the constituent order of a typical Kinaray-a noun phrase (hereafter, NP), the types of nominal markers used in the language, and the functions of these nominal markers. The morphological formation of Kinaray-a nouns will also be presented.

The Kinaray-a variety in the southern part of Antique Province is considered in this study. Ergative case-marking is used in the analysis of the utterances as the language exhibits ergative constructions. The data are drawn from both spoken and written corpora. Spoken data include recorded conversations, homilies, pear stories, and narratives in cooking dishes, while online news reports and literary works are utilized as the written data.

2 Nominal Markers

A Kinaray-a NP may contain a determiner or demonstrative and a head noun or pronoun. The nominal markers take the initial position of NPs. The phrases below are the examples of how Kinaray-a NPs are constructed:

Determiner

1. *ang* *bata*
 det child
 ‘the child’

Demonstrative

2. *dya* *nga* *bata*
 dem lig child
 ‘this child’

The utterances 3 and 4 are examples of nominalized clauses introduced by the determiner *ang*.

3. *Ang ginhambal nanda.*
 det say-pfv erg.3p
 ‘What they said.’
4. *Ang ginahindian na.*
 det like-neg erg.3s
 ‘What he/she doesn’t like.’

Also, Kinaray-a NPs require a ligature to connect the nominal marker and a modifier to the head noun, as in example 5.

5. *Ang manani nga panahon.*
 det nice lig weather
 ‘A nice weather.’

Although it is common for Kinaray-a NP to have determiner as nominal marker, there are some cases in which a marker is not required. These instances are in vocatives and answers to question constructions, as in examples 6 and 7 respectively.

Vocative

6. *Ginoo, kaloy-i kami*
 Lord mercy abs.1p
 ‘Lord, have mercy on us’

Answer to a question

- Question: *Ano ang ngaran mo Ma’am?*
 q-what det name abs.2p hon
 ‘What is your name, Ma’am?’

7. Answer: *Elvie*
 Elvie
 ‘Elvie’

2.1. Determiners

Like other Philippine languages, determiners in Kinaray-a encode number (singular and plural), case (absolute, ergative/genitive, and oblique/locative), and they distinguish between common and personal nouns. Table 1 presents Kinaray-a determiners.

Table 1. The Kinaray-a Determiners

Case	Common		Personal	
	Singular	Plural	Singular	Plural
Absolute	ang	ang mga	si	sanday
Ergative/Genitive	ka(ng)	ka(ng) mga	ni	nanday
Oblique/Locative	sa	sa mga	kay	kanday/kananday

For the determiner *ang*, it may precede a single word noun (8); an abstract noun (9); a relativized clause (10); a verb-like item (11); a preposition (12); an honorific (13); and before an NP with modifiers (14):

8. *ang balay*
det house
'the house'
9. *ang paglaum*
det hope
'the hope'
10. *ang tawo nga tandus*
det person lig industrious
'the industrious person'
11. *ang gingamit nga pangrara*
det use-pfv lig to.weave
'the (material) used for weaving'
12. *ang para kana*
det prep abs.3p
'for her/him'
13. *ang Mayor*
det hon
'the Mayor'
14. *ang mga nami kag tibay nga hilo*
det plu nice conj strong lig thread
'the nice and strong thread'

The markers that introduce a core argument are called *core nominal markers*. If the core argument is a full noun phrase, it is usually introduced by a determiner, as in 15. Otherwise, it may utilize an absolutive pronominal with the absence of a marker, as in 16.

15. *Nagpanaw ang bata.*
walk-pfv det child
'The child walked away.'
16. *Nagpanaw tana.*
walk-pfv abs.3s
'He/She walked away.'

2.1.1. Nominal marker 'sa'

For the oblique marker *sa*, it usually precedes locative nouns, as in sentences 17 and 18.

17. *Duro gemstone ang makita sa suba.*
more gemstone abs find-ipfv obl river
'(You can) find more gemstones in the river.'
18. *Lain gid ang pangabuhi sa uma.*
different par abs lifestyle obl farm
'The lifestyle is really different living in the farm.'

The marker *sa* may also signify temporal marking such as specific times or parts of the day in an oblique phrase, as in sentence 19.

19. *Sanag ang mga bituon sa kagab-ihun.*
Sanag ang mga bituon sa kagab?ihun
 bright abs plu star obl night
 ‘The stars (shine) brightly at night.’

The nominal marker *sa* can also mark the oblique case in a benefactive construction.

20. *Nagtugro tana ka bugas sa mga katawhan.*
 give-pfv abs.3s det rice obl plu people
 ‘He/She gave rice to the people.’

2.1.2. Nominal marker ‘*si/sanday*’

The marker *si* is the counterpart of *ang* for personal nouns. It marks a particular absolutive personal argument. *Sanday* is the plural counterpart of *si*. An example of the nominal singular marker *si* is presented in sentence 21, and the plural counterpart *sanday* is in sentence 22.

21. *Nahulog si Mark sa kahoy.*
 fall-pfv abs Mark loc tree
 ‘Mark fell (from the) tree.’
22. *Sanday nanay kag tatay lang ang tawo sa balay.*
 abs mother conj father par det people loc house
 ‘Only mother and father are at home.’

2.1.3. Nominal marker ‘*ni*’.

The nominal marker *ni* is the counterpart of *ka(ng)* for personal nouns and serves as the agent of a transitive verb or as the possessor in a possessive construction.

In transitive constructions, the common determiner *ka* and personal determiner *ni* introduce the agent, and the determiner *ang* introduces the patient. Interestingly, the determiners *ka* and *kang* (23a and 23b) may be used interchangeably without any semantico-syntactic impact

- 23a. *Ginkaun ka bata ang peras.*
 eat-pfv erg child abs pear
 ‘The child ate the pear.’
- 23b. *Ginkaun kang bata ang peras.*
 eat-pfv erg child abs pear
 ‘The child ate the pear.’
24. *Ginkaun ni Ruzel ang pagkaun ni Allysa.*
 eat-pfv erg Ruzel abs food gen Allysa
 ‘Ruzel ate the food of Allysa.’

When the pronominal counterpart is used, the agent is marked by an ergative determiner and the patient is marked by an absolutive determiner, as in sentence 25.

25. *Ginturuk na ang kanding nga nagapanaw.*
 watch-pfv erg.3s abs goat lig walk-cont
 ‘He/She watched the goat walking.’

Similar to Cebuano (Tanangkingsing 2009) and Hiligaynon (Santos 2012), Kinaray-a NPs may exhibit any of these patterns. It is interesting to note, however, that the ligature *ka* is only used in numerals, as in sentence 29, while the sentences 27, 28, and 30 co-occur with *nga*.

1. Determiner + *mga* + Noun

26. *ang mga tawo*
 det plu people
 ‘the people’

2. Determiner + Possessor Pronoun + *nga* + Noun

27. *ang anda tanan nga gamit*
 abs gen.3p whole lig thing
 ‘(all) of their things’

3. Determiner + Modifier + *nga* + Noun

28. *ang pwerte kalapad nga parayan*
 abs par wide lig rice.field
 ‘the wide rice field’

4. Determiner + Numeral + *ka* + Noun

29. *sa sara ka basket ka peras*
 obl one lig basket gen pear
 ‘in one basket of pear’

5. Modifier_{clausal} + *nga* + Noun

30. *Bahol nga bata*
 big lig kid
 ‘(a) big kid’

2.2. *Demonstratives*

Demonstratives or deictic pronouns in Philippine languages are generally divided into three sets expressing nearness to the speaker, nearness to the addressee, and remoteness from either of the two (McFarland 2008). The same feature is also evident in Kinaray-a. Table 2 enumerates the demonstratives in Kinaray-a.

Table 2. *The Kinaray-a Demonstratives*

Case	Proximal (near the speaker)		Medial (near the listener)		Distal (far from both)	
	<i>Kinaray-a</i>	Gloss	<i>Kinaray-a</i>	Gloss	<i>Kinaray-a</i>	Gloss
Absolutive	dya(ay)	this	ra(ay) ria(n)	that	to(ay)	that
Ergative	dya(ay)	this	ra(ay) ria(n)	that	to(ay)	that
Oblique	(ri/ru)dya (ri/ru)gya	here	(ri/ru)dyan (ri/ru)gyan	there	(rig/rug)to	there

It is interesting to note that Kinararay-a demonstratives have two or more variants for all three spatial orientations. Each deictic pronoun has several forms to express different cases (McFarland 2008). In Kinararay-a, nouns in the absolutive case may be preceded by *dya(ay)* ‘this’, *ra(ay)* ‘that’, or *to(ay)* ‘that’ (yonder)

In Kinararay-a, the demonstrative *dya* (or the variant *dyaay*) ‘this’ is used as a proximal absolutive deictic pronoun. But *dya* may be used with an additional emphatic form *amo*, which seems to be an all-purpose deictic that can stand alone and expresses emphasis without spatial reference. The examples of how these deictic pronouns are used are presented in sentences 31a and 31b.

31a. *Dya ang akun bata.*
 prox abs gen.1s child
 ‘This is my child.’

31b. *Amo dya ang akun bata.*
 emp prox abs gen.1s child
 ‘This is my child.’

32a. *Kanakun ang balay nga dya*
 gen.1s abs house lig prox
 ‘This house is mine.’

Based on the sentences 31a to 32a, it can be claimed that Kinararay-a demonstratives may either be pre-nominal or post-nominal. Two features should be noted here. First is the use of the ligature *nga*. If a noun phrase precedes the demonstrative *dya*, it may be preceded with a ligature *nga*, as shown in 32a. Consequently, if the noun phrase appears after *dya*, *nga* is placed after *dya* as illustrated in 32b.

32b. *Kanakun ang dya nga balay.*
 gen.1s abs prox lig house
 ‘This house is mine.’

A construction where a genitive appears in between the demonstrative and the nominal is also grammatical.

32c. *Dya ang akun nga balay.*
 prox abs gen.1s lig house
 ‘This is my house.’

However, constructions without *nga* are also considered grammatical, but the use of *nga* makes the sentence more emphatic.

32d. *Dya ang akun balay.*
 prox abs gen.1s house
 ‘This is my house.’

For *ra(ay)* ‘that’ in 33 to mark medial demonstratives, and *to(ay)* ‘that’ in 34 to mark distal demonstrative, the same grammatical features are applied with *dya(ay)* ‘this’ in the absolutive case as discussed above.

33. *Amo ra ang niyog siguro.*
 emp med abs coconut maybe
 ‘Maybe, that’s the coconut.’

34. *Naghapus* *run* *to* *ang* *langka*.
 soft-pfv par dis abs jackfruit
 ‘The jackfruit has softened already.’

It is noted that word classification in Philippine languages is based on the interrelationship among various affixes and their use in discourse (Santos 2012). For instance, there are noun roots or nouns derived from verbs, adjectives, or any grammatical category (cf. Dita 2011). There is also no clear lexical distinction between nouns and verbs in Philippine-type languages, which is a unique feature (Amerila 2018) of Philippine languages. Nolasco (2011) also asserts that grammatical categorization of Philippine-type languages, particularly root forms, has proven difficult, especially at their boundaries.

Kinaray-a nominals share these morphological properties with Philippine languages. Notes on the different types of Kinaray-a nouns especially on the affixation process are presented in Appendix A of the paper.

3 Conclusion

This paper describes the constituent order of a typical Kinaray-a NP. Determiners and demonstratives are the two types of Kinaray-a nominal markers. These markers usually introduce the head nouns in NPs. The Kinaray-a determiners are *ang*, *ka(ng)* and *sa* for common nouns, which are basically singular, but they can be made plural by adding the marker *mga* after the determiner, including *ang mga*, *ka(ng) mga*, and *sa mga*, while the personal determiners are *si*, *ni*, and *kay*, and their plural counterparts are *sanday*, *nanday*, and *kanday* respectively. This paper also highlights the types of deictic pronouns based on spatial orientation, that is, proximal, medial, and distal, which can be marked in the absolutive, ergative, and oblique cases. Finally, in the Appendix, notes are provided on types of nouns and nominalizing prefixes.

References

- Abadiano, Ted. 1980. *Kinaray-a Phonemics and Morphophonemics*. Microfiche Order. Summer Institute of Linguistics.
- Amerila, Zara Mazelene A. 2018. A grammatical sketch of Kinaray-a (Unpublished thesis). University of the Philippines, Diliman, Quezon City.
- Atlas ng mga Wika ng Filipinas* (Philippine Language-Atlases). 2016. Komisyon sa Wikang Filipino
- Delos Santos, Alex C. 2010. Kinaray-a. In G. S. Zafra (ed.). *Ambagan, Mga Salita sa Iba't Ibang Wika sa Filipinas*. Quezon City: The University of the Philippines Press, 136-143.
- De los Santos, Alex C. 2003. *The Rise of Kinaray-a: History and Anthology of Contemporary Literature in Antique*. Iloilo City: Libro Agustino.
- Dita, Shirley N. 2011. The Structure of Ibanag Nominals. *Philippine Journal of Linguistics* 42:41-57.
- McFarland, Curtis D. 2008. Deictic pronouns in Philippine languages. *Studies in Philippine Languages and Cultures* 19: 123-141.
- Nolasco, Ricardo. M. 2011. Grammar notes on the national language. Retrieved from <https://vdocuments.net/reader/full/grammar-notes-on-the-philippine-language>
- Santos, Marilou C. 2012. A contemporary grammar of Hiligaynon. (Unpublished doctoral dissertation). De La Salle University, Philippines.
- Simons, Garry. F. and Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the World* (21st ed.). Dallas, Texas: SIL International. Retrieved from Online version: <http://www.ethnologue.com>
- Tanangkingsing, Michael. 2009. A functional grammar of Cebuano (Unpublished doctoral dissertation). National Taiwan University, Taipei, Taiwan.
- Tiongson, Nicanor G. 1994. *CCP Encyclopedia of Philippine Art: Peoples of the Philippines, k to y*. Manila: Cultural Center of the Philippines.

Appendix A - Notes on Types of Nouns in Kinaray-a

In this section of the paper, the two classes of Kinaray-a nominals are presented: unaffixed nouns are depicted in sub-sections A.1 and A.2, which feature bare nouns and borrowed nouns, respectively, whereas section A.3 illustrates how nouns are generated by affixation.

A.1 Bare nouns

In this paper, we refer to nouns without affixes as bare nouns. Body part terms are an example of bare nouns. The Table 3 lists the following Kinaray-a body part words together with their English equivalents will be used as instances of bare nouns.

Table 3. Body Part Terms in Kinaray-a

Root	Gloss	Root	Gloss
<i>talinga</i>	‘ear’	<i>Butkun</i>	arm
<i>siko</i>	‘elbow’	<i>iruk</i>	armpit
<i>mata</i>	‘eye’	<i>likod</i>	back
<i>kiray</i>	‘eyebrow’	<i>lawas</i>	body
<i>pungyahun</i>	‘face’	<i>tul-an</i>	bone
<i>kahig</i>	‘feet’	<i>suso</i>	breasts
<i>tudlo</i>	‘finger’	<i>buli</i>	buttocks
<i>buon-buonan</i>	‘fontanel’	<i>bagi-ing</i>	cheek
<i>dahi</i>	‘forehead’	<i>dughan</i>	chest
<i>alima</i>	‘hand’	<i>silang</i>	chin

The utterances 35 and 36 below use *kahig* (foot) and *alima* (hand) respectively as examples of body part terms as bare nouns. They are unaffixed words and function as the heads of the phrases.

35. *May mga agi kang kahig nga makita sa daray-ahan.*
May mga agi kang kahig nga makita sa daray?ahan.
 exist plu trace det foot lig see-cont obl seashore
 ‘There are traces of footprints which can be seen in the seashore.’
36. *Kinahanglan nga limpyo ang alima bag-o magkaun.*
 stat-should lig clean abs hand before ipfv-eat
 ‘(Your) hands should be clean before eating.’

A.2 Borrowed nouns

Spanish and English loanwords make up the majority of Kinaray-a borrowed nouns. This could be due to the cultural influence of the Spanish and American occupations in the Philippines, as well as the fact that Kinaray-a speakers are bilingual, with English as a second language. Table 4 shows a few examples of borrowed nouns.

Table 4. Kinaray-a Borrowed Nouns

Borrowed Nouns	Origin	Gloss
<i>gobernadora</i>	Spanish	governor
<i>domingo</i>	Spanish	Sunday
<i>softcopy</i>	English	softcopy
<i>telebisyon</i>	English	television
<i>refrigerator</i>	English	refrigerator

37. *Ginpaathag man kang gobernadora ang sabahin*
 pfv-explain also det lady.governor det about
kang mga kabataan nga nadura sa listahan
 det plu youth lig misplace-pfv obl list
kang provincial scholars
 det provincial scholars.
 ‘The lady governor also explained about the youth whose names were misplaced in the list of provincial scholars.’

The Kinaray-a word *governadora* (woman governor) is used in sentence 37. It is derived from the Spanish word ‘*governadora*,’ which is the feminine equivalent to the male ‘*governador*.’

Another Kinaray-a loaned word is used in sentence 38. The word *domingo* (Sunday) is Spanish loanword to refer to a weekday.

38. *Kaduro ka tawo sa tinda kun domingo.*
 a.lot det people obl market during Sunday
 ‘There are a lot people in the market during Sundays.’

A.3 Affixed nouns

Some nouns in Kinaray-a may be formed by affixation. These nominals are derived nouns because they are generated by attaching nominal affixes to roots or stems to form derived nouns. Table 5 presents three samples of Kinaray-a nominalizing prefixes.

Table 5: Morphological Derivations of the Kinaray-a Nominals

Affix	Root	Gloss	Affixed	Gloss
ka-	subu	sad	kasubu	sadness
	lapad	wide	kalapad	wideness
	istorya	to talk	kaistorya	person you are taking to
	imaw	to accompany with	kaimaw	companion
	akay	to ride	kasakay	person together on board
paN-	raha	cook	pangraha	used for cooking
	balay	house	pangbalay	used in the house
	lagaw	walk	panglagaw	used for walking
taga-	Antique	Antique	taga-Antique	from Antique
	banwa	town proper	taga-banwa	from the town proper
	butung	pull	taga-butung	person who pulls something

The prefix *ka-* is an abstract nominalizer or a reciprocal action when it is attached to the root word *subu*. For instance, the root *subu*, which means ‘sad’, becomes *kasubu* ‘sadness’. Meanwhile, when the prefix *ka-* is attached to the root *istorya* or ‘to talk’, it derives the noun *kaistorya* ‘person you are talking to’.

39. *Nakabatyag tana ka kasubo*
 pfv-feel abs.3s det sadness
 ‘He/She felt the sadness.’
40. *Si Lyjoe ang ana kaistorya kang nagaraha sanda.*
 Det Lyjoe abs erg.3s person.taking.to when pfv-cook abs.3pl
 ‘It was Lyjoe who he/she was speaking to when they were cooking.’

The prefix *taga-* refers to a person's place of origin. In this case, the prefix is preceded by a particular name of place, as in sentence 41.

41. *Mga karay-a ang taga-Antique.*
 plu karay-a det ori-Antique
 'The karay-a (speakers) are from Antique.'

In a nominalized term, the prefix *taga-* also denotes the action's doer. Therefore, *taga-* acts as an initiator in this circumstance, as in sentence 42. When a prefix is added to a verb's base form, it indicates that a person has been assigned or hired to carry out the action represented by the base word. The Tagalog equivalent is also *taga-*. Thus, one can use either prefix, and the meaning is the same.

42. *Kinahanglan ka tagabutong ka hilo sa makina.*
 need det ini-pull det thread loc machine
 'A person is needed to pull the thread in the machine.'

Appendix B

1	-	1 st person
2	-	2 nd person
3	-	3 rd person
abs	-	absolutive
conj	-	conjunction
cont	-	continuative
dem	-	demonstrative
det	-	determiner
dis	-	distal
emp	-	emphatic marker
erg	-	ergative
exist	-	existential
gen	-	genitive
hon	-	honorific
ini	-	initiator
ipfv	-	imperfective
lig	-	ligature
loc	-	locative
med	-	medial
neg	-	negative
obl	-	oblique
ori	-	origin
plu	-	plural
par	-	particle
pfv	-	perfective
plu	-	plural marker
prep	-	preposition
prox	-	proximal
q	-	question word
s	-	singular
stat	-	stative

NOMINALIZATION AND ITS VARIOUS FUNCTIONS IN LIANGMAI

Kailadbou DAIMAI
CSLLC, RUSA.2. Utkal University
kailadboudaimai@yahoo.co.in

Abstract

The purpose of this paper is to present the extent of nominalization and its functions in Liangmai. The study includes lexical or derivational nominalization and structures involving the nominalization of clauses. This paper briefly describes the morpho-syntax of nominalization and the constructions in which nominalized forms occur in Liangmai. The nominalizers in this language include the suffixes *-bo*, *-bam*, and *-mai* and the prefixes *kə-* and *pə-*. Lexical nominalization includes derivation of abstract nouns from stative root verbs, agentive nominals from both nouns and verbs, adjectivals from stative intransitive verbs and gerunds. The nominalizing suffix *-bo* is the most common and is highly productive. All verbal roots can be nominalized by the suffix *-bo*, and the resulting forms can be interpreted as abstract nouns, gerunds, attributive adjectives, relative clause, among others. In Liangmai, the suffix *-bo* functions variously as a nominalizer, relativizer or complementizer. Liangmai exhibits externally headed relative clauses in which the head nouns appear to the right of relative clauses.

Keywords: Liangmai, Nominalization, Relativization, Complementation, Derivational, Clausal, Tibeto-Burman

ISO 639-3 codes:njn

1 Introduction

Nominalization is one of the most prominent characteristics and a highly productive phenomenon of the Tibeto-Burman (henceforth TB) languages (Bickel 1999, Watters 2006, 2008, Genetti et al. 2008, 2011, Noonan 1997, Delancey 2002, 2011). It refers to the process by which we derive nominals from a word of another class. Yap et al. observe that:

...nominalization in its core sense refers to the process by which we derive nominal expressions – for example, from verbs (e.g., watch > watcher) or adjectives (e.g., narrow > narrowness, narrowing). Clauses may also be nominalized (e.g., awaken the public conscience > awakening (of) the public conscience. (Yap et al. (2011:3)

The study of nominalization in TB languages begins with Matisoff's seminal paper 'Lahu nominalization, genitivization, and relativization' (1972). He showed that the functions of nominalization, relativization and genitivization in Lahu are marked by the same particle *ve*. A similar complex of functions revolving around a single morpheme occurs in other TB languages (Delancey 2002:56). This morphological convergence of syntactic functions was dubbed the 'Standard Sino-Tibetan Nominalization' (SSTN) pattern (Bickel 1999:271) and has been reported in many studies of a number of TB languages. Two levels of nominalization process are observed: lexical or derivational and clausal. Derivational nominalization refers to the process which creates lexical nouns from words of other lexical categories (usually the verb root), while clausal nominalization operates in the domain of clause and works at syntactic level to allow a grammatical clause to be treated as a noun phrase within a broader syntactic context (Genetti et al. 2008). Nominalized clauses are used in a wide range of functions and syntactic structures in TB languages, including attributive clauses, adverbial clauses, nominal-complement constructions, relative clauses, free-standing independent clauses, and so on.

Both these nominalization processes are attested in Liangmai, a TB language spoken in Manipur and Nagaland. The primary purpose of this paper is to present a brief overview of nominalization in Liangmai. Different nominalizers found in the language are *-bo*, *-bam*, *-mai*, *kə-* and *pə-*. The uses and functions of these nominalizers in the language are discussed in the following sections.

2 Language Classification and Typological Overview

In the classification of Tibeto-Burman languages in the Linguistic Survey of India (LSI), Grierson (1903) was slightly hesitant as to whether Liangmai (called ‘Kwoireng or Liyang’) be placed under Naga-Bodo or Naga-Kuki sub-group. He stated:

Their language appears to be an intermediate one between the Naga-Bodo and the Naga-Kuki group. The pronouns agree best with the latter, and so I class it here, though its geographical position would incline one to put it with the former set of languages (Grierson 1903:462).

Other than some lexicons listed under ‘Kwoireng or Liyang’, no description of the language was provided in LSI. Marrison (1967), in his comprehensive survey of the languages of Northeast India, placed Liangmai in his posited Nruanghmei group, ‘Type B-3’, along with Zeme, Mzieme, Nruanghmei (formerly Kabui), Khoirao, Maram and Puiron. The close affinity among the Zeme, Liangmai and Rongmei is also reflected in Bradley’s (1997) classification of Tibeto-Burman languages in which Liangmai was placed under the Zeliangrong group, which falls under the Southern Naga of the Kuki-Chin-Naga. More recently, Burling (2003) grouped Liangmai, along with Zeme and Rongmei, under the Zeme group, which was again referred to as Western Naga in Post and Burling (2017).

Two major word classes in Liangmai include nouns and verbs. Most nouns are free roots, and they can be monosyllabic, disyllabic or trisyllabic. Nouns with more than three syllables are mostly compounds. There are also bound noun roots, and these take one of the four formatives *mə-*, *tə-*, *tsə-* and *kə-* to become free standing forms. These formatives may or may not be retained in compound formation; however, it is usually dropped when a noun is used with pronominal personal pronouns. The semantics of these formatives is not yet fully known. All verbal roots are bound, and they can be free standing words only if they are minimally affixed (Daimai & Singha 2020:127). There are both monosyllabic and disyllabic verb roots, but trisyllabic roots are rare. Verb roots are also used to derive adverbs and adjectivals. A verb is inflected for tense, aspect and mood. Liangmai is a tonal language having at least three contrastive lexical tones, namely high, mid and low. The language is highly agglutinative.

As with most TB languages, Liangmai is verb-final. The A argument of a transitive clause is typically marked with agentive *-nūu*, though it is not obligatory, whereas the animate O argument is marked with a primary object marker *-tū*. The inanimate O remains unmarked. OAV word order is also permissible, and such word order is used when it involves emphasis on the O argument. Noun phrases are head-medial with variable ordering of modifiers. They consist of an obligatory noun head which is preceded and followed by optional modifiers. Adjectivals can occur in both pre- and post-head positions. Demonstratives occur in pre-head position, while numerals occur in post-head position. To signal possession, personal pronoun pro-clitics are prefixed to the root of a noun, as in *əlù* (1-POSS farm) ‘my farm’ and *nəlù* (2-POSS farm) ‘your field’. The possessor noun phrase can also be marked with the possessive enclitic *-gù*, suffixed to the possessor of the head noun, as in *ə-gù tsəlù* (1-POSS farm) ‘my farm’ and *nə-gù tsəlù* (2-POSS farm) ‘your farm’. The possessive enclitic *-gù* marks the possessor of noun phrase *i* ‘I’ and *nan* ‘you’, which precedes the head noun *tsəlù* ‘farm’. The dependent clause generally precedes the matrix clause.

3 Data and Methodology

According to the Ethnologue (Eberhard et al., 2021), the 2011 census estimates that there are 49,800 speakers of Liangmai. The Liangmai Naga Council, Manipur (LNC, M), an apex organization of the tribe, also gives a similar figure of 50,000. The majority of Liangmai speakers can be found in the bordering area of Tamenglong district of Manipur and Peren district of Nagaland. There are approximately 100 Liangmai villages, most of which are in the state of Manipur. Each Liangmai village

has a distinct variety, but there is intelligibility to a great extent among the speakers of different villages. There is not much of a media platform for Liangmai, and Liangmai literature is still in an early stage of development. There is a Bible and hymnals translated into Liangmai, and the language used in the Bible serves as the standard form when it comes to writing and literature. Nevertheless, the influence of each own variety can be seen in the few books written so far by Liangmai authors.

For the current study, as a native speaker, I have collected most of the data from introspection and elicitation. A list of words adapted from Swadesh list, expressions used in daily life, and basic phrases and sentences in Liangmai elicited from my principal informant, Phenlakbou Marenmai, aged 67, of Tharon village, Manipur, in 2018-2019 were used for the analysis. In addition, I have also extracted relevant data from available written Liangmai sources, mainly the Holy Bible. I also refer to my old field notes for relevant data.

4 Derivational or Lexical Nominalization in Liangmai

Liangmai has multiple nominalizers that are involved in the derivation of nouns. A lexical noun is derived mainly from verbs and sometimes even from nouns. Deverbal nouns are highly productive and are widely attested in normal discourse and speech acts, and they can head NPs in Liangmai. The nominalizing suffixes in Liangmai are *-bo*, *-bam* and *-mai*. The language also has a nominalizing prefix *kə-* and *pə-*. These nominalizers are discussed in detail in the following sections.

4.1. Nominalizer *-bo*

The nominalizing suffix *-bo* is multifunctional. It is used as a general nominalizer in Liangmai. It is highly productive and can be suffixed to all verb roots to derive deverbal nouns that can head an NP. When this nominalizer is attached directly to a verb, the resulting form, taken out of context, may have several possible interpretations and functions. It may be used to derive abstract nouns, for example, the verb root *pɪŋ* means ‘be afraid’, whereas *pɪŋbo* means ‘fear’, as in (1).

- (1) *pa'* *pɪŋbo* *ha-e*
 3S afraid-NMZ N.COP-DECL
 ‘He has no fear.’

Similarly, the following abstract nouns are derived by suffixing nominalizer *-bo* to stative root verbs.

- (2) (a) *luŋsa* ‘to love’ > *luŋsabo* ‘love’
 (b) *sai* ‘to die’ > *saibo* ‘death’
 (c) *məsan* ‘be clean’ > *məsanbo* ‘cleanliness or holy’
 (d) *kim* ‘to satisfy’ > *kimbo* ‘satisfaction’
 (e) *mətʰá* ‘be happy’ > *mətʰábo* ‘happiness’
 (f) *hú* ‘be brave’ > *húbo* ‘courage’
 (g) *tsəliəŋ* ‘be proud’ > *tsəliəŋbo* ‘pride’

Derived abstract nouns can function as heads of noun phrases, as given in (3).

- (3) *...luŋsa-bo* *si-nīu* *dí-tʰu-e*
 love-NMZ EMP-AGT big-SUP-DECL
 ‘...love is the greatest.’

Action nominals can be formed by suffixing *-bo* to action verbs.

- (4)
- | | | |
|-----|----------------------------|-------------|
| (a) | <i>kap-bo</i> | ‘crying’ |
| (b) | <i>tiu-bo</i> | ‘eating’ |
| (c) | <i>pak-bo</i> | ‘running’ |
| (d) | <i>ken-bo</i> | ‘reading’ |
| (e) | <i>rao-bo</i> | ‘writing’ |
| (g) | <i>zi-bo</i> | ‘sleeping’ |
| (h) | <i>mət^henbo</i> | ‘playing’ |
| (i) | <i>mək^hiubo</i> | ‘coughing’ |
| (j) | <i>kəriabo</i> | ‘squeezing’ |

Nominalized action verbs can function as noun modifiers, as given in (5), and can also function as heads of noun phrases, as in (6).

- (5)
- | | | | | | |
|----------------|---------------|--------------|--------------|------------|-----------|
| <i>naŋ-nīu</i> | <i>rao-bo</i> | <i>əriak</i> | <i>delam</i> | <i>bam</i> | <i>lo</i> |
| 2P-AGT | write-NMZ | book | where | EXST | IMP |
- ‘Where is the book you wrote?’
- (6)
- | | | | |
|-------------------|-----------------|-----------|-----------|
| <i>pak-bo-nīu</i> | <i>əliu-leŋ</i> | <i>wi</i> | <i>ne</i> |
| run-NMZ-AGT | 1PL-BEN | good | IRR |
- ‘Running will be good for us.’

Attributes such as dimension, age, value and color are expressed by nominalized stative intransitive verbs when used as noun phrase attributes, or as verbless clause complements in ascriptive clauses. These are derived from stative intransitive verbs using the nominalizing suffix *-bo*; *t^heŋbo* be.long-NMZ ‘long’, *wibo* be.good-NMZ ‘good’, and others.

- (7)
- | | | | | | |
|-----|-----------------------|-------------|---|-------------------------|----------|
| (a) | <i>siam</i> | ‘be small’ | > | <i>siambo</i> | ‘small’ |
| (b) | <i>dí</i> | ‘be big’ | > | <i>díbo</i> | ‘big’ |
| (c) | <i>heŋ</i> | ‘be red’ | > | <i>heŋbo</i> | ‘red’ |
| (d) | <i>sán</i> | ‘be new’ | > | <i>sánbo</i> | ‘new’ |
| (e) | <i>k^ha</i> | ‘be bitter’ | > | <i>k^habo</i> | ‘bitter’ |
| (f) | <i>rai</i> | ‘be first’ | > | <i>raibo</i> | ‘first’ |
- (8)
- | | |
|--------------|---------------|
| <i>əriak</i> | <i>sán-bo</i> |
| book | be.new-NMZ |
- ‘A new book.’
- (9)
- | | |
|--------------|--------------|
| <i>tsəki</i> | <i>dí-bo</i> |
| house | be.big-NMZ |
- ‘A big house.’
- (10)
- | | | | |
|-------------|---------------------------|---------------|---------------------|
| <i>əpui</i> | <i>tsəp^hai</i> | <i>kek-bo</i> | <i>tsərui-bam-e</i> |
| 1.mother | cloth | tear-NMZ | stitch-PROG-DECL |
- ‘My mother is stitching torn cloth.’

Gerunds are also formed by adding the nominalizing suffix *-bo* to verbs.

- (11) *záo sak-bo tsəpúm-leŋ sa-e*
 wine drink-NMZ body-BEN bad-DECL
 ‘Drinking is injurious to health’ (Lit: Drinking wine is bad for body).
- (12) *mət^hiu-nū wí pui kap-bo ŋou-e*
 everyone-AGT DEM woman cry-NMZ see-DECL
 ‘Everyone sees/saw that woman crying.’

Additionally, *-bo* can be used to nominalize clauses as well. Clausal nominalization is discussed in Section 5.

4.2. Nominalizer *-bam*

The nominalizer *-bam* derives a locational noun from action verbs. The derivation expresses a meaning of ‘the place where VERB’. It is a highly productive process applying to both transitive and intransitive verb roots. A few examples of locational nominalized forms are given below.

- (13) (a) *zúŋ* ‘to pee’ > *zúŋ-bam* ‘place to pee’
 (b) *kəhum* ‘to pray’ > *kəhum-bam* ‘place to pray’
 (c) *əláŋ* ‘to cook’ > *əláŋ-bam* ‘place to cook’
 (d) *tiu* ‘to eat’ > *tiu-bam* ‘place to eat’
 (e) *zi* ‘to sleep’ > *zi-bam* ‘place to sleep’
- (14) *zúŋ-bam delam lo*
 pee-NMZ where IMP
 ‘Where is the place to pee?’ or ‘Where is the peeing place?’
- (15) *námai-duŋ-tù kəhum-bam-ga pí-gut tu lo*
 child-PL-DAT pray-NMZ-LOC CAUS-enter PROH IMP
 ‘Do not let children enter the place of praying.’

4.3. Nominalizer *-mai*

In Liangmai, nouns and verbs can be nominalized by the agentive nominalizer *-mai* to derive an agentive noun. This nominalizer is derived from the noun *tsəmai* meaning ‘man or person’. Its function is similar with the Mongsen Ao agentive nominalizer *-eɿ* (Coupe 2007:263). This derivation usually denotes either a type of agent or a referent whose habitual activity is characterized by the meaning denoted by the nominal base in the case of nominalization of noun (literally: the one who has to do with noun) as in (16a) and (16b). In the case of nominalization of a verbal base, it denotes the meaning ‘the one who does (verb)’ as in (16c) to (16f). This type of nominalization is restricted to nouns with human referents.

- (16) (a) *nám-mai*
 village-NMZ
 ‘villager’ (Lit.: the one in or from village)
- (b) *tsəri-mai*
 war-NMZ
 ‘warrior’
- (c) *ken-mai*
 to read-NMZ
 ‘reader’

- (16) (d) *mət^hen-mai*
to play-NMZ
'player'
- (e) *kamsat-mai*
to kill-NMZ
'killer'
- (f) *əláŋ-mai*
to cook-NMZ
'chef/cook'

The scope of this nominalizer includes nominalization of clauses, as given in (17) and (18).

- (17) *əriak-ki* *tat-mai-duŋ-tù* *tət^huan* *pí* *ne*
school go-NMZ-PL-DAT prize give IRR
'The ones who go to school will be given prizes.'
- (18) *naŋ-nīu* *si-mai* *sou-sou* *lo*
2SG-AGT know-NMZ who-who IMP
'Who are the ones you know?'

4.4. The nominalizing prefix *kə-*

The nominalizing prefix *kə-* is found to be attached on a handful of verbs creating abstract nouns.

- | | | | | | | |
|------|-----|------------------------|---------|---|---------------------------|------------------------|
| (19) | (a) | <i>sai</i> | 'die' | > | <i>kə-sai</i> | 'death' |
| | (b) | <i>t^hiu</i> | 'pain' | > | <i>kə-t^hiu</i> | 'boil, furuncle' |
| | (c) | <i>tat</i> | 'go' | > | <i>kə-tát</i> | 'journey, mission' |
| | (d) | <i>tiu</i> | 'eat' | > | <i>kə-tiu</i> | 'eatables' |
| | (e) | <i>sak</i> | 'drink' | > | <i>kə-sak</i> | 'drinks (liquid food)' |

Konnerth (2014) reports a nominalizer *ke-* (with allomorphs *ki* ~ *ka*) in Karbi. She writes, 'this nominalizing velar prefix has many apparent cognates across several branches of Tibeto-Burman both inside and outside Northeast India, which is productive in deriving nouns from verbs' (Konnerth 2014:384). The *kV-* prefix in Karbi is productive in deriving nouns from verbs, but in Liangmai this strategy is not fully productive and cannot be used to derive nouns from all verbs. However, it is possible that the nominalizer *kə-* in Liangmai is cognate with the nominalizing velar prefixes found in Tibeto-Burman languages of different branches spoken in Northeast India (Matisoff 2003; Konnerth 2012, 2014).

Another function of *kə-* in Liangmai is to appear with deverbal nouns when occurring as attributes of head nouns. This prefix functions in a way similar to the attributive derivational prefix *ə-*, which is used to derive adjectives from verbs in Meiteilon (Chelliah 1997:86; Singh 2000:114) and an attributive prefix *a-* which frequently appears on the head noun in adjective constructions in Karbi (Konnerth 2011:121). For example, *kə-* is prefixed to *wibo* 'good' to derived *kəwibo* 'good' as in *kəwibo tijnmik* 'a good day', where *tijnmik* is 'day'.

4.5. The nominalizing prefix *pə-*

When used with stative verb roots, the resulting words denote abstract entities.

- | | | | | | | |
|------|-----|-------------|------------|---|----------------|-----------|
| (20) | (a) | <i>tʰeŋ</i> | ‘be long’ | > | <i>pə-tʰeŋ</i> | ‘length’ |
| | (b) | <i>dí</i> | ‘be big’ | > | <i>pə-dí</i> | ‘breadth’ |
| | (c) | <i>suk</i> | ‘be deep’ | > | <i>pə-súk</i> | ‘depth’ |
| | (d) | <i>kú</i> | ‘be tall’ | > | <i>pə-kú</i> | ‘height’ |
| | (e) | <i>rit</i> | ‘be heavy’ | > | <i>pə-rít</i> | ‘weight’ |
-
- | | | | | | |
|------|--|----------------|---------------|--------------|-----------------|
| (21) | <i>pui-piú-gu</i> | <i>luŋsiat</i> | <i>pə-súk</i> | <i>mədat</i> | <i>si-lak-e</i> |
| | mother-father-GEN | love | NMZ-be deep | measure | know-NEG-DECL |
| | ‘The depth of parent’s love cannot be measured,’ | | | | |

When attached to any action-oriented roots, it indicates ‘the manner of’ or ‘the way of’. Note that the root morphs in nominalized forms have a rising tone, as those in (22).

- | | | | | | | |
|------|-----|--------------|---------|---|-----------------|-------------------------------|
| (22) | (a) | <i>tat</i> | ‘walk’ | > | <i>pə-tát</i> | ‘the way of his/her walking’ |
| | (b) | <i>zi</i> | ‘sleep’ | > | <i>pə-zí</i> | ‘the way of his/her sleeping’ |
| | (c) | <i>tiu</i> | ‘eat’ | > | <i>pə-tíu</i> | ‘the way of his/her eating’ |
| | (d) | <i>kap</i> | ‘cry’ | > | <i>pə-káp</i> | ‘the way of his/her crying’ |
| | (e) | <i>məlai</i> | ‘move’ | > | <i>pə-məlai</i> | ‘the way of his/her moving’ |

5 Clausal Nominalization

Nominalization is a major tool for creating various types of syntactic structures, such as attributive phrases, nominal complement clauses and relative clauses. Each nominalized clause is discussed as follows.

5.1. Attributive phrases

Liangmai employs attributive phrases in which the root of the verb is affixed by the nominalizing suffix *-bo*, as illustrated in (23) and (24).

- | | | | |
|------|--|-----------------|------------|
| (23) | <i>tsəlat</i> | <i>dinsi-bo</i> | <i>píu</i> |
| | language | speak.know-NMZ | man |
| | ‘A good orator (Lit. the man who can speak well).’ | | |
-
- | | | | |
|------|------------------------------------|---------------|--------------|
| (24) | <i>ənai-nīu</i> | <i>liu-bo</i> | <i>tsəki</i> |
| | 1DU-AGT | buy-NMZ | house |
| | ‘The house that we (both) bought.’ | | |

The phrases with these nominalized forms modify the head noun.

5.2. Nominal complement clauses

Nominal complement clauses are also derived with *-bo*. A nominal complement clause is embedded directly into a noun phrase without any further alternation of its structure.

- | | | | |
|------|---|---------------|-------------|
| (25) | <i>tsəpuaŋ</i> | <i>kep-bo</i> | <i>məri</i> |
| | elephant | shoot-NMZ | story |
| | ‘The story of the shooting of an elephant.’ | | |

Such constructions also function as noun phrase complements of verbs as illustrated in (26).

- (26) *pá-niu mənɪŋ-bo i si-e*
 3-AGT think-NMZ I know-DECL
 ‘I know what he is thinking.’

5.3. Relative clauses

Like in many other TB languages, nominalization is the main strategy to form relative clauses in Liangmai. The general nominalizer *-bo* provides the principal means of nominalizing the verb stem of relative clauses. Relative clauses functioning as noun phrase modifiers occur before the heads. The elicited (27) to (29) demonstrate the pre-head positions of relative clauses in verbless clauses and verb clauses respectively.

- (27) [*ə-tu` dap-bo*] *píu si*
 1-DAT beat-NMZ man DEF
 ‘The man who beat me’
- (28) [*zaó sak-bam-bo*] *píu si əkina-e*
 wine drink-NMZ man DEF 1.younger.sibling-DECL
 ‘The man drinking wine is my younger brother.’
- (29) [*danai i-nū liu-bo*] *əriak si ken wi-e*
 yesterdayI-AGT buy-NMZ book DEF read good-DECL
 ‘The book which I bought yesterday is good (to read).’

As seen in examples (27) to (29), relative clauses in Liangmai are distinguished by the presence of a definite article *si* after the head noun, following the nominalized clause. Head nouns in Liangmai relative constructions commonly appear to the right of the relative clause; however, they may also be clause-internal, as in (30).

- (30) *maipíu-nū əriak məzenna-leŋ pi-bo si*
 man-AGT book orphan-BEN give-NMZ DEF
 ‘The book the man gave to the orphan.’

Left-headed and headless relatives in Liangmai are also found in data elicitation. For example, the construction in (31) is ambiguous, depending on whether the relative head is taken to be the overt nominal *əriak* ‘book’, or whether the clause is interpreted as headless (‘the one’).

- (31) *əriak [liu-bo] si* or *[əriak liu-bo] Ø si*
 book buy-NMZ DEF book buy-NMZ DEF
 ‘The book that was bought’ or ‘The one who bought the book’

Relative clause formed with agentive nominalizer *-mai* is also found in the language and such clauses are also found in the pre-head position.

- (32) *məlum-mak-mai tsəpiu kep piŋ-e*
 believe-NEG-NMZ medicine shoot scare-DECL
 ‘The one that didn’t believe is scared to take injection’

5.4. Sentential Complements

Another function of Liangmai *-bo* is the nominalization of sentential complements. The nominalizer *-bo* is suffixed to the predicate of the complement clauses. The same construction is employed regardless of whether the complement clause functions as a subject (33) or as an object (34) of the matrix clause.

- (33) [tsəmi tiu kiŋ-bo] tsəpúm-leŋ wi-mak-e
 meat eat a lot-NMZ body-BEN good-NEG-DECL
 ‘[Eating too much meat] is not good for body/health.’

- (34) [pá-nūu din-bo] i məlum-e
 3SG-AGT say-NMZ I believe-DECL
 ‘I believe [in what he said].’

6 Conclusion

This paper briefly discussed different nominalizers and their functions in Liangmai. Like other Tibeto-Burman languages of the Himalayan region, Liangmai makes use of derivational (i.e., lexical) and clausal nominalization at the morphological and syntactic levels of grammar. We noticed that prefixation is limited to the process of lexical derivation whereas suffixation is used in both processes of lexical and clausal nominalization. The nominalizer *-bo* is the most common and most productive nominalizer. It is used to derive abstract nouns, action nominals, adjectivals and gerunds. In clausal nominalization, it is employed extensively in nominalized clauses, attributive phrases, complement clauses and relative clauses. The nominalizers *-bam* and *-mai* are used to derive locational and agentive nouns respectively. Agentive nominalization is also found to derive relative clauses, though not as productively as the general nominalizer *-bo*. The nominalizing prefixes *kə-* and *pə-* are found to occur only in the derivational process.

Abbreviations

1	first person
2	second person
3	third person
AGT	agentive
BEN	benefactive
CAUS	causative
DAT	dative
DECL	declarative
DEF	definite
DEM	demonstrative
DU	dual
EMP	emphatic
EXST	existential
GEN	genitive
IMP	imperative
IRR	irrealis
LOC	locative
N.COP	negative copula
NEG	negative
NMZ	nominalizer
PERF	perfective
PL	plural
POSS	possessive
PROG	progressive
PROH	prohibitive

SG singular
 SUP superlative

References

- Bickel, Balthasar. 1999. Nominalization and focus constructions in some Kiranti languages. *Topics in Nepalese Linguistics*, ed. by Yogendra P. Yadava & Warren G. Glover, 271-296. Kathmandu: Royal Nepal Academy.
- Bradley, David. 1997. Tibeto-Burman languages and classification. *Papers in Southeast Asian Linguistics 14: Tibeto-Burman languages of the Himalayas*, ed. by David Bradley, 1-72. Canberra: Australian National University.
- Burling, Robbins. 2003. The Tibeto-Burman Languages of Northeastern India. *Sino-Tibetan Languages*, ed. by G. Thurgood and R J Lapolla, 169-192. London/ New York: Routledge.
- Chelliah, S. L. 1997. *A Grammar of Meithei*. New York: Mouton de Gruyter.
- Coupe, Alexander. 2003. *A phonetic and phonological description of Ao: a Tibeto-Burman language of Nagaland, North-east India*. Canberra: Pacific Linguistics.
- Daimai, Kailadbou & Kh. Dhiren Singha. 2020. Expressing negative meaning in Liangmai. *Linguistics of the Tibeto-Burman Area* 43.1:124-146. <https://doi.org/10.1075/ltba.19008.dai>
- DeLancey, Scott. 2002. Relativization and nominalization in Bodic. *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society: Special Session on Tibeto-Burman and Southeast Asian Linguistics* (2002), 55-72.
- Delancey, Scott. 2011. Finite structures from clausal nominalization in Tibeto-Burman. *Nominalization in Asian languages: Diachronic and typological perspective*, ed. by Foong Ha Yap, Karen Grunow-Harsta and Janick Wrona, 343-359. John Benjamins Publishing Company.
- Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the World*, 24th edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.
- Genetti, Carol. 2011. Nominalization in Tibeto-Burman languages of the Himalayan area. *Nominalization in Asian languages: Diachronic and typological perspective*, ed. by Foong Ha Yap, Karen Grunow-Harsta and Janick Wrona, 163-193. John Benjamins Publishing Company
- Genetti, Carol, Coupe, Alec R., Bartee, Ellen, Hilderbrandt, Kristine and Lin, You-Jing. 2008. Syntactic aspects of nominalization in five Tibeto-Burman languages of the Himalayan area. *Linguistics of the Tibeto-Burman area* 31.2:97-143.
- Grierson, George A. 1903. *Linguistic Survey of India, TibetoBurman family, Bodo-Naga and Kachin Groups*. Vol.III, Part II. Calcutta (reprint Delhi: Low Price Publications; 2005).
- Konnerth, Linda. 2011. Functions of nominalization in Karbi. *North East Indian Linguistics*, Vol. 3., ed. by Gwendolyn Hyslop, Stephen Morey, and Mark W. Post, 120-134. New Delhi: Foundation/Cambridge University Press, India.
- Konnerth, Linda. 2012. The nominalizing velar prefix *gV- in Tibeto-Burman languages of Northeast India. *North East Indian Linguistics*, Vol.4., ed. By Gwendolyn Hyslop, Stephen Morey, and Mark W. Post, 58-80. New Delhi: Foundation/Cambridge University Press, India.
- Konnerth, Linda. 2014. *A Grammar of Karbi*. PhD Thesis. University of Oregon.
- Marrison, G. Edward. 1967. *The Classification of the Naga languages of North-East India*. PhD Thesis. Department of Phonetics and linguistics, SOAS.
- Matisoff, James A. 1972. Lahu nominalization, relativization, and genetivization. *Syntax and Semantics* ed. by John P. Kimball, 237–257. New York: Academic Press.
- Matisoff, James A. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. Berkeley, CA: University of California Press.

- Noonan, Michael. 1997. Versatile nominalizations. *Essays on Language Function and Language Type* ed. by Joan Bybee, John Haiman, and Sandra A. Thompson, 374–394. Amsterdam: John Benjamins.
- Post, Mark W. and Robbins Burling. 2017. “The Tibeto-Burman Languages of Northeast India.” *The Sino-Tibetan Languages*, ed. by Randy J. LaPolla and Graham Thurgood, Second edition, 213-42. Abingdon/ New York: Routledge.
- Singh, C. Y. 2000. *Manipuri Grammar*. New Delhi. Rajesh Publications.
- Watters, David. 2006. Nominalization in Himalayish Languages. In *Encyclopedia of Nepal's Languages*. (Draft version).
- Watters, David. 2008. Nominalization in the Kiranti and Central Himalayish languages of Nepal. *Linguistics of the Tibeto-Burman Area* 31.2:1-43.
- Yap, Foong Ha, Grunow-Harsta Karen and Wrona. 2011. Nominalisation strategies in Asian languages. *Nominalization in Asian languages: Diachronic and typological perspective*, ed. by Foong Ha Yap, Karen Grunow-Harsta and Janick Wrona, 163-193. John Benjamins Publishing Company.

SYLLABLE STRUCTURE AND MORPHEMICITY IN TONE PATTERNS ON VERBS IN KANISE KHUMI

Elissa IKEDA
Payap University
elissa.ikeda@gmail.com

Abstract

In the Khomic group within Kuki-Chin, there have been conflicting representations of whether all morphemes bear tone, or whether tonelessness is related to syllable type (Herr 2011; Hornéy 2012; Peterson 2019a). While verb roots in these languages can be monosyllabic or sesquisyllabic, they are always bound by affixes and clitics. This study offers a fine-grained analysis of tone variation on various syllable types and morphemes. I examine verbs in Kanise Khumi using archived word list data (Bryant 2020). In addition to pitch specification in various contexts, I focus on the durability of voice quality cues that are associated with Kanise tones (Ikeda 2021). Contextual variation appears to accord with acoustic studies of minor syllables that have shown that not all minor syllables are created equal (Butler 2014), and tonal specification is sensitive to whether the syllable is part of a compound, part of the lexical root, or a functional morpheme (West 2014).

Keywords: tones, verbs, morphophonemics, phonology, phonetics, language description, acoustics

ISO 639-3 codes: cek

1 Introduction

This paper focuses on surface tone patterns of citation verbs in Kanise Khumi. My underlying assumption is that tone in this language cannot be adequately analyzed without reference to the morphological and prosodic structure of words. However, defining these structures is itself no simple task (Dixon & Aikhenvald 2002). In the beginning, we can only build initial hypotheses about these structures and their tonal patterns based on descriptions of closely related varieties within a language cluster or subgroup. We then test those hypotheses with data from a specific language.

Kanise Khumi is spoken mainly in villages near Sami town, Paletwa Township, Chin State, Myanmar. The number of speakers is unknown, but an internal census based on ethnicity has counted 5,776 Kanise Khumi (Bryant, p.c.). Kanise Khumi has been previously known by the names of some of the larger clans, such as Nideun (Eberhard, Simons & Fennig 2020), Tahaensae (So-Hartmann 1988), or Uiphaw. It is associated with the ISO language code [cek], a label based on lexical similarity, not shared sense of identity. Peterson's (2017; 2019a; 2012) proposed classification of Kuki-Chin is based on a more detailed analysis of the Khumi cluster of languages. He situates the Khumi cluster in the Southwestern group within a peripheral branch of Southwest Tibeto-Burman (=Kuki-Chin). Other languages in the cluster include those spoken in geographically adjacent areas such as Lemi Chin and Mro Khimi as well as those spoken further away in the Chittagong Hill Tracts of Bangladesh such as Bangladesh Khumi and Rengmitca.

An analysis of tone in a Southwest Tibeto-Burman language must take into account several well-established premises. One is that tonal variation can mark morphosyntactic relations, and even be the sole morphological marker (Hyman 2007; Peterson 2019a; Henderson 1967; Coupe 2007). In addition, the syllable is often the tone bearing unit (Lotven et al. 2020; Hyman & VanBik 2002; VanBik 2006; Peterson 2019a). However, not all syllables in these languages must bear tone, especially affixes and clitics (Hyman 2006; Hornéy 2012; Hyman & VanBik 2004). Aside from morphosyntax, tonality can also be dependent on syllable type (e.g., toneless minor syllables vs. tonal major syllables (Herr 2011; Hornéy 2012; Hyman & VanBik 2002)). Furthermore, some tones may be constrained by context, for example, contour tones restricted to word-final position or prohibitions against adjacent tones (Hyman

2006). At the same time, contextual effects can be limited to specific tones (So-Hartmann 1989; Sarmah, Dihingia & Lalminghlui 2015). Finally, tone is not synonymous with pitch. In many languages in Southeast Asia, tone is signaled through a bundle of phonetic features, including pitch height, pitch shape, duration, and voice quality (Brunelle & Kirby 2016; Henderson 1965).

Descriptions of Khomic languages differ with respect to tone on verbal affixes and clitics. Herr's (2011) aim is to interpret the phonological status of minor syllables in Lemi Chin, not analyze tonal variation. She treats all minor syllables as phonologically toneless and all major syllables as tone bearing. In contrast, Hornéy (2012) does not base tonelessness on syllable structure, but rather on morphological structure. In her phonological analysis of Mro Khimi, she analyzes tone melodies of verbs. She describes verb “prefixes” and “suffixes” as underlyingly toneless, regardless of whether they are minor or major syllables. Using a lexical phonology approach (Kiparsky 1982), she posits that the surface tone on the “clause final particle” is derived through melody association and spreading from the verb root. The surface tone of “prefixes” is derived through processes that include spreading, polar tone assignment, and lowering. Both of these studies are based primarily on word list data. In comparison, Peterson's analyses of Bangladesh Khumi are constructed from a large corpus that includes long texts and elicited data. He has paid extensive attention to tonal variation in diverse morphosyntactic environments. In Peterson (2019a), syllables are specified with an underlying tone; however, the “half-syllables” found in sesquisyllabic forms are treated as toneless. When he makes a claim about the tone of a major syllable that is an affix or clitic (2013; 2019a), it appears to be based on the behavior of that specific marker, rather than a general property of all preverbal elements or all postverbal elements.

Thus, it makes sense to begin analyzing tone patterns on Kanise Khumi verbs with the following questions in mind. Do the tones on affixes/clitics appear to be consistent or variable? Do the tones on Kanise Khumi affixes/clitics appear to be sensitive to the tone of the verb root? Do any Kanise Khumi minor syllables appear to be specified for tone? Following Peterson, this paper takes a morpheme-specific approach. I focus on the morphemes that appear most frequently on citation verbs in Kanise Khumi word list recordings. First, we look at the Kanise postverbal marker *-ta* which behaves more like Lemi *-te* than Mro Khimi *-de* in terms of tone patterns. The rest of the paper is devoted to prefixes. With the valency-affecting prefixes *p-*, *t-*, and *a(ŋ)-*, syllable structure appears to differentiate tonal patterns as in Lemi. A similar argument could be made for the adjectivizers *k-* and *ka(ŋ)-*. In contrast, the tone patterns with the highly frequent 3rd-person singular participant marker, *ʔɜ-*, resist a simple distinction between toneless minor syllables and tone-bearing major syllables. An argument based on polar tone assignment of prefixes also does not seem appropriate. As such, the tone pattern data raises questions about how to define or categorize minor syllables in Kanise Khumi and about morpheme productivity.

2 The Data

The data for this paper comes from an elicited word list that includes 2,076 items (Bryant 2020). The .wav files and Excel database were downloaded from Zenodo. The word list is based on the EFEO-CNRS-SOAS Word List for Linguistic Fieldwork in Southeast Asia (Pain et al. 2019). Most of the recordings in the list are from one male speaker in his 50s. About 193 of the items were recorded with a younger male speaker. Some of these items are duplicates of those recorded by the main speaker. Both men are multilingual. The word list was elicited using Burmese glosses. During elicitation, Bryant transcribed the words phonetically and the speaker wrote the words in the current orthography. Three tokens were recorded for each item. The speaker was also asked to parse the response into what he perceived as words wherever possible. When he did so, he pronounced the entire item three times, then three tokens of each separate “word.” The main speaker generally used very slow, careful speech.

2.1 Syllable structure

Most Kanise Khumi syllables have CV structure. Table 1 shows attested syllable structures.

Table 1: Kanise Khumi syllable types

CV	ʔɛ̃l	‘house’	kəɫ.dɔɫ	‘to boil’ (intransitive)
CVV	suiɫ	‘gold’	təɫ.ʔauɫ	‘to yell, speak loudly’
CCV	amɫ.praɫ	‘uncle’	ʔəmɫ.pɾɛɫ	‘to divorce’
CCVV	təɫ.praũɫ	‘cemetery, graveyard’	tsaɫ.p ^h raɪɫ tɔɫ	‘to read’
CVN.CV	nɪɫ.nɪɫ	‘they two’	ʔəmɫ.baɪɫ	‘to lie prone, on one’s stomach’
N.CV	nɪ.tsɔ̃ɫ	‘skewer’	m.bəɫ jaɫ	‘to give birth’

Simple onsets are shown in Table 2. Quite a few allophones occur with fricatives. The affricate /tʃ/ may freely be pronounced [s]. The palatal fricative is variably realized as [j ~j~ɟ]. The velar fricative /ɣ/ also varies in manner of articulation, [ɣ ~ɰ~g]. Permissible complex onsets included /kl, kɪ, k^hɪ, p^hɪ, pɪ, mj/. The main speaker exhibits free variation between [kl] and [tl], but the younger speaker produced only [kl]. Coda nasals are permitted in non-final syllables. In word-final position, oral closure is rarely heard, and most often the corresponding vowel is nasalized. Glottalization in codas is considered a feature of tone. Syllabic nasals occur in non-final position only. Minor syllable onsets include /m, p, n, t, d, ts, s, k, ʔ/ with /p, t, k, ʔ/ occurring most frequently (Baleno 2020).

Table 2: Kanise Khumi simple onsets

p p ^h b	t t ^h d		k k ^h	ʔ
m m̃	n ñ		ŋ ŋ̃	
(f) v	s ts	j	ɣ	h
(w)	l l̃ l̃̃			
	tʃ			

Kanise Khumi has nine simple vowels as shown in Table 3. All simple vowels occur with oral and nasal phonemes. Complex vowels include /ai, au, ui, oe/. There are no examples of nasalized /ui/ or /oe/. The open-mid vowels are usually realized as [e^ɛ], [ɜ^ɔ], and [ɔ^o] with a vowel height transition during the vowel. As with Lemi Chin (Herr 2011) and VanBik’s Khumi (2006) vowel length does not appear to be contrastive. Bryant often transcribed the mid-central vowel in minor syllables as [ə]; however, Tan (2021) finds no significant difference between [ə] and [ɜ].

Table 3: Kanise Khumi simple vowels

i ĩ	ɪ ɪ̃	u ũ
e ɛ̃		o ɔ̃
ɛ ɛ̃	ɜ ɜ̃	ɔ ɔ̃
	a ɶ̃	

2.2 Phonetic tones

Tone appears to bear a low functional load in terms of lexical distinctions. Kanise Khumi speakers report three phonetic tones (Bryant, p.c.). Analyzing monosyllabic nouns, Ikeda (2021) described these as (1) high level modal, (2) falling low breathy, and (3) short mid glottalized.

Table 4 provides visualizations of the tone contours of the three phonetic tones in similar environments. In Table 4, the verb root is preceded by a complement noun (object or instrument). In isolation each complement noun bears Tone 1, the high-level modal tone. The vertical red lines bisect the verb roots. In Table 5, the root is prefixed with the adjectivizer *k-*. The prefix is a mid or neutral pitch See section 3.3.

Table 4. Kanise Khumi phonetic tones on verb roots preceded by a noun

Tone 1 (H)	Tone 2 (F)	Tone 3 (M)
[tuiḷ nē:ḷ] ‘drink water’	[pɜ:ḷ r̥iḷ] ‘to draw in a casting net’	[ʔē:ḷ s̥aḷ] ‘to build a house’

Table 5. Kanise Khumi phonetic tones on verb roots preceded by adjectivizer *k-*

Tone 1 (H)	Tone 2 (F)	Tone 3 (M)
[kəḷ prē:ḷ] ‘dense, thick (of forests)’	[kəḷ j̥ɔḷ] ‘ticklish, sensitive’	[kəḷ l̥ēḷ] ‘big’

The six tone combinations presented in Tables 4 and 5 are the most frequent in sesquisyllabic and disyllabic verbs. Table 6 shows all the combinations that are found in citation verbs in the word list. There appears to be no rule prohibiting sequences of like tones within a compound. Second, although Tone 2 occurs less frequently, it can precede or follow Tone 1 within a compound. In non-final position, Tone 2 does not fall as steeply or as low as the Tone 2 observed in monosyllables or final syllables. However, the initial pitch height and the breathy voice quality are similar. Finally, the three empty cells in Table 6 show sequences that are not observed in citation verbs. These sequences are attested in trisyllables where *-ta* follows the verb stem and in noun compounds.

Table 6: Adjacent tones in Kanise Khumi

[HH] ¹ [tui ¹ nē ¹] 'to drink water'	[HF] [pɜ: ¹ rɪ ¹] 'to draw in a casting net'	[HM] [ʔɛ: ¹ sɑ ¹] 'to build a house'
[MH] [rɑ ¹ ne ¹] 'to drink alcohol'	[MF] 'to be old' (age)	[MM] [maɪ ¹ tsɔ ¹] 'to be old' (age)
[FH] [lɪ ¹ p ^h i: ¹] 'bald'	[FF]	[FM]

There was also a sequence with Tone 2 followed by a high glottalized final syllable in [tɛ̃² ʔɜ²] 'false'. This presents the possibility of a high glottalized tone. However, the syllable also begins with a glottal stop, which seems to condition glottalization of the rhyme in many other instances. At this point, it is unclear whether this represents a fourth phonetic tone.

2.3 Structure of citation verbs

In the Kanise Khumi word list, citation verbs never appear as isolated monosyllables. In fact, isolated monosyllables are quite rare in the word list, with only fifty-six nouns, two pronouns, and two numbers. The simplest forms of verbs given in the word list are isolation forms that are composed of two syllables, either sesquisyllabic or disyllabic. There are 283 of these forms. This number includes a few duplicates where (a) related semantic ideas yielded the same form or (b) the same word was elicited from both speakers. All duplicates were included in this paper because they provide evidence relevant to tonal variation. Three types of structures are observed: arguably dimorphemic sesquisyllables, arguably dimorphemic disyllables, and arguably monomorphemic disyllables.

(1) Arguably dimorphemic sesquisyllables

a. <i>t</i> -root:	[tə ¹ ʔau ¹]	'to yell, speak loudly'	(n=21)
b. <i>p</i> -root:	[pə ¹ ŋɪ ¹]	'to forget'	(n=15)
c. <i>k</i> -root:	[kə ¹ dɔ ¹]	'to boil' (intransitive)	(n=9)
d. <i>s</i> -root:	[sə ¹ gui ¹]	'to follow'	(n=5)

(2) Arguably dimorphemic disyllables

a. ʔɜ-root	[ʔɜ ¹ ŋo: ¹]	'to see'	(n=109)
b. <i>a</i> -root	[ʔa ¹ bi ¹]	'to hide oneself'	(n=49)
c. <i>aŋ</i> -root	[ʔã ¹ prɛ ¹]	'to run away, flee'	(n=33)
d. <i>ka(ŋ)</i> -root	[ka ¹ kau ¹]	'empty'	(n=15)
e. Noun – Verb	[tui ¹ nē: ¹]	'drink water'	(n=10)
f. Verbal Compound	[ku ¹ ŋau ¹]	'to tame'	(n=13)

(3) Arguably monomorphemic disyllables

a. [dɛ ¹ tlã ¹]	'bland (flavor)'	(n = 4)
b. [kɔ ¹ le ¹]	'to tickle'	
c. [tã ¹ pɔ ¹]	'thin'	

It is important to note that some of these forms can be combined as trisyllables. In (4), ʔɜ- and *a*-precede all four of the arguably dimorphemic sesquisyllable types.

¹ HH = high-high, MH = mid-high, FH = falling-high, HF = high-falling, MF = mid-falling, FF = falling-falling, HM = high-mid, MM = mid-mid, FM = falling-mid

(4) Arguably trimorphemic trisyllables

a. ʔ ₃ - <i>t</i> -root:	[ʔ ₃ l təɫ hɛɫ]	‘to stir up (water)’
b. ʔ ₃ - <i>p</i> -root:	[ʔ ₃ l pəɫ ɲɛɫ]	‘to smell’
c. ʔ ₃ - <i>k</i> -root:	[ʔ ₃ l kəɫ tsɿɫ]	‘to polish’
d. ʔ ₃ - <i>s</i> -root:	[ʔ ₃ l sə ʔɛɫ]	‘to nip, to pinch with one’s fingers’
e. <i>a</i> - <i>t</i> -root:	[ʔaɫ təɫ mɔɫ]	‘to dry yourself (with a towel)’
f. <i>a</i> - <i>p</i> -root:	[ʔaɫ pəɫ tɛɫ]	‘to tease’ (parsed as <i>a p</i> -root)
g. <i>a</i> - <i>k</i> -root:	[ʔaɫ kəɫ tsuɫ]	‘to gather, assemble’
h. <i>a</i> - <i>s</i> -root:	[ʔaɫ səɫ lɛɫ]	‘to dive’

There is a postverbal morpheme, *-ta* that occurs frequently with citation verbs (n=90). It appears to be parallel to, *-dɛ*, the obligatory suffix described by Hornéy for Mro Khimi and *-te³*, the nearly obligatory suffix that appears in Herr’s transcriptions of Lemi Chin verbs. Unlike Mro Khimi, *-ta* does not occur in dimorphemic disyllables in the Kanise Khumi word list. Rather, *-ta* is appended on sesquisyllabic, disyllabic, and trisyllabic verbs as in (5). It is unclear why *-ta* is less obligatory in the Kanise Khumi word list data than the Lemi and Mro Khimi word list data.

(5) *-ta* postverbal marker

a. ʔ-root- <i>ta</i>	[ʔəɫ dɛɫ t̚aʔɫ]	‘to mash (a tuber)’ (0353)
b. <i>t</i> -root- <i>ta</i>	[təɫ p̚ɿɫ t̚aʔɫ]	‘to squeeze’ (parsed as <i>t</i> -root <i>ta</i>) (0354)
c. <i>p</i> -root- <i>ta</i>	[pəɫ jɔɫ t̚aɫ]	‘to suck’
d. <i>k</i> -root- <i>ta</i>	[kəɫ sɛɛɫ t̚aɫ]	‘tremble’ (parsed as <i>k</i> -root <i>ta</i>) (2209)
e. ʔ ₃ -root- <i>ta</i>	[ʔ ₃ l kʒɿɫ t̚aʔɫ]	‘to grow’
f. <i>a</i> -root- <i>ta</i>	[ʔaɫ d̚ɔɫ t̚aɫ]	‘to think about’ (2072a)
g. Noun—Verb- <i>ta</i>	[tsɔɫ hɛɫ t̚aʔɫ]	‘to sow rice’
h. NOUN- <i>a</i> -root- <i>ta</i>	[tsɔɫ ʔaɫ t̚ɿɫ t̚aʔɫ]	‘to thresh rice by stamping’
(parsed as NOUN <i>a</i> -root <i>ta</i>) (0290)		
i. ʔ ₃ - <i>p</i> -root- <i>ta</i>	[ʔ ₃ l pəɫ t̚houɫ t̚aʔɫ]	‘to transplant rice seedlings’
(parsed as ʔ ₃ - <i>p</i> -root <i>ta</i>) (0280)		

3 Tone analysis

The transcriptions in this article are phonetic. Most are from Bryant’s Excel spreadsheet database. In several words, I have revised the transcriptions based on auditory perception and acoustic evidence. I charted the tone patterns based on Bryant’s transcriptions and checked them against the sound files using auditory perception and Praat visualization.

3.1 Postverbal *-ta*

This marker appears to be parallel to postverbal markers found in Lemi Chin, Mro Khimi, and possibly Bangladesh Khumi. In Herr’s Lemi Chin word list, [-te³] is present on almost every verb. In her phonetic transcriptions, the form is always specified with the mid tone and often accompanied by a glottal stop. A similar morpheme /dɛ/ follows every verb stem in Hornéy’s Mro-Khimi word list data. While Herr focuses on describing minor syllables, Hornéy attempts a thorough phonological analysis of the tone melodies found in verbs. She considers /dɛ/ to be an obligatory clause-final particle. She proposes that it is underlyingly toneless, gaining its tone from the root through a process of melody association and rightward spreading. According to Peterson (2019a; 2018), Bangladesh Khumi has a post-verbal hearsay evidential clitic represented as =te⁵. In his examples and sample text, =te⁵ surfaces with low and high falling tones but not with the checked tones. It is unclear whether the Bangladesh evidential clitic is analogous with the postverbal morphemes in the other three languages. Yet it appears that there is a contrast between Lemi on the one hand, in which the tone is fixed, and Mro Khimi and Bangladesh Khumi on the other hand, in which the tone varies.

Kanise Khumi seems to follow the same pattern as Lemi. The phonetic realization is consistent with Tone 3: mid pitch with either a level or slightly falling shape as well as glottalization. The tone of

-ta is not sensitive to the tone of the verb stem. An example of *-ta* following each tone is provided in (6).

- | | | | |
|-----|---------------------|-----------------------|------------------------------|
| (6) | a. Following Tone 1 | [tui̯ ɛ:̯ t̩a̯ʔ] | ‘to water’ |
| | b. Following Tone 2 | [tsɔ̯ ʔa̯ t̩i̯ t̩a̯ʔ] | ‘to thresh rice by stamping’ |
| | c. Following Tone 3 | [tsɔ̯ ɛ̯ t̩a̯ʔ] | ‘to sow rice’ |

Any variation observed in the tone of *-ta* is consistent with general observations of Tone 3 in monosyllabic nouns (Ikeda 2021). In monosyllables, Tone 3 varies between level and slightly falling. This variation is also shown in (7) where most often the mid pitch is level (7a), but a slightly falling pitch is also frequent (7b). In seven words (7c), the pitch shape varies freely between the two shapes, indicating that the difference is not significant. Furthermore, right edge contour effects have been widely observed in languages related to Kanise (Evans 2009; Ikeda 2021; Herr 2011; Hornéy 2012). Generally, Speaker 2 tends to show the right-most edge falling contour as in (7d). In such cases, voice quality becomes the most reliable cue.

- | | | | |
|-----|----|-------------------------|--------------------------------|
| (7) | a. | -t̩a̯ | (n=50) |
| | b. | -t̩a̯ | (n=26) |
| | c. | -t̩a̯ varies with -t̩a̯ | (n =7) |
| | d. | -t̩a̯ | (n=7: Speaker 2; creaky voice) |

Moreover, it is precisely at the right edge where voice quality is most noticeable. The verb ‘to grow’ has Tone 3 in both the verb stem and *-ta*. When the speaker pronounced the entire item [ʔɛ̯ k̩ʔ t̩a̯ʔ] ‘to grow’, the final syllable was more glottalized than the penultimate syllable. When he isolated the verb stem [ʔɛ̯ k̩ʔ], the final syllable [k̩ʔ] was strongly glottalized. Then [t̩a̯ʔ] was repeated three times in isolation with strong glottalization.

In a few other cases, the speaker parsed *-ta* separately from the verb stem. Figure 1 shows the pitch curve of [tui̯ ɛ:̯ t̩a̯ʔ] ‘to water’ in blue dots. The vertical red line cuts through the center of *-ta*. The pitch of *-ta* is mid, slightly falling. Glottalization is evident in the breakup of F0 near the end of the syllable. Figure 2 presents three repetitions of *-ta* when the speaker parsed ‘to water’ into three parts: [tui̯] ‘water’, the verb stem [ɛ:̯], and [t̩a̯ʔ]. This suggests that the speaker treats *-ta* as a functional morpheme.

Figure 1: [tui̯ lɛ:̯ tɕʔʌ] ‘to water’

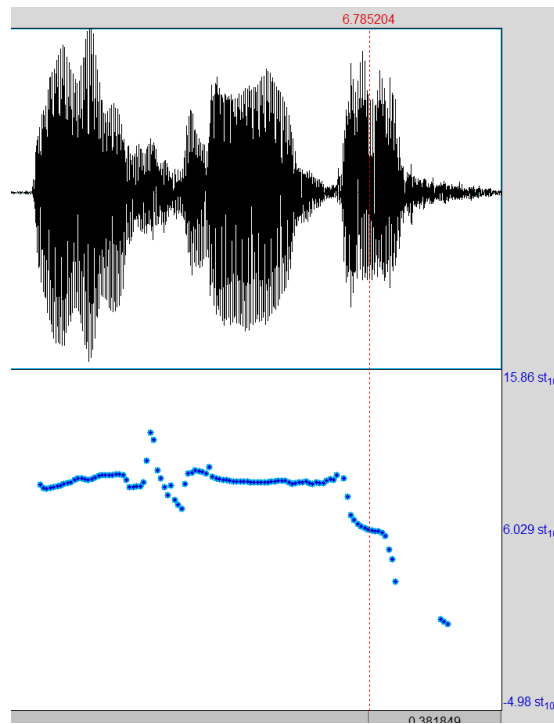
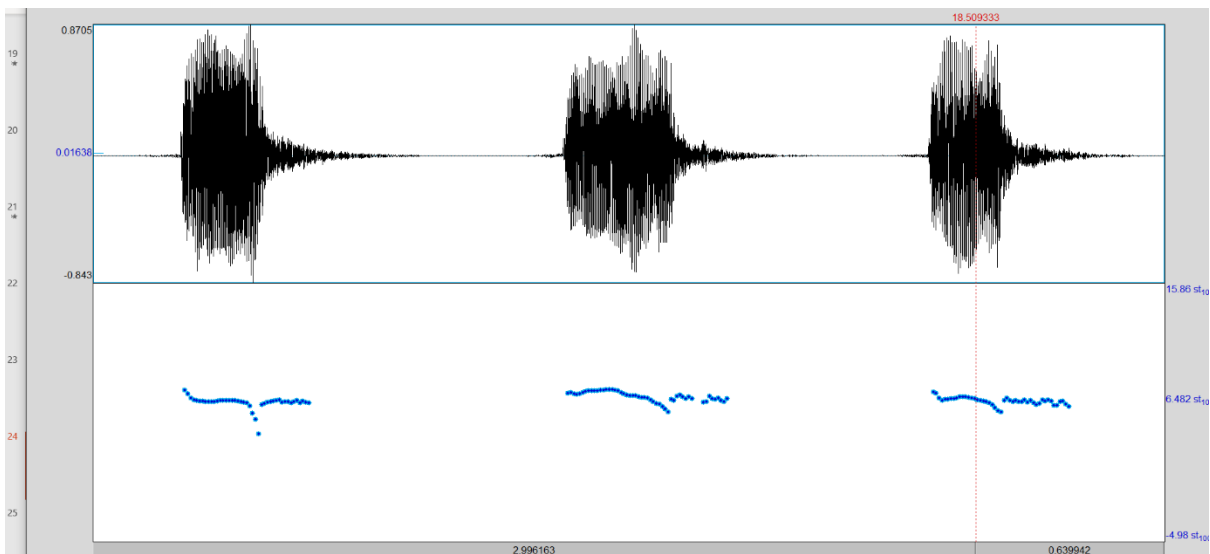


Figure 2: -ta repeated three times in isolation



3.2 Valency affecting prefixes

3.2.1 Valency increasing p- and t-

What about the tones of preverbal affixes? Valency increasing prefixes have been observed across a variety of Chin languages and quite consistently within the Khumi cluster. The most common forms are bilabials *m-* and *p-*. These are attested in Lamkang, Mara, Daai, K’Cho, Mro Khimi, Lemi, Bangladesh Khumi, and Rengmitca (So-Hartmann 2013; Peterson 2019b; Peterson 2013).

In Lemi, causative and transitive verbs can be derived by prefixing free verb roots that denote intransitive states and activities. According to So-Hartmann (2013), *m-* and *b-* are in free variation, and *t-* is another form with lower productivity. She also gives examples revealing that *m-* and *t-* have been

lexicalized and frozen with verb roots and noun roots. These are treated as toneless minor syllables by Herr (2011) as in /m.nã³.te³/ ‘to kiss’. Herr treats all Lemi minor syllables as phonologically toneless and states that they surface with a default mid-tone. So-Hartmann (2013) presents similar causative-transitive prefixes for Mro Khimi, with both *m-* and *t-*. Hornéy (2012) analyzes all Mro Khimi verb prefixes as toneless minor syllables that undergo vowel insertion and polar tone assignment as in /m.sø^H-dɛ/ → [màsõdɛ]. Rengmitca also has a causative *m-* prefix (Peterson 2019b).

Bangladesh Khumi has a semi-productive causative *p-* prefix and a less frequent, less transparent causative *t-* prefix which may be fossilized (Peterson 2013). Peterson (2013) suggests that *t-* may be an allomorph of *p-* that primarily occurs before bilabial and *h* initials, and So-Hartmann (2013) comments that this may be the case in Lemi and Mro as well. Peterson represents these valency increasing prefixes as toneless minor syllables in Bangladesh Khumi. Similar to Lemi and Mro Khimi, they are “more transitive” than their free verb root counterparts. Peterson gives examples such as *tlâng⁴/p⁴tlâng⁴* ‘melt’ and *më⁴/t⁴më⁴* ‘twist’. Peterson also provides evidence of pairs with the highly frequent middle marker such as *acã²/p²cã²* ‘bounce’ and *apew¹/t¹pew¹* ‘explode’. (See Section 3.2.2.)

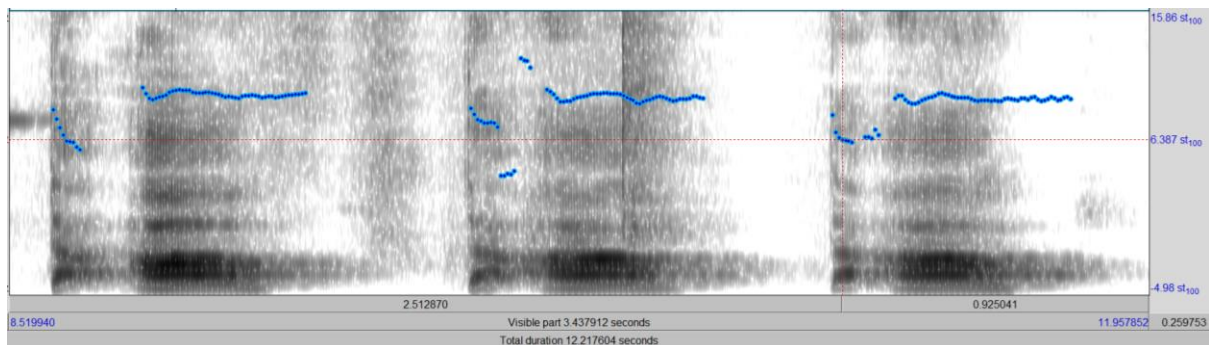
These same valency increasing prefixes are attested in Kanise Khumi. A similar pair was elicited for ‘to yell, speak loudly’, *a*-root [ʔaɭ ʔauɭ] and *t*-root [təɭʔauɭ]. In Kanise Khumi, *p-* never precedes a bilabial initial or /h/, but *t-* does. The pitch patterns for *p-* and *t-* are given in Table 7. Three surface pitch patterns are observed. The pitch of the prefix is predictably mid, regardless of the tone of the following syllable. There is no indication of polar tone. Unlike *-ta*, the valency increasing prefixes are never separated from the verb root.

Table 7: Pitch patterns with prefixes *p-* and *t-*

Pitch Pattern	Examples	# of items
HM	[pəɭ ʔəɭ] ‘to bake clay’ [təɭ me:ɭ] ‘to forget’	<i>p-</i> = 6 <i>t-</i> = 3
MF	[pəɭ k ^h ɛɭ] ‘to fetter’ [təɭ ɣɛɭ] ‘to war, to wage war’	<i>p-</i> = 1 <i>t-</i> = 3
MM	[pəɭ ŋɛɭ] ‘to kiss’ [təɭ p ^h ɾɛɭ] ‘to erase, delete’	<i>p-</i> = 8 <i>t-</i> = 15

The prefix syllable tends to be of noticeably short duration (roughly 50-90 ms), so pitch is often significantly perturbed by surrounding consonants. The differences in pitch height and pitch shape across words are similar to the differences across tokens in the same word (See Figure 3). Therefore, it is difficult to say that these differences merit analysis as different tone patterns.

Figure 3: Three tokens of [təɭʔauɭ] ‘to yell, speak loudly’



Voice quality cues on the major syllable (verb root) were consistent with the bundle of features observed in monosyllabic nouns. The high-level pitch is correlated with modal voice (Tone 1). The pitch fall from mid to low is correlated with breathy voice (Tone 2). The mid tone is shorter and glottalized and often creaky with either level or slightly falling pitch shape (Tone 3). Speaker 2 continued to exhibit a right edge boundary effect on Tone 3 with voice quality the most reliable cue. Yet with the prefix

syllable, it is difficult to perceive voice quality cues. Perhaps this is due to the short duration. When the prefix precedes an initial glottal stop, it sounds quite glottalized.

Interestingly, the speaker does not associate the vowel in the presyllable with the same orthographic representation as a main syllable mid-central vowel. He always writes the presyllable vowel as <a> whereas in final syllables, a mid-central vowel is written as <ay>. Perhaps these prefixes are fossilized and now form “reduced” syllables. Although the presyllables are short and predictably have mid pitch, we cannot conclude that they are lexically specified with Tone 3. Like Lemi Chin, minor syllables may be toneless and simply realized with a “neutral” pitch.

3.2.2 Valency-decreasing a(ŋ)-

The pitch patterns associated with the full syllable prefix a(ŋ)- are quite different in that the prefix has predictably high pitch, much like Tone 1.

Peterson (2013) describes Bangladesh Khumi as having a highly frequent, toneless middle marker *a-* with reciprocal, passive, and anticausative senses. There are verb pairs where the verb prefixed with *a-* is less transitive than the verb root as in *abang³/bang³* ‘hang’. In addition, verb pairs with *a-* vs. *p-/t-* can be compared in terms of valency as in *acã²/p’cã²* ‘bounce’ and *apew¹/t’pew¹* ‘explode’. Finally, Bangladesh Khumi has optional nasals associated with *a-* as in *a(m)lë¹/p’lë¹* ‘roll’, *a(ng)hã²/p’ha²* ‘slide’ and *a(ng)hüng²/t’hüng³* ‘shake’. While Peterson treats this marker as a full syllable, he describes it as atonal.

Equivalent valency-decreasing prefixes have been observed in Lemi and Mro Khimi (So-Hartmann 2013). Lemi intransitive, reflexive, and reciprocal verbs can be derived from transitive verbs with the prefix *ae-*. The Lemi transitive verbs can simply be verb roots as in *aehi* ‘y is spread’/*hi* ‘x spreads y’. However, the transitive verbs may also have the *m-/t-* prefixes to which *ae-* is added as in *aemäcaw* ‘x feeds himself’/*mäcaw* ‘y feeds x’ and *aetämüing* ‘x+y suppress each other’/*tämüing* ‘x suppresses y’. With some verbs expressing body posture, the *ae-* intransitives are lexicalized and there is no more transitive counterpart.

So-Hartmann (2013) gives examples of a valency-decreasing *ka-* in Mro Khimi, such as *kamshie* ‘x washes himself’/*mshie* ‘y washes x’ and *kabraan* ‘x+y fight with each other’/*braan* ‘x quarrels with y’. However, she states that the *ka-* for deriving reflexives is not really productive and that reciprocal verb derivations are rare.

Again, we find a dichotomy between Lemi and Mro Khimi, where Lemi full syllable prefixes seem to carry tone and Mro Khimi prefixes display polar tone assignment. Hornéy (2012) seems to have included *ka-* in her analysis of polar tone prefixes, for she has [kat^høde] ‘boast’ and [kas^hide] ‘wear’ as examples of pitch patterns with low and high pitch respectively. Table 8 provides a few comparisons, illustrating that Herr analyzed Lemi’s *ae-* as a full syllable, surfacing with a mid glottalized tone.

Table 8: Comparative data for Lemi (Herr 2011), Bangladesh Khumi (Peterson 2013) and Kanise Khumi

	Lemi	Bangladesh Khumi	Kanise Khumi	
a.	ʔɛ ³ .lɔ ³ .te ³	a(m)lë ¹	ʔã ¹ l ɬɛ ¹	‘roll’
b.	ʔɛ ³ .pɔ̃ ⁵ .te ³	apung ²		‘marry’
c.	ʔɛ ³ .pre ³ .te ³	a(ng)pre ²	ʔãm ¹ l pɾɛ ¹	‘divorce’

Again, Kanise Khumi aligns with Lemi Chin. In 79 out of 82 items, the prefix is pronounced with high pitch. Voice quality sounds modal, except when preceding a verb root that begins with a glottal plosive. There are three exceptions where *a-* appears with a mid pitch. One of those cases is the verb ‘to sing’ [ʔa¹ ʔoe¹] #2150; however, when the same verb appears at another point in the word list as ‘to sing (of birds)’ [ʔa¹ ʔoe¹] #784, the prefix had high pitch. The three exceptions may indicate that the pitch of the prefix is non-contrastive, even if tone is lexically specified.

Table 9: Pitch patterns with prefixes *a-* and *aŋ-*

Pitch Pattern	Examples	# of items
HH	[ʔa] bi] ‘to hide (oneself)’	<i>a-</i> = 12
	[ʔã] prẽ] ‘to run away, to flee’	<i>aŋ-</i> = 14
HF	[ʔa] gẽ] ‘to wait’	<i>a-</i> = 10
	[ʔã] pã] ‘to cross over, traverse’	<i>aŋ-</i> = 6
HM	[ʔa] bĩ] ‘to wear in your ears’	<i>a-</i> = 25
	[ʔã] prẽ] ‘to divorce’	<i>aŋ-</i> = 12
MH	[ʔa] ʔoe] ‘to sing’	<i>a-</i> = 2
	[ʔa] mã:] ‘to dream’	<i>aŋ-</i> = 1
	[ʔã] rẽ] ‘rich’	

The speaker seems to associate the prefix with Tone 1. Generally, the speaker parses ‘words’ so that the prefix is bound to the verb root. However, it is possible for him to isolate the valency-decreasing prefix. In two items, the speaker isolated the prefix from the verb root. For example, #712 ‘to kick (of a horse)’ [ʔa] pi] was repeated twice and then [ʔa:] and [pi] were each repeated three times. When the prefix was pronounced in isolation, the duration was long, the pitch was high, and the voice quality was modal. These are the features of Tone 1. Similarly, the form of the prefix with the nasal [ʔã:] was repeated three times in isolation with the verb [ʔã] ʔã] ‘to carry together (two people carry objects tied to a pole that rests their shoulders)’ #2431. Again, the phonetic cues were like Tone 1, long duration, high pitch, and modal voice quality. Further evidence for lexical specification with Tone 1 comes from trisyllabic forms. Like Lemi, Kanise Khumi has forms where the valency-decreasing prefix is added to transitive verbs that begin with *p-* and *t-*. The speaker parsed [ʔa] pã] tã] ‘to tease’ #2293 as [ʔa] and [pã] tã], with [ʔa] carrying Tone 1.

As Peterson (2013) noted, it is difficult to explain the optional nasal. It is unclear why it appears on some verbs and not others. However, the phonotactic restrictions on coda consonants are consistent. In final position, the nasal plosive is deleted and pronounced as a nasalized vowel. In non-final position, the place of articulation of the nasal assimilates to the place of articulation of the following consonant. Before an initial glottal plosive, the nasal is velar, a pattern that has also been observed for presyllables in Katic languages of the Austroasiatic phylum (Gehrmann 2017). Bangladesh Khumi, on the other hand, appears to have both *am-* and *ang-* forms that are not sensitive to the place of articulation of the following syllable.

3.3 Adjectivizers *k-* and *ka(ŋ)-*

The adjectivizers *k-* and *ka(ŋ)-* provide another comparison of tone patterns in prefixes. Synchronically, Kanise Khumi appears to have multiple forms with similar functions. Several adjectives are formed with either *k-* or *kã(ŋ)-*. They can modify a noun as in (8). The nasal is often, but not always, homorganic with the onset that follows. Sometimes *ka-* is simply nasalized and there are even a couple of cases where *ka-* is not nasalized.

- (8) a. [tui] kã] nu] ‘turbid water’
 b. [tui] ka] tsai] pi:] ‘clear water’

In unpublished notes, entitled “Lemi Functional Prefixes”, So-Hartmann lists three prefixes as having adjectivizer functions: *kã-*, *kae-*, *kang-*. The first two also derive nominals in Lemi. In addition, Mro has a prefix *ka-* that derives agentive nominals from activity verbs (So-Hartmann unpublished; So-Hartmann 2008).

Whether the adjectivizer is a minor syllable or a full syllable, the pitch is almost always mid as shown in Tables 10 and 11. There are only two exceptions. The tone of the prefix does not appear to be sensitive to the tone of the following syllable.

Table 10: Pitch patterns with *k-*

Pitch Pattern	Examples	# of items
MH	[kəɫ tʰɔ̃ɿ] ‘small’	4
MF	[kəɫ jɔ̃ɿ] ‘ticklish, sensitive’	2
MM	[kəɫ lɛ̃ɿ] ‘big’	2
HF	[kəɫ ʔãɿ] ‘shine (the sun shines)’	1

In many ways, *k-* behaves phonetically like *p-* and *t-*. The prefix has very short duration and the pitch is almost invariably mid. Glottalization is noticeable when the following onset is a glottal plosive. There is no way to determine whether the prefix is toneless with neutral pitch or whether the prefix carries Tone 3 with the voice quality feature neutralized. On the other hand, the full syllable *ka(ŋ)-* appears to be specified as Tone 3. In four cases, the speaker isolated the prefix and pronounced it very clearly with the features of Tone 3: short duration, mid pitch, glottalization.

Table 10: Pitch patterns with *kã(ŋ)-*

Pitch Pattern	Examples	# of items
MH	[kãɿ sɛ̃ɿ] ‘yellow’	5
MM	[kãɿ kãɿɿ] ‘empty’	4
FH	[kãɿ hɛ̃ɿ] ‘raw (not cooked)’	1
MF	[kã(ŋ)ɿ lɔ̃ɿ] ‘white’	5

3.4. 3rd-person participant marker ʔɜ-

With the previous sections in mind, it would be tempting to argue that Kanise Khumi patterns like Lemi. In other words, minor syllables are toneless and major syllables are tone-bearing. In contrast to Mro Khimi, we do not find the evidence to support the hypothesis that all affixes are deriving their surface tone in relation to the verb root. Yet the preverbal 3rd-person participant marker ʔɜ- creates a problem for an analysis based on syllable structure.

In some Tibeto-Burman languages, verbs can be marked for pronominal agreement (Matisoff 2003; Bauman 1974). Matisoff remarks that a language may often use a verb prefix showing agreement with a 3rd-person subject that corresponds with the possessive function of the same prefix on nouns. He gives Lai Chin as an example. Kanise Khumi has a parallel set where the 3rd-person agreement verbal agreement marker ʔɜ- (9a) resembles the minor syllable prefix ʔə- on nouns with inalienable possession (such as parts of the body, parts of a plant, bodily fluids). The relationship between the verbal marker ʔɜ- and the 3rd-person independent pronouns is highly transparent as illustrated in Table 11.

- (9) a. [ʔɜ] dɛɿ ‘to die’
 b. [ʔəɿ tʰajʔ] ‘fruit’

Table 11: Independent pronominal elements in Kanise Khumi

	Singular	Dual		Plural
		Inclusive	Exclusive	Inclusive
1st	<i>kai</i> [kaiɿ]	<i>ai hni</i>	<i>kai hni</i> [kaiɿ ɲɿɿ]	<i>ai ci</i>
2nd	<i>nang</i> [nãɿ]	<i>nang hni</i> [nãɿ ɲɿɿ]		<i>nang ci</i> [nãɿ tsɿɿ]
3rd	<i>y ni</i> [ʔɜ(ɿ) niɿ]	<i>(y) ni hni</i> [nĩɿ ɲɿɿ]		<i>(y) hni ci</i> [niɿ tsɿɿ]

While many Chin languages have robust pronominal agreement, such agreement is weaker in Khomic languages. So-Hartmann wrote that there is no pre-verbal participant reference in Lemi and only object agreement marking in Mro. Peterson (Peterson 2002) explains that participant reference marking can be elicited in Bangladesh Khumi, but it is not frequent in his corpus of texts. It is likeliest to be found

as encoding speech act participants in reported speech in narratives, in perspective shifting, and in speech act participants in conversation. He specifically states that he cannot speak for varieties in Myanmar. In the Kanise word list elicitation, ʔ₃- was the most frequent preverbal element by far, though not obligatory.

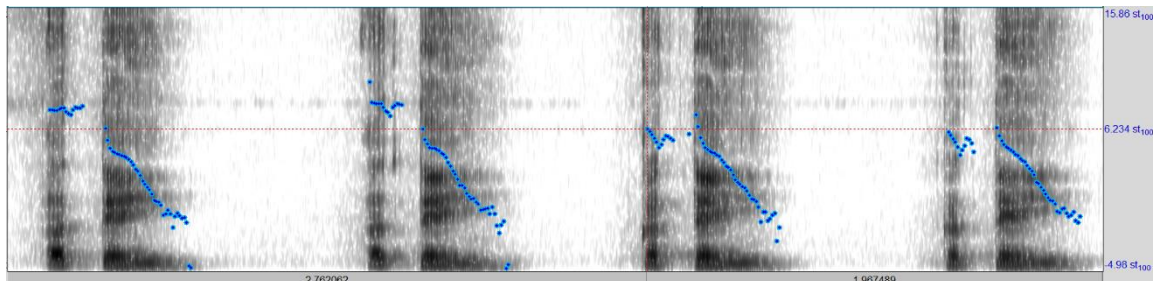
The speaker parsed ʔ₃- as a separate element a few times during the word list elicitation, but rarely. The pitch patterns are displayed in Table 11. There was no indication of polar tone. Overwhelmingly, the peak of the first syllable is higher than the peak of the second syllable. There was more variation within the HM pattern than in any other. Voice quality of the first syllable tone sounds glottalized when preceding a glottal plosive onset.

Table 11: Pitch patterns with ʔ₃-

Pitch Pattern	Examples	# of items
HH	[ʔ₃] tʰo] ‘sour’	18
HF	[ʔ₃] so] ‘spicy, hot (of peppers)’	29
HM	[[ʔ₃] tʰo] ‘to whittle, cut out, to shape, to sculpt’	45
varies HM & MM	[ʔ₃] tsaj] ‘clean’	16
MM	[ʔ₃] ke] ‘to gnaw’	1

Bryant generally transcribed the marker as ʔ₃-, but sometimes he used ʔ₃-. Thus far, acoustic analysis has not indicated a distinction between the two (Wen this volume). In terms of duration and vowel quality, the marker appears remarkably like the minor syllables occurring with *p-*, *t-*, *k-*, and *s-*. Yet pitch seems to be the feature that sets ʔ₃- off from the others. Figure 4 illustrates this difference. The speaker gave two alternatives for ‘true, real, factual’: first [ʔ₃] tẽ] and then [kə] tẽ]. It is not clear how these differ semantically, but the crucial point is that the figure displays the difference in pitch between ʔ₃- and *k-*. The first two repetitions are [ʔ₃] tẽ] and the second two are [kə] tẽ].

Figure 4: Praat visualizations of [ʔ₃] tẽ] and [kə] tẽ] ‘true, real, factual’



The tendency for ʔ₃- to be realized with a high pitch begs the question of whether it bears Tone 1, the high-level modal tone. Indeed, the phonetic cues for Tone 1 were evident when the speaker parsed ʔ₃- separately from the rest of the verb in trisyllabic verbs. In (10), the speaker first pronounced the verb ‘to chop down, to knock down’ as [ʔ₃] pə] tɔ]. He then parsed the verb into two parts: [ʔ₃] and [pə] tɔ]. Next, he gave the verb in context with an object complement ‘tree’ [tʰɛ:] kũ:] ʔ₃] pə] tɔ]. Finally, he parsed the object separately from the verb: [tʰɛ:] kũ:] was repeated three times and then [ʔ₃] pə] tɔ] was repeated three times.

- (10) tʰɛ:] kũ:] ʔ₃] pə] tɔ]
‘tree’ ‘chop down’

In this way, the speaker made units of analysis visible. ʔ₃- is a unit that can be pronounced in isolation. When it is spoken in isolation, it carries the phonetic cues of Tone 1. The duration is consistent with Tone 1 monosyllables; the voice quality is modal, and the pitch is high level. Thus, ʔ₃- differs from the other non-final syllables with a short, mid-central vowel, namely *p-*, *t-*, *k-*, *s-*, in that it is separable from the verb stem and in that it seems to be specified for tone at some level. A psychological distinction is

also evident in the orthography. In word-final syllables, a mid-central vowel is represented as <ay> and a mid-central vowel in non-final “minor syllables” *p-*, *t-*, *k-*, *s-* is represented as <a>. On the other hand, the 3rd-person agreement verbal agreement marker is represented as <y>.

4 Conclusion

The analysis of tone patterns in Kanise Khumi citation verbs differ from the analysis of Mro Khimi and Lemi Chin. In Mro Khimi, there were varied tone melodies with each morpheme, such as postverbal *-de* and preverbal elements like *mə-* and *kə-*. This variation motivated Hornéy’s analysis that affixes are underlyingly toneless, with tone being specified through post-lexical processes like melody association and polar tone assignment. While there are a few exceptions in each category, the Kanise Khumi morphemes examined in this paper tend to surface with predictable pitch, not variable pitch. The pitch of an affix does not appear to be sensitive to the pitch of the verb root. Two morphemes, *-ta* and *ʔə-*, do involve a greater proportion of variation than the other morphemes discussed here. It is important to note, however, that with these two morphemes, the variation in pitch across words is matched with a high amount of variation across tokens of the same word.

It is also difficult to characterize Kanise Khumi based on syllable structure. In Lemi Chin, full syllables bear tone, whereas minor syllables are phonologically toneless, surfacing with a neutral mid pitch. Minor syllables in Lemi Chin also reflect other constraints that are widespread in Southeast Asian languages. The duration is noticeably short in comparison to full syllables, producing an iambic rhythm. Onset consonants in these syllables are restricted in number and complexity of phonemic features. Phonetic vowel quality is loosely restricted to a range of mid-central vowel qualities. These same features are evident in several of the Kanise Khumi prefixes, notably the valency-increasing prefixes *p-* and *t-*, and the adjectivizer *k-*. With predictable mid pitch, these prefixes support an analysis based on syllable structure in which full syllables are tone-bearing and minor syllables are toneless. Prefixes *a(ŋ)-* and *ka(ŋ)-* would fit with this analysis in that the vowel is not “reduced” and they allow coda consonants. Furthermore, they do not share the same pitch patterns as *p-*, *t-*, and *k-*, even though they have similar valency affecting and adjectivizing functions. Valency-decreasing *a(ŋ)-* could be analyzed as a full syllable that bears Tone 1, the high-level modal tone. The adjectivizer *ka(ŋ)-* could be interpreted as a full syllable that bears Tone 2, the short mid glottalized tone. Such an analysis would be supported by the rare occasions when the speaker parsed *a(ŋ)-* and *ka(ŋ)-* separately from the verb stem. In these cases, *a(ŋ)-* carried Tone 1 and *ka(ŋ)-* carried Tone 2.

Nevertheless, the pitch patterns involving the 3rd-person agreement marker *ʔə-* pose a problem for such an analysis based on syllable structure. This marker shares many of the same features as a minor syllable: (1) non-final, (2) short duration, (3) mid-central vowel quality, (4) no coda consonant. Yet, the pitch of *ʔə-* cannot be predicted in the same way that it is for the other minor syllable preverbal elements. The pitch of the other minor syllables *p-*, *t-*, *k-*, *s-* is usually mid. While the pitch of *ʔə-* is mostly predictable, it is generally high, not mid. Unlike the other minor syllables, *ʔə-* is separable from the verb stem. When separated, it clearly bears Tone 1, a high-level modal tone. Table 12 highlights the transitional status of *ʔə-*. In some ways, *ʔə-* patterns like the minor syllable prefixes *p-*, *t-*, *k-*, *s-*. In other ways, *ʔə-* patterns like the full syllables *a(ŋ)-* and *ka(ŋ)-*.

Table 12: Summary of non-final syllable features

	<i>p-</i> , <i>t-</i>	<i>k-</i>	<i>s-</i>	ʔ_3-	<i>a(ŋ)-</i>	<i>ka(ŋ)-</i>
semantic function	✓	✓	?	✓	✓	✓
iambic rhythm	✓	✓	✓	✓	✓	✓
restricted onset /p,t,k,ʔ,s/	✓	✓	✓	✓	✓	✓
coda permitted	✗	✗	✗	✗	✓	✓
vowel quality	MCV	MCV	MCV	MCV	a	a
separable	✗	✗	✗	✓	✓	✓
orthographic vowel	<a>	<a>	<a>	<y>	<a>	<a>
predictable pitch	M	M	M	H	H	M
voice quality (in isolation)	✗	✗	✗	modal (Tone 1)	modal (Tone 1)	glottalized (Tone 3)

(✗ = unattested)

The strong preference for iambic rhythm in Kanise Khumi words means that non-final syllables almost always get “reduced” in lexical derivation. Lower amplitude, intensity, and shorter duration is usually evident even with compounds in Kanise Khumi. So-Hartmann and Peterson commented on the productivity of the minor syllable prefixes in other Khomic languages, arguing that there was evidence that they may be fossilized. The ability to separate ʔ_3- , *a(ŋ)-*, and *ka(ŋ)-* from the base may be an indication that these are still productive morphemes in Kanise Khumi. It may also be evidence that ʔ_3- relates to the verb stem differently than *p-*, *t-*, *k-*, *s-*. This conclusion is supported by the difference in pitch realizations where ʔ_3- is generally high and the others are usually mid.

It does not seem likely that all affixes in Kanise Khumi are toneless. Preverbal *a(ŋ)-*, *ka(ŋ)-*, ʔ_3- as well as postverbal *-ta* appear to be specified for tone, at least phonetically. Similarly, tone patterns on Kanise Khumi verbs cannot be explained by syllable structure alone, due to the fact that ʔ_3- patterns like *a(ŋ)-* and *ka(ŋ)-* rather than *p-*, *t-*, *k-*, *s-*. Instead, tone patterns on Kanise Khumi affixes need to be analyzed with reference to each affix. Furthermore, the morphological structure cannot be neglected. West (2014) proposed a typology of minor syllables based on gradient realizations of pitch: (1) non-TBU, optionally epenthized; (2) morpheme-internal, lexically specified TBU; (3) stem-external, lexically specified TBU; (4) reduced, not phonemically distinct; (5) reduced, phonemically distinct. It is possible that ʔ_3- patterns differently because it is a stem-external minor syllable, whereas *p-*, *t-*, *k-*, *s-* may no longer be functional morphemes. Perhaps they have become fused with verb root through a process of lexicalization. A great deal more data would need to be analyzed to investigate this hypothesis. At this point, it is enough to say that we must take each morpheme on its own terms when it comes to tone analysis.

References

- Baleno, Grace. 2020. Surface pitch patterns on Kanise Khumi sesquisyllabic nouns. Research Report. Trinity Western University: Canada Institute of Linguistics.
- Bauman, James. 1974. Pronominal verb morphology in Tibeto-Burman. *Linguistics of the Tibeto-Burman Area* 1.1:108–55.
- Brunelle, Marc & James Kirby. 2016. Tone and phonation in Southeast Asian languages. *Language and Linguistics Compass*. Wiley Online Library 10. 4:191–207.
- Bryant, Joseph. 2020. Kanise Khumi Wordlist Data. Archive. <https://zenodo.org/record/4029982#.X7c7c2gzaUk>.
- Butler, Becky Ann. 2014. Deconstructing the Southeast Asian sesquisyllable: a gestural account. Ithaca, New York: Cornell University PhD dissertation.
- Coupe, Alec. 2007. *A grammar of Mongsen Ao*. Berlin: Mouton de Gruyter. Cambridge: Cambridge University press.

- Dixon, Robert MW & Alexandra Y Aikhenvald. 2002. Word: a typological framework. In Robert MW Dixon & Alexandra Y Aikhenvald (eds.), *Word: a typological framework*, vol. 1, 1–41. Cambridge University Press.
- Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2020. *Ethnologue: Languages of the World*. Twenty-third edition. Dallas, Texas: Twenty-third. Dallas, TX: SIL International. Online version: <http://www.ethnologue.com>. (25 August, 2020).
- Evans, Jonathan P. 2009. Is there a Himalayan tone typology? *Senri Ethnological Studies*.
- Gehrmann, Ryan. 2017. The historical phonology of Kriang, a Katuic language. *Journal of the Southeast Asian Linguistics Society* 10. 1: 114–139.
- Henderson, Eugénie JA. 1965. The topography of certain phonetic and morphological characteristics of South East Asian languages. *Lingua*. Elsevier 15:400–434. [https://doi.org/10.1016/0024-3841\(65\)90020-3](https://doi.org/10.1016/0024-3841(65)90020-3).
- Henderson, Eugénie JA. 1967. Grammar and tone in South-East Asian languages. *Wissenschaftliche Zeitschrift der Karl-Marx-Universität-Leipzig* 16:171–78.
- Herr, Kristen Elizabeth. 2011. The phonological interpretation of minor syllables, applied to Lemi Chin. Chiang Mai, Thailand: Payap University MA thesis.
- Hornéy, Christina Scotte. 2012. A Phonological Analysis of Mro Khimi. Grand Forks, ND: University of North Dakota MA thesis.
- Hyman, Larry M. 2006. Elicitation as experimental phonology: Thlantlang Lai tonology. *UC Berkeley PhonLab Annual Report* 2. 2.
- Hyman, Larry M. 2007. Kuki-Thaadow: An African tone system in Southeast Asia. *UC Berkeley Phonology Lab Annual Report* 3.
- Hyman, Larry M & Kenneth VanBik. 2002. Tone and syllable structure in Hakha-Lai. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 28, 15–28.
- Hyman, Larry M & Kenneth VanBik. 2004. Directional rule application and output problems in Hakha Lai tone. *Language and Linguistics* 5.4:821–861.
- Hyman, Larry M. & Kenneth VanBik. 2004. Directional Rule Application and Output Problems in Hakha Lai Tone. *Language and Linguistics* 5.4:821–861.
- Ikeda, Elissa. 2021. Tone patterns in Kanise Khumi monosyllables. In *Proceedings of the 11th Payap University and UNRN Research Symposium 2021*, 1223–1234. Chiang Mai, Thailand: Research and Academic Service Affairs.
- Kiparsky, Paul. 1982. Lexical phonology and morphology. *Linguistics in the morning calm*. Hansin.
- Lotven, Samson, Kelly Berkson, James C Wamsley, Jillian Danaher, Kenneth Van Bik & Stuart Davis. 2020. The syllable in Kuki-Chin. *Journal of South Asian Languages and Linguistics*. De Gruyter 6. 2:281–308.
- Matisoff, James A. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University of California Press. <https://escholarship.org/uc/item/19d79619>.
- Meyase, Savio Megolhuto. 2018. Tenyidie: Another African Tone System in Southeast Asia? In *Proc. TAL 2018*, Sixth International Symposium on Tonal Aspects of Languages, 169–172.
- Pain, Frederic, Michel Ferlus, Alexis Michaud, Thi Thu Hà Phạm & Ryan Gehrmann. 2019. EFEO-CNRS-SOAS Word List for Linguistic Fieldwork in Southeast Asia. Identifier: oai:halshs.archives-ouvertes.fr:halshs-01068533. <http://halshs.archivesouvertes.fr/halshs-01068533>.
- Peterson, David A. 2002. On Khumi Verbal Pronominal Morphology. *Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society: Special Session on Tibeto-Burman and Southeast Asian Linguistics* 99–110.
- Peterson, David A. 2012. The Khumi cluster and it's place in Kuki-Chin. Paper presented at ICSTLL 45. Nanyang Technological University, Singapore.

- Peterson, David A. 2013. Affecting valence in Khumi. *Language Typology and Historical Contingency: In honor of Johanna Nichols* 171–194.
- Peterson, David A. 2017. On Kuki-Chin subgrouping. In Picus Sizhi Ding & Jamin Pelkey (eds.), *Sociohistorical linguistics in Southeast Asia*, 189–209. Leiden: Brill.
- Peterson, David A. 2018. Kuki-Chin utterance-final particles. In *Proceedings of the 51st International Conference on Sino-Tibetan Languages and Linguistics*. 第 51 回国際漢蔵語学会実行委員会・京都大学白眉センター. https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/235303/1/proc_icstll51_51.pdf.
- Peterson, David A. 2019a. Bangladesh Khumi. In Alice Vittrant & Justin Watkins (eds.), *The Mainland Southeast Asia Linguistic Area* (Trends in Linguistics Studies and Monographs), vol. 314, 12–55. Berlin & Boston: Walter de Gruyter GmbH & Co KG.
- Peterson, David A. 2019b. Progress in the analysis of Kuki-Chin Morphosyntax. Presented at the Workshop on Recent Advances in Kuki-Chin Linguistics, SOAS, University of London. <https://zenodo.org/record/3612149/files/soas2019petersonhdt.pdf?download=1>.
- Sarmah, Priyankoo, Leena Dihingia & Wendy Lalhminghlui. 2015. Contextual variation of tones in Mizo. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- So-Hartmann, Helga. 1988. Notes on the Southern Chin languages. *Linguistics of the Tibeto-Burman Area* 11.2:98–119.
- So-Hartmann, Helga. 1989. Morphophonemic changes in Daai Chin. *Linguistics of the Tibeto-Burman Area* 12. 2:51–66.
- So-Hartmann, Helga. 2008. Nominalization and Relativization in Mro. In. SOAS, University of London.
- So-Hartmann, Helga. 2013. Valence-changing Prefixes in Kuki-Chin. Paper presented at ICSTLL 46. Dartmouth College.
- So-Hartmann, Helga. Lemi Functional Prefixes.
- Tan, Wenhao. 2021. Preliminary acoustic analysis of minor syllables in Kanise Khumi. Presented at the SEALS30.
- VanBik, Kenneth. 2006. Proto-Kuki-Chin. Berkeley, CA: University of California, Berkeley PhD dissertation. <https://escholarship.org/content/qt5283829m/qt5283829m.pdf>.
- West, Luke. 2014. Pitch specification of Sgaw minor syllables. University of Toronto PhD dissertation.

SYNTACTIC AND SEMANTIC BOUNDNESS OF NOUN PHRASES IN BURMESE

Mathias JENNY¹

University of Zurich
mathias.jenny@uzh.ch

Abstract

This paper investigates the correlation of form and function in Burmese noun phrases. Burmese has several ways of combining nouns with verbal and other modifiers, which exhibit different degrees of syntactic boundness and semantic integration. The claim is that syntactic and semantic boundness are predictive in that syntactically more tightly bound expressions coincide with closer semantic integration. The results partly confirm this claim though there are other factors determining the choice of a specific construction in a given context in Burmese. The study, which concentrates on verbal modifiers, is based on a wide range of sources, including grammatical descriptions, lexica, as well as formal and colloquial texts. Additional examples have been constructed and checked with native speakers.

Keywords: Burmese, Noun phrase, word order, semantics and syntax

ISO 639-3 codes: mya

1 Introduction

In the present study, we investigate different types of noun-modifier collocations in Burmese. These constructions exhibit clearly different degrees of syntactic boundness, and they are in most cases not, or not completely, coextensive in their functional ranges. That is, in many cases, the speaker has to choose one among several available constructions, the choices often being not interchangeable. Of special interest are the NPs containing a noun and a verbal modifier, which exhibit the widest range of formal possibilities, each with its own inherent semantics. Although the expressions are not generally mutually exchangeable, the syntax-semantics correlation is not always obvious and requires an in-depth analysis. One hypothesis that is looked into in this study is the idea that formal (morphosyntactic) boundness iconically corresponds to functional (semantic) boundness (e.g., papers in Haiman 1985). There is evidence that this claim is at least partly true in Burmese, though it is not the only factor determining the choice of a specific construction over another one in a given context.

The noun attributes investigated in this paper are preposed and postposed verbal and nominal modifiers of the types *mjìN-p^hju* (*mjìN-bju*)² ‘horse-be.white’,³ *p^hju-ðò-mjìN* ‘be.white-ATTR-horse’, and *mjìN-ʔəp^hju* ‘horse-NML.be.white’, all translated as ‘a/the white horse’. These expressions are treated as synonyms by Pe Maung Tin (1963), and are generally not analyzed in indigenous grammars, such as the latest edition of the *Myanmar Thadda* (Myanmar Language Commission 2005). Obviously for the Burmese, they pose no problem and do not appear to deserve room for elaboration in prescriptive grammars.

¹ This paper was planned to be written in collaboration with Rudolf Yanson who unexpectedly passed away in May 2021. It is a sad honor for me to finish this piece of work on which we had been working together for over a year. The draft was well advanced when we last were in contact in late April, and I made every effort to finish it to reflect Rudolf’s perspective as much as my own.

² Voicing occurs in close juncture after plain and nasalized vowels. It is regular in grammatical morphemes, but less so in lexical items. Where voicing is consistent (and in some cases phonemic), it is indicated in this paper. In cases where there is variation, either individual or regional, the unvoiced form is given.

³ In glossing Burmese verbs, we use the copula ‘be.V’ for stative verbs (corresponding to adjectives in English) and ‘to.V’ for active verbs.

2 Types of noun modifiers

In Burmese, there are several types of constructions linking nouns with different types of attributes, some with clear morpho-syntactic structure and transparent semantic composition. Other constructions are more challenging, both in terms of syntax and semantics. In particular, the latter deserve a closer investigation.

2.1 Nominal modifiers

The first type is represented by constructions containing two nouns [N1N2], where N2 unambiguously functions as head and N1 as an attribute. This pattern is consistent with the general head-final structure of Burmese. Examples of this type are *wɛʔ-θà* ‘pork’ (lit. ‘pig-flesh’) and *cè-jouʔ* ‘statue’ (‘brass-figure’) (Jenny & Hnin Tun 2016:112ff). [N1N2] constructions are different from superficially similar expressions where N1 is a possessive modifier. In this case, N1 is marked as possessive (or dependent) by the “induced creaky tone” (Okell 1969) where possible. This morphological distinction is neutralized if N1 end in a glottal stop or creaky tone. Compare *lu-mjò* ‘ethnicity’ (‘man⁴-kind’) with *lú-bəwá* ‘human life’ (‘man.DEP-life’). The former may be termed a NOMINAL COMPOUND, the latter a NOUN PHRASE. Another type of [N1N2] contains two nouns in appositional relationship. That is, they both are syntactically speaking heads, both referring to the same entity, one being more general, the other more specific. This type of construction is seen in kinship and social terms with proper names, as in *s^həja wɪn-ʔaun* ‘teacher Win Aung’. A further type of [N1N2] construction is seen in coordinate compounds, where none of the nouns function as modifier. An example of this type of construction is *ʔəp^he-ʔəme* ‘parents’ (‘father-mother’). This last type is clearly beyond the scope of this paper.

2.2 Verbal modifiers

Verbal modifiers can be combined with nouns in four different constructions. In the following subsections, the four types of N-V collocations will be presented with their formal and functional properties.

2.2.1 Marked attributives

The most frequent and flexible of the modifying constructions is the one consisting of a verb attached to the following noun by a special syntactic marker *θə/ðə* in formal or *té/dé* in colloquial language. These two markers will be glossed as ATTR in all examples, and the construction is given here as [VATTRN]. Relevant examples are *cì-dé/ðə ʔeɪN* ‘a/the big house’ (‘be.big-ATTR house’) and *nì-dé/ðə kà* ‘a/the red car’ (‘be.red-ATTR car’).

The syntactic structure of [VATTRN] is identical to more general attributive (relative) clauses modifying a nominal head, with the clause consisting minimally of a verbal predicate. In the same fashion, a more complex clause can be attached to a noun, including full arguments, TAM, polarity, and other categories. Attributive clauses in the pattern [CLAUSEATTRN] also allow tense modification of the clause, with the attributivizer appearing either as non-future *té/dé* (literary Burmese *θə/ðə* or *θi/ði*) or future *mé* (literary Burmese *mji*).⁵ The examples above could therefore be translated as ‘a/the house that is big’ and ‘a/the car that is red’, respectively. In this paper, we will discuss only unmarked attributive verbs of this pattern, without overt arguments and TAM/polarity marking.

2.2.2 Nominalized verb as modifier

Another type of construction, the most challenging one in terms of form and function, consists of a noun and a postnominal verb in nominalized form, [VaN] (see also Nosova 1974). The verbal noun is formed by adding the prefix *ʔə-* to any verb, the most common and flexible means to form nouns from verbs.

⁴ The gloss ‘man’ for *lu* is to be understood in the sense ‘human being’, irrespective of gender.

⁵ Speakers of Burmese employ two different styles of the language, viz. literary and colloquial, distinguished mainly by their grammatical forms (Okell 1994). The former is used in formal contexts, the latter in informal spoken settings. While literary Burmese used to be the only language of writing, colloquial Burmese appears increasingly in modern prose, sometimes in a mixed style.

This type of construction is seen in expressions like *ʔeɪN ʔəci* ‘a/the big house’ (‘house-NML.be.big’) and *kà ʔəni* ‘a/the red car’ (‘car-NML.be.red’).

The nominalizing prefix *ʔə-* is very old in Burmese, going back to proto-Tibeto-Burman origins, where a nominalizing prefix **ʔa-* can be reconstructed with nominalizing and other functions (Matisoff 2003:104ff).⁶ It has a general nominalizing function, with no specification as to whether the nominalized form is a verbal abstract, an instrument, an actor, object, or any other function with a logical relation to the verbal semantics. The following examples, mostly taken from Okell and Allott (2001), illustrate the different functions of [aV] nominalizations.

Base verb	Gloss	Derived noun	gloss
<i>louʔ</i>	‘to do’	<i>ʔəlouʔ</i>	‘work’
<i>wɪN</i>	‘to enter’	<i>ʔəwɪN</i>	‘entry’
<i>tʰù-sʰàN</i>	‘be unusual’	<i>ʔətʰù-ʔəsʰàN</i>	‘something unusual’
<i>θi</i>	‘to know’	<i>ʔəθi</i>	‘knowledge, acquaintance’
<i>θi</i>	to.bear.fruit	<i>ʔəθi</i>	‘fruit’
<i>kàUN</i>	‘be good’	<i>ʔəkàUN</i>	‘goodness, a good one’
<i>sʰo</i>	‘to sing, say’	<i>ʔəsʰo-dɔ</i>	‘singer’ ⁷

The questions arising about the [NaV] constructions are both syntactic and semantic. It is not at first sight obvious what the syntactic make-up of [NaV] expressions is, that is, which part (if any) is to be seen as head, which as modifier, and how the two parts relate semantically. These points will be addressed in more detail in section 4.

2.2.3 Bare verb as modifier

While [NaV] are frequent and productive without any restrictions other than those caused by semantic clashes, they seem to be in competition with what looks like reduced forms [NV], that is, a noun modified by a bare verb without the nominalizing prefix. Typical examples of [NV] constructions include lexicalized expressions with specific meaning, such as *ʔeɪN-bju-dɔ* ‘the White House’ (‘house-white-HONORIFIC’) and *ɛaʔ-ni* ‘Red Shirts’, *ɛaʔ-wa* ‘Yellow Shirts’ (former political movements in Thailand), literally ‘shirt-be.red’ and ‘shirt-be.yellow’, respectively. Apart from clearly lexicalized expressions with partly idiosyncratic meaning not fully derivable from the parts, there are cases where [NV] appears interchangeably with [NaV], as in *kà-hàUN* ‘used car’ and *kà-θiʔ* ‘new car’ besides *kà-ʔəhàUN* and *kà-ʔəθiʔ*, with the same meanings. As with [NaV], the internal syntactic structure and semantic composition of these constructions is not obvious and deserves more detailed investigation. This should lead to a better understanding of the factors favoring one or another construction in a given context.

Superficially, there are constructions containing active verbs of the [NV] type where V is not a modifier of N, but these [NV] expressions are (usually non-headed) lexicalized compounds, and the relation ‘noun-modifier’ is not traceable within such constructions. Relevant examples are *tʰəmɪN-jeʔ* ‘a cook’ (from *tʰəmɪN cʰeʔ* ‘to cook rice’), *jɪN-màUN* ‘driver’ (from *jɪN màUN* ‘to drive a vehicle’) and *zəbwè-dò* ‘waiter’ (from *zəbwè tʰò* ‘to strike tables’). In other, less frequent [NV] compounds, the V functions as a semantic head, as in *lu-bjɔ̀-θu-bjɔ̀* ‘hearsay, rumor’ (from *lu pjɔ̀ θu pjɔ̀* ‘man to.speak person to.speak’) and *lu-zú* ‘group, crowd’ (from *lu sú* ‘man to.gather’).⁸ N in these examples denotes the patient (object) or agent (subject) of the verb, unlike the NV compounds under consideration here, where N is the semantic head and always denotes the subject of a stative V. One group of [NV]

⁶ The label “nominalizing” is not entirely adequate, as the prefix is also added to already nominal bases, as in kinship terms (*ʔəpʰe* ‘father’, *ʔədɔ* ‘aunt’) and classifiers (*ʔəkaUN* ‘body; CLF for animals’, *ʔəpin* ‘trunk,tree; CLF for plants), among others. The gloss NML for ‘nominal’ is used here in the examples, not specifying the exact semantics of the nominalization.

⁷ *ʔəsʰo-dɔ* is lexicalized with the HONORIFIC/ROYAL suffix *-tɔ/dɔ*.

⁸ In all these lexicalized items, the second part is always voiced; there is a distinction between voiced and non-voiced V in all these cases, the latter being a V with O combination (phrase), the former an NP.

constructions that falls somewhere between the two is exemplified by expressions like *je-dε?* ‘rising tide’ (from *je tε?* ‘water to.rise’) and *ne-bu* ‘heat of the sun’ (from *ne pu* ‘sun be.hot’). In the latter expressions, N is the subject of V, which may or may not be stative, and V is not necessarily a modifier to N. All of these constructions are beyond the scope of this paper and will not be discussed further here.

2.2.4 Reduplicated modifier

The last type of N-V construction, consisting of a noun with an adjoined reduplicated verb, is semantically different from the previous ones, as it adds an emotional or subjective connotation to the modifier. This value can be augmentative (‘very X’) or diminutive (‘rather/quite X’). In both senses, the reduplicated attributive is frequently introduced by the particle *k^ha?*. The formula used in this paper is [NVV], reflecting the more common order of elements. Examples are *ʔeɪN cì-ʔì* ‘a rather/very big house’, lit. ‘house be.big-be.big’, and *kà ni-ni* ‘a reddish car, a quite/very red car’, lit. ‘car be.red-be.red’.⁹ In both cases, *k^ha?* ‘rather’ may be added, yielding *ʔeɪN k^ha?-cì-ʔì* and *kà k^ha?-ni-ni*, respectively.

The order of elements is generally a noun followed by the reduplicated verbal modifier, with or without *k^ha?*, though Okell (1969) also gives examples with the opposite order for [NVV] and [NaV] constructions, as seen in the following examples.¹⁰ Okell (1969:81) states that:

both *ǎV* attributes (when not tightly linked) and those derived by other formatives are sometimes found in the reverse order, i.e., with the attribute before the head, e.g.,

<i>pja-bja ʔèɪŋjì</i>	as well as	<i>ʔèɪŋjì pja-bja</i> ,
<i>k^ha?-pja-bja ʔèɪŋjì</i>	as well as	<i>ʔèɪŋjì k^ha?-pja-bja</i> ,
<i>ʔəpja ʔèɪŋjì</i>	as well as	<i>ʔèɪŋjì ʔəpja</i> ¹¹

The reduplicated attributives, when preposed as suggested by Okell (1969), can be ambiguous as to their function as noun attributes or adverbs. The expression *kàUN-gàUN ʔeɪN-hma nei-dε* is most likely interpreted as ‘they stay home well’, rather than ‘they live in a good house’. The latter is more likely to be expressed as *ʔeɪN kàUN-gàUN-hma ne-dε*, the former unambiguously as *ʔeɪN-hma kàUN-gàUN ne-dε*. This type of construction, which is semantically clearly different from the other three, is not further investigated in this paper and left here for further analysis. In this study, we will concentrate on the constructions of the types [VATTRN], [NAV], and [NV].

3 Syntactic and semantic comparison of noun-modifier expressions

3.1 Different form - different function?

The four types of verbal modifiers with nouns can be summarized as follows:

[VATTRN]	<i>cì-dεì ʔeɪN</i>	‘a/the house that is big’
[NaV]	<i>ʔeɪN-ʔəcì</i>	‘a/the big house’
[NV]	<i>ʔeɪN-ʔì</i>	‘a/the big house’
[NVV]	<i>ʔeɪN-cì-ʔì</i>	‘a/the rather/very big house’

As can be seen from the translations given above of the different constructions, they seem to be semantically identical and in relevant publications are usually treated as synonymous. Thus, Okell (1969:68ff) translates the expressions *pu-dε-je* [VATTRN], *je-bu* [NV], and *je-ʔəpu* [NaV] all invariably

⁹ The voiceless initial of the first part of the reduplicated verb suggests that there is no close juncture here, unlike in [NV] compounds seen in 2.2.3.

¹⁰ Expressions of the type [VVN] and [aVN] are not accepted as grammatical by all speakers and are hardly (if at all) found in spontaneous conversation.

¹¹ The transcription and glosses of examples from all sources have been unified.

as ‘hot water’. Hopple (2011:71f) also does not differentiate between the constructions [NV] and [NaV], as is evident from her translation of both *ʔeiN-bju* and *ʔeiN-ʔəp^hju* as ‘white house’.

In a recently published descriptive grammar of Burmese (Jenny & Hnin Tun 2016) the three constructions are translated as follows: *ʔeiN-hàUN* [NV] ‘old home, place of origin’ (lit. ‘house-be.old’), *ʔeiN-ʔəhàUN* [NaV] ‘a/the old house’, and *hàUN-dé ʔeiN* [VATTRN] ‘a/the old house’ (Ibid.:152). The difference between [NV] and [NaV] is given as “unlike bare verbs following nouns, the nominalized forms are not restricted to fixed expressions and are less lexicalized” (Jenny & Hnin Tun 2016:152). This suggests that [NV] expressions are only found in conventionalized or lexicalized compounds, not as a productive syntactic process. The data given below will show this to be too narrow an explanation, as there is some overlap between [NV] and [NaV], the former being formed spontaneously and productively. Most published grammars and descriptions of Burmese are silent on the question whether [NV], [NaV] or [VATTRN] is used in which context, presenting them as homonymous without further explanation. The impossibility of substituting one for the other in most cases speaks differently, though. In a publication about NPs in Burmese, Tagunova (1971) writes with respect to the difference between [VATTRN] and [NV] that

prepositional attributes (i.e., [VATTRN] constructions, R.Y.) express an individual quality related to a concrete surrounding and are used to underline a quality or peculiarity of the concrete person or subject. Postpositional attributes (i.e., [NV] constructions, R.Y.) are used to describe qualities or peculiarities of the whole class of the persons or subjects. They indicate a permanent quality, and are of descriptive, not determining nature. (Tagunova 1971:49; translation R.Y.)

It is worth mentioning that the author ignores constructions of the [NaV] type all together, which, besides the vagueness of the definitions given above, does not contribute much to solving the problem of the interrelation between the three constructions under discussion here. Regarding the semantics of the [VATTRN] constructions, the author mainly bases her conclusion on the fact, that such attributes can have their own grammatical markers (see above, 2.2.1), which naturally makes them more concrete and contextually adaptive than the other constructions, and as such they may sometimes correspond to situations where an English definite article is used.

Despite the apparent synonymy of the three constructions, they almost never allow substitution of one by another in a concrete speech situation. The main task of this study is to find out the semantic differences between the constructions and the reasons why native speakers use this or that construction in a given context, as well as how the syntactic and semantic differences correlate.

There are some structural differences between the three constructions. The construction [VATTRN] is different from other two in that it can contain any verb, active or stative, preserving grammatical markers such as TAM and negation, e.g., *sèiN-dè-ATTR jwɛʔ* (‘green-still-ATTR leaf’) ‘leaf which is still green’. Syntactically, this construction is clearly clausal, and semantically transparent. In the constructions [NV] and [NaV] only stative verbs can occur, and only in isolated form without any verbal modifiers. They are more similar to European attributive adjective expressions, and it is their syntactic status and semantic range that is the topic of analysis in the following.

We now turn to the syntactic features underlying [NV] and [NaV] constructions.

3.2 [NV] and [NaV] - background

As seen above in 2.1, nouns can be modified by other nouns, either in attributive/associative or possessive relationship. The latter is usually indicated morphologically by the creaky tone on the last syllable of the possessor, if the syllable is creakable, although with non-human possessors, the induced creaky tone is rather rare, neutralizing the formal distinction between attributive and possessive relations. Thus, the pattern [N1N2] can encode a nominal head with another noun as its modifier. In this case it is invariably N2 that functions as head of the compounds, N1 being the modifier. Other possibilities are appositional collocations and additive compounds. When analyzing the [VaN] constructions, where [aV] is taken to be a noun, these possibilities have to be taken into consideration.

Verbs, both active and stative, can take the general nominalizing prefix *ʔə-*, resulting in any of several possible nominal readings. The semantics of a given [aV] as event or participant nominalization depends on the semantics of the verb as well as the concrete context, and in many cases different

meanings apply in different contexts, as illustrated above in 2.2.2. This semantic variation of [aV] nominalizations is an important factor to be counted in when analyzing the syntactic and semantic structure of [NaV] expressions. Despite the wide range of possible functions of [aV] forms, they can all be analyzed as underlyingly nominal, with different contextual readings, including adverbial and attributive.

The third point that is relevant to the discussion is the fact that the nominalizing prefix *ʔə-* is dropped in certain contexts, such as compounds and NUMERAL-CLASSIFIER combinations, where the classifier regularly loses the prefix: *ʔəθi* ‘fruit’ (from the verb *θi* ‘to bear fruit’), but *θəjeʔ-θi* ‘mango’ (lit. ‘mango-fruit’) (Okell & Allott 2001:256). Furthermore, there is the tendency in colloquial Burmese, both spoken and written in informal contexts such as chats on social media, to drop the nominalizing prefix *ʔə-*, resulting in bare verbs appearing in the place of nominalized verbs. The following examples are taken from personal informal chats on messengers: *sə-ji tʰá-mε*. ‘I will get up early.’ (lit. ‘be.early-be.big to.get.up-FUT’) for standard *ʔəsə-ji tʰá-mε*, where *sə* ‘be early’ is nominalized (lit. ‘big earliness’). This drop of the prefix in colloquial Burmese is especially frequent in two-part compounds of nominalized verbs, where the second part appears as bare stem, as in *ʔəsi-wè* ‘meeting’ (from the verb *sì-wè* ‘to gather, meet’) for standard *ʔəsi-ʔəwè*. A quick Google search returns 2,7 million results for the full form *ʔəsi-ʔəwè* ‘meeting’, and 210,000 for the reduced form *ʔəsi-wè*, the former pointing to official announcements and state media, the latter mostly to Facebook and other informal media.

3.3 Semantic range of [NV] and [NaV]

The constructions of the [NV] type are peculiar in that quite often they represent lexicalized units, some of which are listed as distinct words with idiosyncratic meanings, e.g.: *lu-bjeʔ* ‘clown’ (lit. ‘man-to.joke’), *θu-zèin (dəzèin)/lu-zèin* ‘stranger’ (lit. ‘person/man-be.green/fresh’), and *θan-bju* ‘tin’ (lit. ‘iron-be.white’). The degree of lexicalization is rather subjective and variable among speakers. Thus, we can assume that the compilers of Burmese dictionaries include in the dictionary those [NV] constructions which they think have become lexicalized, but in different dictionaries there are likely to be different approaches to the same constructions. For example, in the Burmese-Russian Dictionary (Minina & U Kyo Zaw 1976) we find such entries as: *lu-bú* ‘dwarf’ (lit. ‘man-be.short’), *lu-wá* ‘fat person’ (lit. ‘man-be.fat/full’), both of which are absent from the Burmese-English dictionary (Myanmar Language Commission 1976). On the other hand, the Burmese-Russian Dictionary omits the following entries which are present in the Burmese-English Dictionary: *lu-luʔ* ‘unmarried person, person at large’ (lit. ‘man-be.free’) and *lu-ηε* ‘youth’ (lit. ‘man-be.young/small’). Such examples can be multiplied and show the uncertain lexical status of these compounds. This suggests that [NV] is not restricted to fully conventionalized or lexicalized expressions, but speakers are rather free to employ the construction in productive ways, and we can assume that every individual speaker has his or her own conception as to which of the [NV] constructions got lexicalized to the extent that the parts of the construction lost their original meaning and therefore may not express the meaning the speaker intends to convey. These individual differences of how lexicalized (and semantically idiosyncratic) a given [NV] expression is seen may be one of the reasons why speakers choose one or another construction in a concrete speech context.

4 The syntax of [NaV] and [NV]

4.1 The syntax of [NaV]

We now proceed to the syntax of the [NaV] construction. It is a rather infrequent in the formal language, used mostly in colloquial style, but is mentioned in all relevant descriptions of Burmese. The problem with this construction is that the component [aV], as seen above, is formally a noun and therefore according to basic syntactic rules should be treated as a head within the construction [N1N2]. Syntactically, N1 should be treated as a modifier to N2, as in other [N1N2] compounds. Still, in all relevant publications [aV] is treated as an attribute to preceding noun. Semantically, this treatment is perfectly logical. Indeed, in such constructions as *ʔeiN-ʔəpʰju* ‘a/the white house’ (‘house-NMLZ.be.white’), *mi-ʔəni* ‘a/the red light’ (‘fire-NML.be.red’) or *ηà-ʔəci* ‘a/the big fish’ (‘fish-

NML.be.big’) it seems natural to interpret the second component of the constructions as attribute to the preceding one as the translations suggest, although this interpretation contradicts basic head-final syntactic rule of Burmese. Contrary to this view, C. F. Lehman (p.c.) insisted that [aV] is to be treated as a noun in all cases, and an expression like *lu-ʔəci* ‘a/the big man’ must be analyzed as ‘a person-type bigness’, that is, with N modifying the [aV]-head. Although this seems semantically counterintuitive, it is syntactically sound.

Adding to the difficulty is the presence of clearly attributive [aV] in prenominal position [aVN], as in *ʔəmjan-làn* ‘expressway’ (‘NML.be.fast-road’), *ʔəmjan-ʔəʰà* ‘express train’ (‘NML.be.fast-train’), and *ʔəni-jaun* ‘red’ (‘NML.be.red-color’). In these expressions, there is no problem in identifying the prenominal [aV] as a modifying abstract noun, as in ‘speed-road ~ road of speed’, ‘speed-train ~ train of/with speed’, and ‘redness-color ~ color of redness’, respectively. The order is fixed in these cases, that is, there is no **làn-ʔəmjan*, **ʔəʰà-ʔəmjan*, or *jaun-ʔəni*, which suggests that these constructions are not relevant to the development of [NaV] expressions even if they are also to be analyzed as containing a nominalized verb modifying the noun.

Supporting evidence of [aV] to be taken as the head of [NaV] seems to be the fact that in [NaV] constructions the N can be omitted in all contexts without the expression changing its meaning. If we accept the rule that syntactic heads are the parts of a phrase that cannot be omitted, N should be considered a dependent, rather than the head. If in the following example (1a) *ʔəin* ‘house’ is the head, (1b) should be ungrammatical, though it is perfectly grammatical and idiomatic if the noun ‘house’ is present in the speech context. Note that the English translation in this case requires the dummy head ‘one’.

- 1a. *θu ʔəin ʔəθiʔ sʰauʔ-ne-de.*
 3 house NML.new¹² to.build-stay-NFUT
 ‘He is building a new house.’

- 1b. *θu ʔəθiʔ sʰauʔ-ne-de.*
 3 NML.new to.build-stay-NFUT
 ‘He is building a new one.’

Of course, such transformation presupposes a context, from which the ‘reduced’ form of the utterance appears. But the form [aV] can occur on its own without any concrete context, as in (2).

2. *θu ʔəθiʔ caiʔ-te.*
 3 NML.new to.like-NFUT
 ‘He likes new stuff.’

A noun may also take two or more [aV] expressions at the same time, all of them, having the same referent, as in example (3) taken from modern Burmese prose written in rather colloquial style.¹³

3. *cənoʔ-go ləpʰeʔ-je ʔəcʰo-jə ʔəpʰan-jə taiʔ-te.*
 Ism.DEP-OBJ tea-liquid NML.be.sweet-ENUM NML.be.acrid-ENUM give.to.drink-NFUT
 ‘He gave me tea, both sweet and strong.’

Thus, we may draw the preliminary conclusion that the unit [aV] exhibits all relevant features of a noun but treating it as a head of the construction [NaV], which seems natural for the nouns in this position, goes against semantic logic. Let us therefore further investigate the syntactic nature of [NaV]

¹² The lexeme *θiʔ* hardly occurs in predicative function ‘be new/fresh’ but seems to be restricted to attributive function. Therefore, we prefer the gloss ‘new’ over ‘be.new’ in this case.

¹³ Ludu U Hla. 2016. *cəno hle əθi.* [I am the captain]. 2nd edition. Mandalay: Lu Du, 258.

expressions. It is necessary to distinguish between the several types of the [NaV] constructions. One type is obviously with a possessive relation between N1 and N2. This type of construction does not pose any syntactic or semantic problems, and the interpretation is straightforward. This is not the case in [NaV] expressions with modifying function, which present problems in interpreting their semantics. As the surface structure of the different [NaV] expressions is identical, ambiguity can arise without context, as in the following examples (4) and (5) adapted from Okell (1969:92f).

4. *k^houN ʔəmjin behnə-pe lɛ̃.*
 bench NML.be.high how.many-foot CQ
 ‘How many feet high is the bench?’
5. *k^houN ʔəmjin tə-lòun lo-jin-dɛ̃.*
 bench NML.be.high one-CLF require-DES-NFUT
 ‘(He) wants a high bench.’

There is no problem with the first example as a possessive construction which could be overtly marked as such by the possessive marker *jé*. In the second one, [aV] obviously has attributive meaning. Let us consider another example (6).

6. *θu ʔein ʔaci s^hauʔ-ne-dɛ̃.*
 3 house NML.be.big to.build-stay-NFUT
 ‘He is building a big house.’

As it stands, this sentence can also be interpreted differently, taking *ʔaci* as independent from *ʔein*, having adverbial or secondary predicative (DEPICTIVE) function, ‘he is building the house big’ (see Himmelmann & Schultze-Berndt 2005).¹⁴ The difference becomes overt if the quantifier phrase *tə-lòun* ‘one-CLF’ is added. The [NUM-CLF] phrase occurs directly after the noun *ʔein* in depictive contexts, but after [NaV] in attributive contexts, as in (7a) and (7b).

- 7a. *θu ʔein tə-lòun ʔaci s^hauʔ-ne-dɛ̃.*
 3 house one-CLF NML.be.big to.build-stay-NFUT
 ‘He is building a house, and he’s making it big.’
- 7b. *θu ʔein ʔaci tə-lòun s^hauʔ-ne-dɛ̃.*
 3 house nml.be.big one-clf to.build-stay-nfut
 ‘He is building a big house.’

Two other revealing examples are given in (8) and (9).

8. *kɔzɔ-go ʔəθiʔ tə-k^hú k^hin-dɛ̃.*
 carpet-OBJ NML.new one-CLF to.spread.out-NFUT
 ‘He changed the carpet for a new one.’
9. *mjeʔ.hna-θouʔ-pəwa-go ʔəθiʔ tə-t^hɛ̃ lɛ̃-dɛ̃.*
 face-to.wipe-towel-OBJ NML.new one-CLF to.exchange-NFUT
 ‘He changed the towel for a new one.’

These examples are relevant to the present discussion in two points. First, the [aV] expression (‘NML.new’) is here separated from the noun ‘carpet’ by the object marker *-ko/go*. This leads to the

¹⁴ A similar function is seen in [aV] expressions like *ʔəwá sà* ‘eat one’s fill’ (lit. ‘NML.be.full to.eat’), *ʔəpi louʔ* ‘to complete, finish’ (lit. ‘NML.be.finished to.do’), and *ʔəsɔ̃ pjan* ‘return early’ (lit. ‘NML.be.early to.return’).

conclusion that the sequences *kɔzɔ̌ ʔəθiʔ* ‘new carpet’ and *mjɛʔ.hna-θouʔ-pəwa ʔəθiʔ* ‘new towel’ do not constitute a syntactic unit (constituent) which may be called noun phrase. Second, the component *ʔəθiʔ* (‘NML.new’) in both cases unambiguously functions as noun since it is followed by the numeral-classifier *tə-kʰú* and *tə-tʰɛ*, respectively. Unlike *ʔəci* ‘NML.be.big’ in (7) above, *ʔəθiʔ* cannot be interpreted here as depictive. Rather, this example shows that [aV] has some degree of syntactic independence of N even if it has attributive (modifying or specifying) function. This independence is further shown by the fact that the object marker *-ko/go* can appear either after the noun or after the N or after [aV] (with the [NUM-CLF] phrase), as seen in (10) and (11).

10. *kɔzɔ̌ ʔəθiʔ tə-kʰú-i-go kʰiŋN-dɛ.*
 carpet NML.new one-CLF-OBJ to.spread.out-NFUT
 ‘He changed the carpet for a new one.’
11. *mjɛʔ.hna-θouʔ-pəwa ʔəθiʔ tə-tʰɛ-go lɛ-dɛ.*
 face-to.wipe-towel NML.new one-CLF-OBJ to.exchange-NFUT
 ‘He changed the towel for a new one.’

A further piece of evidence for the syntactic independence of [aV] in [NaV] constructions comes from expressions where the verb is negated. As the verbal prefix slot is occupied by the nominal prefix *ʔə-* in [NaV], the negator *mə-* is blocked from this position. The alternative nominalizing strategy with suffixed *-ta/da*, commonly used to nominalize clauses, occurs in this context, as seen in (12).

12. *je ʔəʔè tə-lòun, mə-ʔè-da tə-lòun pè-ba.*
 water NML.be.cold one-CLF NEG-be.cold-NFUT.NML one-CLF to.give-IMPORT
 ‘Give me one bottle of cold water, one not cold.’

The parallel occurrence of *ʔəʔè* ‘cold one’ and *mə-ʔè-da* ‘not cold one’, both specifying *je* ‘water’, show that they are not syntactically speaking modifiers (i.e., dependents) of the nominal head, but heads themselves. The clausal nominalization suffix *-ta/da* is a merged form of the attributive suffix *-té/dé* and a dummy noun *ha* ‘thing’ (Yanson 2005:233f). This type of nominal expression never occurs as a modifier of another noun, unlike [aV], which in prenominal position is best analyzed as modifier of a following (but not of a preceding!) noun, as in *ʔəmjən-làn* ‘expressway’ seen above.

This leads to the conclusion that [aV] is a syntactically free element in a clause, rather than a bound part of a [NaV] construction. Semantically, it may be more or less closely bound to the N, or it may have adverbial or depictive function in the clause. Following this argumentation, we may say that [NaV] constructions are not NPs in which either N or [aV] is the head, the other the dependent, but rather that both are independent syntactic units. This means that, unlike in Lehman’s analysis presented above, [NaV] as in *ʔəiv ʔəci* ‘a/the big house’ should not be interpreted as ‘house-like bigness’, but rather as ‘house, a big one’. N and [aV] are thus seen as standing in an APPositionAL relation, rather than a MODIFIER-MODIFIED relation. This analysis is perfectly consistent with the wide range of meanings of [aV] derivations, as seen above in 2.2.2. It also explains the fact that the choice of classifier is triggered by N, rather than by an abstract [aV], as seen in examples (4) and (5) above. If *ʔəmjən* refers to the height of the bench, the appropriate quantifier complement is the measure word *pɛ* ‘foot’, but if the same [aV] expression functions as modifier to ‘bench’, the classifier *lòun* (for counting round objects, fruit, pieces of furniture, etc.) must be used with the numeral. The analysis of [NaV] as an appositional expression ‘X, a/the Y one’ also explains the possibility to drop N and have only [aV] without change in meaning. Furthermore, appositional expressions may become more tightly bound in some cases, resulting in compounds that behave like lexicalized items. The change may be gradual and affect only some, usually very frequent, collocations. This process can be seen in kinship terms combined with proper names, which may become lexicalized into a single intonation unit, as in English ‘my uncle, John’, becoming ‘my Uncle John’. In Burmese, the appositional expression *dó sʰəja pʰe wìn* ‘our teacher, Hpe Win’, with an intonational break between *sʰəja* and *pʰe wìn* may be compared to the more closely bound *sʰəja pʰe wìn* ‘Teacher Hpe Win’ with no intonational break. The latter is still less tightly

integrated phonologically than an [NV] expression like *s^həja-jì* ‘headmaster, big teacher’ (‘teacher - be.big’), as seen by the different voicing behavior.

To sum up the syntactic structure of [NaV] expressions, we propose that [aV] is a syntactically independent nominal form relating semantically in different ways with N (as attributive or depictive) or V (as adverbial). In attributive function in [NaV] expressions, it is formally an apposition to N, with which it may become more closely bound in some cases.

4.2 The syntax of [NV]

As seen in 2.2.3 above, [NV] expressions can have several functions, some with rather idiosyncratic semantics. Here we only consider [NV] noun phrases where V modifies N. Unlike [NaV], [NV] constructions do not show any overt morphosyntactic linker between the noun and the modifying verb. Semantically, as in [NaV], the expression is a noun, with the postnominal verb as a modifier. Therefore, the question of the internal structure arises, similarly to the [NaV] constructions. Again, if N is the head and V the modifying dependent, the structure violates the basic head-final rule of Burmese as seen in [N1N2] constructions. Unlike [NaV], though, [NV] can hardly be analyzed as symmetric appositional construction, as the two components are not of the same category and cannot be taken as coreferential. The only hint of syntactic boundness in [NV] expressions is the fact that V is generally voiced in these constructions, showing it to be in close “juncture” with N (Okell 1969:12). As mentioned above, voicing is not always a sure diagnostic for syntactic relations, as, firstly, it is not indicated in the orthography, and, secondly, there is possible regional and individual variation in many cases. On the other hand, standard dictionaries do indicate voicing in the pronunciation (e.g., Myanmar Language Commission 1976, San Lwin 2010), so at least for the [NV] expressions listed in the dictionaries, there is safe evidence for the role of voicing. The close morphosyntactic boundness indicated by voicing in [NV] expressions generally corresponds to their semantic closeness.

There are two possible ways to explain the internal structure of closely bound [NV]. First, one may take them as derived from fuller [NaV] expressions, with the nominalizer *ʔə-* dropped. We have seen above that this omission of *ʔə-* occurs frequently in the informal language. The development would then be from appositional [N]+[aV] to attributive [NaV] and compound [NV], the latest step commonly occurring in expressions with conventionalized meaning. If this path of development is correct, we might expect several in-between cases, that is, variation of [NaV] and [NV], the latter possibly with varying voicing.

The other possible explanation of [NV] attributive expressions is to take V as direct verbal modifier, indicating an inherent or permanent feature of N, which explains the close boundness. The position of adjectival modifiers with respect to their nominal head is variable in many languages across the world and does not necessarily correspond with the general headedness of the language, allowing for asymmetric orders (Dryer 2013). The different orders may have different historical reasons in different languages, and they may be indicative of different degrees of lexicalization.

Both analyses are plausible and with no obvious counterevidence, it is possible that both are adequate for different [NV] expressions, with not a single development leading to all cases in the present language. The available diachronic data suggests that [NV] is the earliest pattern of modified nouns, with [NaV] appearing only later.¹⁵ Together with the idiosyncratic semantic development of at least some [NV] expressions, this makes a strong case for [NV] actually being an old pattern in the language. The synchronic data from colloquial Burmese on the other hand show how the pattern can arise spontaneously in the present-day language. We may therefore conclude that two distinct paths of development merge in modern Burmese [NV] expressions, one from adjectival function of V in original [NV] compounds, the other from shortening of [NaV] to [NV] in non-formal speech. The question that remains to be answered is when and why [NV] and [NaV] are chosen by a speaker. This will be discussed in the next section.

¹⁵ The inscriptional data of Burmese must be taken with care, as only a very specific (high) genre is represented in writing until very recently. The absence of a construction from the data does not necessarily mean the construction did not exist in the language at the time.

5. Factors triggering the choice of [VattrN], [NV], and [NaV]

With three noun-modifier constructions that are treated as synonyms by most grammatical descriptions, the factors triggering the choice of one over the others are not obvious. Taking the premise for granted that difference in form entails difference in meaning/function, however subtle, we should be able to detect the differences in the semantics of the three forms from their distributional behavior.

The case of [VATTRN] seems to be the clearest, as this construction allows for all kinds of grammatical markers, making the modifying part syntactically and semantically more independent and transparent than in the other two types. [VATTRN] constructions are the most common and are adequate in situations where a specific referent is described, modified, or specified. Verbal modifiers can be added only in [VATTRN] constructions, as in *ʔəlun mjín-ðð taiʔ* ‘the/a very tall building’ (‘very be.high-ATTR building’) and *sèin-ðè-ðð ʔəjwɛʔ* ‘the/a still green leaf’ (‘be.green-yet-ATTR leaf’).

With the possibility of adding aspectual and other verbal modifiers, [VATTR] does not necessarily refer to permanent or inherent qualities of N but may rather express momentary attributes. In formal Burmese, [VATTRN] is the default construction, corresponding to both relative clauses and attributive adjectives in English. The N modified by [VATTR] may be either specific or non-specific, corresponding to English definite or indefinite article constructions, depending on the context. [VATTRN] may therefore be seen as default choice, which is theoretically always possible, but not conversationally the most adequate choice in some contexts.

The construction [VATTRN] is usually used in situations where the speaker wants to point out a specific quality of a specific referent or group of referents, often in direct opposition to another possible quality, as in (13).

13. *cì-dé* *ɲà* *mə-jàUN-bù.*
 be.big-ATTR fish NEG-to.sell-NEG
 ‘I don’t sell the big fish (but I do sell the small ones).’

A customer in the market chooses fish and points on the big one, on some reason the seller does not want to sell this particular fish. Using the construction [NV] *ɲà-ji* (‘fish-big’) in this situation would mean that the seller does not sell big fish in general, just small ones. Less clear is the difference with [NaV], which might be interpreted that the seller is talking about a specific big fish or the fish he sells in general.

In (14), the conventionalized expressions *lu-zò* ‘bad person, bandit, villain’ is used with its specific meaning, while in the second clause the more transparent [VATTRN] *kàUN-dé lu* ‘a/the man who is good’ appears instead of *lu-gàUN*, which is a lexicalized expression meaning a ‘dignified person’, rather than just ordinary good man. In this situation it would sound too elevated.

14. *lu-zò-dwe-né* *mə-sʰeʔ.sʰAN-né,* *kàUN-dé* *lu-dwe-né-bè.*
 man-bad-PL-with NEG-to.associate-PROH be.good-ATTR man-PL-with-EXCL
sʰeʔ.sʰAN *pó.*
 to.associate INSIST
 ‘Don’t associate with scoundrels, associate with good people.’

Similarly, in (15), the [VATTRN] construction is appropriate in a concrete context where the mother warns her child not to drink water that is hot at the moment, rather than referring to hot water in general which the child should avoid. The latter would be expressed by [NV] as *je-bu* (‘water-hot’), rather than [VATTRN].

15. *pu-dé* *je* *mə-θauʔ-né.*
 be.hot-ATTR water NEG-to.drink-PROH
 ‘Don’t drink (the) hot water.’

Sentence (16) is appropriate in a situation where a person has a choice of cars, but the one he usually drives is under repair, so he decides to drive the red one instead.

16. *di-lo s^ho-jin ŋa ni-dé kà-bè sì-mɛ.*
 PROX-as to.say-if 1s be.red-ATTR car-EXCL to.ride-FUT
 ‘In this case, I will take the red car.’

The construction [NV] *kà-ni* (‘car-be.red’) would be proper in the situation that a person is dreaming about a high post and the privileges that come with it, one being a ‘red car’ as a recognized symbol of prosperity. Not all possible [NV] or [NaV] collocations do occur in the language which also affects the choice of type of the construction to be used in concrete context. For example, the following constructions do not exist, **lu-mjan* (‘man-quick’) or **lu-zò* (‘man-early’) although they have typical [NV] structure and seem semantically possible. In both cases, only [VATTRN] is available to represent the meaning ‘the/a fast person’ and ‘the/an early comer’, respectively.

While the semantic difference between [VATTRN] and [NV] is clear in most cases, the same cannot be said for [VATTRN] and [NaV], nor for [NV] and [NaV]. In most examples above, [VATTRN] could be substituted by [NaV] without obvious change of meaning, though the former tends to sound more formal than the latter. Based on the distribution of [NV] and [NaV] seen so far, it is possible to provisionally hypothesize, that the constructions of the [NaV] type are used in situations, when the speaker considers the construction [NV] to be lexicalized, and as such, its components have lost their direct meaning, which the speaker wants to express. Compare the following expressions in (17a-b) and (18a-b).

- 17a. *θu ʔein ʔaci s^hauʔ-ne-dɛ.*
 3 house NML.be.big to.build-stay-NFUT
 ‘He is building a big house.’
- 17b. *θu ʔein-ji s^hauʔ-ne-dɛ.*
 3 house-big to.build-stay-NFUT
 ‘He is building a big house.’
- 18a. *kà-ʔəhàun kà-ʔəθiʔ jàun-wɛ-ʔin*
 car-NML.be.old car-NML.new to.sell-to.buy-NML
 ‘sale of used and new cars’
- 18b. *kà-hàun kà-θiʔ ʔəjàun-ʔəwɛ.*
 car-be.old car-new NML.to.sell-NML.to.buy
 ‘sale of used and new cars’

In both cases, the [NV] patterns (*ʔein-ji* and *kà-hàun/kà-θiʔ*) are not seen by the speaker as having idiosyncratic (lexicalized) meaning, so they are free to choose either form.¹⁶ A different situation can be seen in the following examples: *θà ʔaci* ‘the/an elder son’ (‘son-NML.be.big’), *θəmi ʔəŋɛ* ‘the/a younger daughter’ (‘daughter-NML.be.young/small’). The speaker uses the construction [NaV] instead of the seemingly natural [NV] for the obvious reason that the construction [NV] would be interpreted as ‘big son’ and ‘small daughter’, respectively, that is with a specific kinship meaning, rather than as a descriptive attribute. The latter are comparable to proper nouns, mainly used when addressing one’s children. This means that [NV] is blocked where there is an [NV] expression with a conventionalized meaning, but it is possibly available where no lexicalized [NV] exist in the speaker’s lexical inventory.

¹⁶ The choice of [NV] and [NaV] in examples 18a and 18b may be triggered by the form of the nominalizer of the activity verbs, avoiding non-parallel *ʔə*-nominalizations in the same expression (**kà-ʔəhàun kà-ʔəθiʔ əjàun-ʔəwɛ*).

Another relevant example is the difference between *kà-jì* and *kà ʔəci* ‘big car’, as well as between *kà-ŋɛ* and *kà ʔəŋɛ* ‘small car’. The forms *kà-jì* and *kà-ŋɛ* are used on signboards at toll gates in Myanmar, referring generally to ‘big cars’ (i.e., ‘trucks and buses’) and ‘small cars’ (i.e., personal cars), respectively, as distinct categories of vehicles. The [NaV] constructions are used when talking about big and small specimens of the general category ‘car’. Again, the compounds *kà-jì* and *kà-ŋɛ* have specific meaning and can be taken as lexicalized, although they do not necessarily appear in dictionaries.

The following pairs of expressions further illustrate the semantic difference between [NV] and [NaV].

[NV]	gloss	[NaV]	gloss	Components
<i>lu-jì</i>	‘adult’	<i>lu ʔəci</i>	‘big person’	<i>lu</i> ‘man’, <i>ci</i> ‘be big’
<i>lu-ŋɛ</i>	‘youth’	<i>lu ʔəŋɛ</i>	‘young person’	<i>lu</i> ‘man’, <i>ŋɛ</i> ‘be young/small’
<i>mjó-hàun</i>	‘ancient city’	<i>mjó ʔəhàun</i>	‘old city, former city’	<i>mjó</i> ‘city’, <i>hàun</i> ‘be old’
<i>je-nwè</i>	‘hot green tea’	<i>je ʔənwè</i>	‘hot/warm water’	<i>je</i> ‘water’, <i>nwè</i> ‘be hot, warm’
<i>lɛʔ-èiʔ</i>	‘copy’	<i>lɛʔ ʔəèiʔ</i>	‘new hand’	<i>èiʔ</i> (not <i>éiʔ</i>)
<i>lɛʔ-hàun</i>	‘original’	<i>lɛʔ ʔəhàun</i>	‘old hand’	<i>lɛʔ</i> ‘hand’, <i>hàun</i> ‘be old’
<i>hìn-jo</i>	‘soup, broth’	<i>hìn ʔəcʰo</i>	‘mild, sweet curry’	<i>hìn</i> ‘curry’, <i>cʰo</i> ‘be sweet’

The last example is especially telling, as it is not uncommon to hear orders in restaurants like *hìn-jo ʔəsaʔ* ‘hot, spicy soup’, where there is an apparent semantic clash between the opposites *cʰo* ‘be sweet’ and *saʔ* ‘be hot, spicy’.

From the examples given above, it appears that generally [NV] are lexicalized compounds with specific meaning, functioning in some cases like proper nouns (*θà-jì* ‘older son’), categories (*kà-jì* ‘big vehicles, trucks’), or have idiosyncratic meaning (*hìn-jo* ‘soup, broth’). In this, [NV] expressions are not fully productive, though there is individual variation in terms of what is acceptable as [NV]. [NaV] on the other hand is an open construction that can be used in a wide range of contexts when referring to a specimen with the described qualities. [NaV] is the colloquial counterpart of the formal [VATTRN] construction, which in colloquial style occurs mostly when the V is modified in some way. Having said this, we have to accept the fact that colloquially [NaV] expressions may become [NV] spontaneously by dropping the nominal prefix *ʔə-*, as happens frequently in informal speech and writing. This shortening of [NaV] to [NV] may or may not show voicing of V, depending on the speech style and speed, and the shortening may be blocked if there is a clearly lexicalized [NV] expression with conventionalized meaning and the speaker consciously wants to avoid ambiguity.

6 Conclusion and outlook

We can conclude that the three [NOUN-ATTRIBUTE] constructions investigated in this paper, while being almost synonyms, exhibit important differences in their syntax and contextual semantics. The most transparent type, [VATTRN], is the default construction, especially in the formal and literary language, to combine a noun with any kind of verbal attribute. [NV] constructions in the literary language are always lexicalized compounds with specific meaning. If the verbal attribute of a noun is not further modified or specified by aspectual or other markers, the colloquial style prefers the construction [NaV], where [aV] is syntactically a nominalized verb, which can have a wide range of meanings. The most common meaning in the constructions discussed in this paper can be labeled ‘the/a X one’, referring to a participant, rather than to the quality itself (as in ‘X-ness’), which stand in appositional relation to the N. In the colloquial language, [NaV] may be reduced to [NV], except in cases where there is a conventionalized [NV] compound.

The fact that formal and informal styles are not always clearly separate, and the styles may be mixed in many contexts leads to an overlap of [VATTRN] and [NaV] without obvious semantic difference between the two. On the other hand, the possible shortening of [NaV] to [NV] results in the two forms not being clearly distinguished in the colloquial style.

To further substantiate the analysis given in this paper, a larger corpus-based investigation covering different language styles should be made in the future. The actual use and distribution of the forms under discussion can only be verified by in-depth research in a wide range of textual genres, including spontaneous speech in informal settings.

References

- Matthew S. Dryer. 2013. Order of Adjective and Noun. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/87>, Accessed on 2020-06-22.)
- Haiman, John (ed.) 1985. *Iconicity in syntax*. Amsterdam/Philadelphia: John Benjamins.
- Himmelman, Niklaus P. and Eva Schultze-Berndt. 2005. Issues in the syntax and semantics of participant-oriented adjuncts: an introduction. In Himmelman, Niklaus P. and Eva Schultze-Berndt (eds.) *Secondary predication and adverbial modification. The typology of depictives*. Oxford: Oxford University Press, 1-67.
- Hopple, Paulette M. 2011. *The Structure of Nominalization in Burmese*. SIL e-books 22.
- Jenny, Mathias and San San Hnin Tun. 2016. *Burmese. A Comprehensive Grammar*. London: Routledge.
- Matisoff, James SA. 2003. *Handbook of proto-Tibeto-Burman: system and philosophy of Sino-Tibetan reconstruction*. Berkeley: University of California Press.
- Myanmar Language Commission. 1976. *Myanmar-English Dictionary* (6th edition). Yangon: Myanmar Language Commission.
- Myanmar Language Commission. 2005. *Myanmar Thadda* [Myanmar Grammar]. Yangon: Myanmar Language Commission.
- Nosova, N. I. 1974. Word and form shaping by prefix –a in Burmese. In Olderogge, Dmitrij A. (ed.) *Countries and peoples of the East*. Moscow: Nauka, 230-234. [in Russian]
- Okell, John F. 1969. *Reference Grammar of Colloquial Burmese*. London: Oxford University Press.
- Okell, John. 1994. *Burmese (Myanmar). An introduction to the literary style*. DeKalb: Northern Illinois University.
- Okell, John F. and Anna Allott. 2001. *Burmese/Myanmar dictionary of grammatical forms*. Richmond: Curzon.
- Pe Maung Tin, U. 1963. *Myanma thadda hniin athon-ahnon, taIn myiin*. [Myanmar grammar and usage, high level]. Rangoon,
- San Lwin. 2010. *Pyi Zone myanma-ingaleik abidan* [Pyi Zone Myanmar-English dictionary]. Yangon: Pyi Zone.
- Yanson, Rudolf. 2005. Tense in Burmese: a diachronic account. In Watkins, Justin (ed.) *Studies in Burmese linguistics*. Canberra: Pacific Linguistics, 221-240.
- Г. Ф. Мининой и У Чо Зо. 1976. Бирманско-русский словарь, под редакцией. [G. F. Minina and U Kyo Zaw (eds.) 1976. Burmese-Russian Dictionary. Moscow: Russian Language
- Тагунова И.М. 1971. Глаголы качества в функции определения к имени в бирманском языке, Языки Китая и Юго-Восточной Азии: проблемы синтаксиса. [Tagunova I. M. 1971. Stative verbs in noun-attribute function in Burmese. In Languages of China and Southeast Asia: problems of syntax, Moscow: Nauka]

ON THE USAGES AND FUNCTIONS OF THE CASE MARKER *-nīu* IN COLLOQUIAL LIANGMAI

Wichamdinbo MATAINA
CEL, Sikkim University
wicham.mataina@gmail.com

Kailadbou DAIMAI
CSLLC, Utkal University
kailadboudaimai@yahoo.co.in

Abstract

This paper describes the morphosyntactic functions of the case marker *-nīu* in Liangmai, a Tibeto-Burman language, spoken in the Northeast India. The agentive marker *-nīu* is used optionally with both the subjects of transitive and intransitive clauses especially when the subject is acting in volition. We attempt to explain the difference between when *-nīu* is omitted and when *-nīu* is present. It also serves as an update to the previous reports on Liangmai case marking in Charengna (2014:399), Mataina (2018:18) and Daimai (2018).

Keywords: Liangmai, case, agentive, volition
ISO 639-3 code: njn

1 Introduction

In this paper, we will describe the system of case marking associated with *-nīu* in Liangmai and specify its syntactic and pragmatic functions. Many Tibeto-Burman languages have been described to have a system of ‘optional’ or ‘pragmatic’ case marking (Coupe 2011, DeLancey 2011, LaPolla 1995, and Chelliah 1997). The case marker *-nīu* is also used optionally in Liangmai and it is found to mark both A and S arguments depending on pragmatic conditions. It is typically found on A(gent) arguments of transitive clauses and on the A arguments of a causative clauses formed with intransitive verbs (Daimai & Raguibou 2020:35). However, it is also reported to be omitted from a highly agentive A arguments and can even be used with non-agentive A/S(subject) arguments (Mataina 2018:19). The marker *-nīu* is usually absent in intransitive clauses. Earlier works in Liangmai case marking system by Charengna (2014), Mataina (2018) and Daimai (2018) have described *-nīu* as nominative, agentive and ergative respectively. The paper is an update to earlier reports on the Liangmai case-marking system, including our SEALS XXX 2021 video presentation, and an attempt to gain more understanding in the case marking system of Liangmai.

1.1. The People

Liangmai is one of the Naga tribes living in India’s northeastern states of Manipur and Nagaland. In Manipur, they live in Tamenglong, Senapati and Kangpokpi districts, and in Nagaland, they are found in the Peren district. The total population of the Liangmai as per the census report of India 2011 is 49,811. The Liangmai people were previously known by an incorrect ethnonym, and till recently, the term ‘Kacha Naga’ was used to refer to Liangmai and Zeme in Manipur. This misnomer was, however, officially done away in Manipur with the Constitution (Scheduled Tribes) Order (Amendment) Act in 2011. The Liangmai in Nagaland, however, are still known as the Zeliang, along with the Zeme. The Liangmai share ethnic, cultural and linguistic affinity with Zeme and Rongmei. These three tribes are commonly referred to as Zeliangrong, a term formed by putting together the first syllables of the three tribes. They are believed to be descendants of a common ancestor. There are numerous cognate root words shared among their languages as well. They live in a contiguous area spreading across Assam, Manipur and Nagaland. Bradley (1997) also used the term ‘Zeliangrong’ as a subgroup within the Southern Naga of the Kuki-Chin-Naga branch.

1.2. The Language

In Grierson's Linguistic Survey of India (1903), Liangmai, called Kwoireng or Liyang, was treated as belonging to the Naga-Kuki sub-group, though no description of the language was provided. More recently, Burling (2007) grouped Liangmai under his Zeme group along with Zeme and Rongmei.

Map 1. North East India maps showing Liangmai inhabited area (adapted from <https://www.mapsofindia.com/maps/northeast/sevensisters.htm>)



In our data elicitation, the basic constituent order found in pragmatically neutral declarative clauses is SOV. We rely on elicited examples because such examples tend to have all of the participants overtly realized, which is frequently not the case in natural discourse. In example (1), we find an intransitive construction with the third-person subject preceding the verb 'laugh'. In example (2), we see the first-person agent preceding the object: the verb is in the clause-final position. When there are two objects, the primary object precedes the secondary object followed by the verb, and the primary object is overtly marked. In example (3), the agent is marked with *-nūu*, and the primary object is overtly marked with the dative morpheme *-tū*.

1. *pǎ* *núi-bām-mei*
 1SG laugh-PROG-DECL
 'He is laughing'

2. *ī* *tsəlù* *tàt-néi*
 1SG field go-IRR
 'I will go to the field'

3. *naŋ-nīu* *ə-tū* *rāŋkáj* *pì-jēi*
 2SG-AGT 1SG-DAT money give-DECL
 ‘You give the money to me.’

Liangmai has a binary mood system contrasting a zero-marked realis with an overtly-marked irrealis mood. The unmarked realis mood corresponds to actualized events as in (3), while the irrealis suffixes *-néi* and *-rabo* is used to encode non-actualized events as in (2). The language is also marked for aspects, such as progressive as in (1) and perfective. As is commonly found in other Tibeto-Burman languages, the agentive marker *-nīu* in Liangmai has the same form as the instrumental marker as illustrated in (4).

4. *pǎ-nīu* *tətʰi-tù* *sìŋ-tāŋ-nīu* *dàp-jēi*
 2SG-AGT dog-DAT wood-CL-INSTR beat-DECL
 ‘He beat the dog with a stick’

1.3. Data sources

There are dialectal differences in Liangmai but not to an extent that they cannot understand varieties of one another. We observed that in terms of understanding each other between the southern Liangmai village varieties like Thalon and Namtiram against the Northern Liangmai village varieties like Kuilong and Maguilong, the difference accounts to only a small amount. These differences are basically in accents and some lexical words. Each Liangmai village has a different accent. For instance, in the lexical domain, the Thalon village calls ‘uncooked rice’ as [kə̀bàn], while the northern village variety uses the term [tsə̀bàn] for the same. There is no significant difference in the segmental phonology as per our observation. In this study, most of the sentence data were created by us as native speakers of Liangmai, and the spoken variety of these sentence data were based on the northern village varieties like Kuilong, Maguilong and Lemta and the southern varieties like Thalon. In addition, we also used sentences produced by Mataina (2018) and Daimai and Raguibou (2020) and cited them in the sentence examples wherever required.

2 Overview of case markers in Liangmai

We claim that there are ten case markers in Liangmai. Charengna (2014:398) reported nine case markers as reproduced in Table 2. Mataina (2018:18) reported ten case markers with a variation and one additional case marker called the terminative case from earlier reports, as listed in Table 1. The label or the terminology of case markers used in both the studies are partially different.

Table 1: Case markers in Liangmai (as reported in Mataina 2018:18)

Case name	Markers	Semantic function
Agentive	<i>niu</i>	agent; contrast
Accusative/Dative	<i>tu</i>	affected patient / recipient
Genitive	<i>gu</i>	possessor
Instrumental	<i>niu/ k^huiluziu</i>	instrument
Comitative	<i>nai</i>	accompany
Benefactive	<i>lej</i>	benefit
Locative	<i>ga</i>	location
Ablative	<i>gasu ganiu</i>	source, source of action
Allative	<i>lam</i>	movement towards a place
Terminative	<i>kətaŋ</i>	movement towards a time/place

Table 2: Case markers in Liangmai (as reported in Charengna 2014:398)

Case name	Markers
Nominative	<i>niu</i>
Accusative	<i>tu</i>
Instrumental	<i>niu</i>
Dative	<i>lej</i>
Locative	<i>gə-ləm</i>
Genitive	<i>gu</i>
Ablative	<i>gasu/gəniu-ləmsu</i>
Sociative	<i>nai</i>
Benefactive	<i>lej</i>

3 Usages of the marker *-nīu* in Liangmai

For the purpose of this paper, before we look at the role of the marker *-nīu*, it would be useful to look at the four patterns of core case marking as given by Dixon (1979, 1994) and cited in Coupe (2007). Also, in this section, we discussed the various terminologies used to label the case marker *-nīu* in earlier reports. The four patterns of core case marking that are common to the world languages are the following:

- **Nominative-Accusative** - The A argument and S arguments are marked in similar ways but differently from the O argument.
- **Ergative-absolutive** - The O argument of a transitive clause and the S argument of an intransitive clause have similar marking, while the A argument is marked differently.
- **Split-S system** - S arguments of ONE SET of intransitive verbs get a morphosyntactic treatment similar to the A argument of a transitive verb, while the S argument of ANOTHER SET of intransitive verbs gets a morphosyntactic treatment similar to O arguments of transitive clauses.
- **Fluid-S system** - Similar to split-S (i.e., having a set of intransitive verbs that marks S like A, and another set of intransitive verbs that marks S like O) except that in the fluid-S, the S argument gets case marking based on semantic factors.

The position of the Liangmai agentive case marker in relation to the above four types given by Dixon varies. It does not completely follow any one of the above patterns strictly, but it seems to be following partially all the above patterns, and this can be seen from the examples discussed in the following sections.

Charengna (2014) labelled *-nū* as a nominative-case marker. In Meitei, Bhat and Ningomba (1997) also labelled *nə* as a nominative case marker, but for the same language, Chelliah (1997) considers it to be an agentive marker. In Liangmai, Mataina (2018) labelled the same marker as an agentive marker, while Daimai (2018) labelled the same marker an ergative marker. Following LaPolla (1995), Coupe (2007, 2011) and Teo (2012), we use the label ‘agentive’ in this paper. LaPolla uses the term agentive to refer to a subject argument that acts willfully or is done in volition, but he uses the term ergative when transitive marking is done systematically based on the syntactic pattern. Also, we do not label the marker *-nū* here as a nominative case because, as a nominative marker, it does not mark subject experiencer of A or S arguments. The nominative case is therefore not morphologically marked in Liangmai. As stated earlier, it may be noted that the agentive morpheme *-nū* is homophonous with an instrumental case marker in Liangmai.

3.1 Commendation and accountability

The marker *-nū* in Liangmai is used to essentially recognize the ‘doer’ as such to give a person a credit or a blame. Otherwise, in a pragmatically neutral situation, the marking of *-nū* is optional (Mataina 2018:21) and Daimai (2018), as shown in (5).

5. *nātsī-pīu* ∅ *mərūi kāmsāt-lù k'āi bām-mēi, nāŋ rū-k'āi-lāu*
 2SG.POSS-brother-GEN-AGT hen kill-PERF keep PROG-DECL, 2SG burn-PO-IMP
 ‘Your brother had killed a chicken and kept, you burn (it). (lit. Clean it by burning to cook).

However, the agentive marker is mandatory in a pragmatic situation such as in a response to a question ‘Who did it?’. This required usage is reported in Ao (Coupe, 2007). This is the context in which the agentive marker is obligatory. Example (6) is an interrogative sentence and (7) is an answer to the question of (6). In a context other than the content question, a speaker would insert the marker *-nū* if he or she wants to intentionally indicate the doer (to give a credit or a blame).

6. *sāu (-nū)* *mərūi kámsát lau*
 who (-AGT) chicken kill Q
 ‘Who killed the chicken?’
7. *lúŋt^hənbāu-nū¹* *mərūi kāmsāt-le.*
 Lungthonbou-AGT chicken kill-DECL
 ‘Lungthonbou killed the chicken (yesterday).’

¹ The tone of the agentive marker is generally influenced by the tone of the noun it marks (Mataina 2018)

In (6), it is normal to omit the agentive marker by the questioner. That does not necessarily but generally mean the questioner is not worried about who prepared the chicken for the family. However, inserting the agentive marker in such a context as (6) definitely implies that the questioner wants to identify who did the work. In a pragmatically neutral situation, the agentive is not required at all. This is demonstrated in a habitual statement in (8) and (9).

8. *i nái mǎtsǎŋ kǎbàk-tù gūi-jēi*
 1S day-alway pig-ACC feed-DECL
 ‘I feed the pig every day.’
9. *ǎ-kìk’ün-léŋ t̃ŋkúm mǎtsǎŋ pastor-píu kǎk’éŋ thū-jēi*
 1S-POSS-family-BEN year always pastor-male prayer do-DECL
 ‘The pastor prays for our family every year.’

Coupe (2007:159) has shown a context of how an agentive marker was not required in Mongsen Ao, a Naga language, even when a questioner was inquiring what Coupe had done with the tea that had been made for him (as the tea was no more). Upon replying with the Ao example (10), Coupe was corrected by the questioner (his Ao language consultant) that it was wrong to use the agentive marker in that context. This is similarly true with Liangmai. However, it is also not wrong to use the agentive marker in such context in the case of Liangmai and can be optional, as in (11). Yet, our understanding is that the agentive marker *-nū* in Liangmai in particular and North East Indian languages like Mongsen Ao (Coupe, 2007) and Sumi (Teo, 2014) in general is used to indicate that the doer is pointed out clearly.

10. *nì nə səŋa t̃fəmu-kū (Mongsen Ao)*
 nì nə səŋa t̃fəmu-ukū
 1SG AGT tea drink-ANT
 ‘I’ve drunk the tea.’ (Coupe 2007:159)
11. *ī-(nū) tsá sàk-lu-dei*
 1SG (AGT) tea drink-PST PERF-DECL
 ‘I’ve drunk the tea.’

3.2. Volition and power

In earlier studies on North East Indian languages, among others, Bhat and Ningomba (1997) in Meetei, and Teo (2014) in Sumi reported that the agentive marker is used to indicate volition. We observe that in Liangmai as well the agentive marker is also used to indicate volition, and in addition, an ability and power. However, there are some exceptions to it which we will see in a moment. For an intransitive clause as in (12), the agentive marker is not used, but for the same example in (13), it is used to show the intention of the speaker. He intentionally coughs to get an attention or to distract someone.

12. *i mǎkhū-jēi*
 1SG cough-DECL
 ‘I coughed.’
13. *i-nū mǎkhū-jēi*
 1SG-AGT cough-DECL
 ‘I coughed (intentionally).’

The agentive marker *-nū* is also used to indicate power or ability. It occurs, for instance, in a situation where no one is willing to perform certain tasks due to its challenging or dangerous nature, or a kind of work that people are too lazy to perform. Examples are shown in (14) through (18).

14. *ə-nà-nīu* *sefai* *māi-tù* *ə-lām-nēi*
 1SG-POSS-son-AGT army man-ACC contact-IRR
 ‘My son will contact the army.’
15. *i-nīu* *tsəkóŋ-tsepian* *k^hiu-néi*
 1SG-AGT plate-cup wash-IRR
 ‘I will do the dishes.’
16. *zīsú-nīu* *sŋtsǎi-gá* *sái-k^hái-jēi*
 Jesus-AGT cross-PP die-POL-DECL
 ‘Jesus died on the cross.’
17. *kək^hiu* *māi-nīu* *kám-ra-bau* *gū-jēi*
 rich man-AGT do-PURP-NOM POSS-DECL
 ‘This is (something) supposed to be done by a wealthy person.’
18. *nə-pǎu* *kəríŋ-bāu* *tīŋ-hāi*, *pǎ-nīu* *tsəkúi* *mīn-lū-jēi*
 2SG grandpa PRE-alive-NOM time-DEF 2SG PRE-lion catch-PST-PERF-DECL
 ‘While your grandpa was alive, he (even) caught a lion.’

However, the claim that the marker *-nīu* in Liangmai is used to indicate volition and ability is complicated by (19). In (19), the subject experiencer is not indicating any volition to be sweating, but the agentive marker *-nīu* is still used. Also, in (19), the agentive marker can be optional, but it is more natural to mark the subject. However, this must be taken as an exception because this type of occurrence is very rare. Chelliah (1997:124) also pointed out that in Meetei, the agentive marker *nə* is easily found in other sentences that describe unintentional situations, which goes against Bhat and Ningomba’s (1997) claim that *nə* is used to mark a volition .

19. *hāiga* *ī-nīu* *kəsōŋziu* *sáira* *kúm* *bām-bau-gā*, *dekám* *tsəmì* *t^hūziu* *lau?*
hāi-ga *ī-nīu* *kəsōŋ-ziu* *sái-ra*
 PROX-LOC 1SG-AGT hot-COPM die-PURP
- kúm-bām-bau-gā* *dekám* *tsə-mì* *t^hū-ziu* *lau?*
 like-PROG-NOM-CONJ why NRL-fire make-CPM-QPTCL
 ‘Here, I am about to die due to heat and why are (you) making fire?’ (Mataina 2018)

3.3. Subject experiencer

An agentive marker cannot be present when the A occurs as a subject experiencer and has no control over the action (Mataina 2018:19). In examples (20) to (22), all the actions are caused by a natural phenomenon or by an accident.

20. *ə-wán* *thiu-jei*
 1S-stomach pain-DECL
 ‘My stomach aches.’
21. *luŋthónbāu* *tsə-khāu-gā* *kəlūi* *wāŋ-mí-dei*
 PN NRL-cleft-LOC roll go-PERF-DECL
 ‘Lunthonbou has fallen on the cleft.’ (Literally- rolled and went)

22. *tə-düi* *zāu-bām-mei*
 NRL-water drip-PROG-DECL
 ‘Water is dripping (from the roof).’

In contrast, the agentive can be optional as well in a subject experiencer as shown in (23) in which the S argument is a subject experiencer. However, in this kind of construction, the speaker intends to point out the reason why someone is upset. Otherwise, in a normal context, the agentive *-nūu* will not be required, as shown in (23).

23. *pə -nūu* *nə-tū* *pə-tsūn* *sà-bāu* *rà,*
 3S-AGT 2-ACC 3SG-heart bad-NOM DEF

nāŋ *mālīu* *mák-bau* *dōnnū-jēi*
 2S talk NEG-NOM reason-DECL
 ‘The reason he’s upset with you is because you are silent.’

24. *pə* *nə-tū* *pə-tsūn* *sà* *bām-mei*
 3S-AGT 2-ACC 3SG-heart bad PROG-DECL
 ‘He is upset with you.’

There are certain types of intransitive verbs (i.e., those with S arguments-subject experiencer) that will not allow the agentive marker. Such verbs are *kək^hún* ‘shiver’, *ŋk^həŋ* ‘to fall’, *kàp* ‘to cry’, and *núi* ‘to laugh’ as shown in (25). No specific reason is known as to why these verbs do not allow the agentive marker. One of the reasons could be that these verbs cause the subject argument to experience the action of the outside agent. However, if any of the actions created by such *intransitive verbs* can cause distraction or trouble to others, the agentive marker may be inserted. In (26), the questioner was disturbed by the noise of a cry the previous night, so he asked someone a content question as in (26). Someone responded that it was *lúnthóŋbau* who cried the previous night. Here, *lúnthóŋbau*, the subject experiencer does not have a fault of his own in disturbing someone the previous night, but in order to get the blame or indicate the doer, the agentive marker *-nūu* is used. This is related to the discussion noted in §3.2, particularly in the response statement to a query like ‘who did it?’.

25. *lúnthóŋbau* *kàp-jēi*
 Lungthonbou cry-DECL
 ‘Lungthonbou cries.’
26. *ī* *nəlū-tù* *kàp-tū-lāu* *tūu* *dín-nēi,* *sāu-nūu* *kālútsún* *kāp-zū-lāu*
 1S 2P-ACC cry-NEG-IMP CONJ tell-DECL who-AGT last night cry-COMP-Q
 ‘I told you (all) to not cry, who cried last night? (it disturbed me).’
27. *lúnthóŋbau-nūu* *kàp-jēi*
 PN-AGT cry-DECL
 ‘Lungthonbou cried (last night).’

3.4 Causative construction

Additionally, marking A arguments looks more natural in causative sentences, but they can be still optional as shown in (28) and (29) where (28) is more natural. More examples on causative sentences are given in (30) and (31).

28. *ape-nūu* *kātsa* *kám-thiú-e*
 grandma-AGT tea CAUS-hot-DECL
 ‘Grandma made the tea hot/warmed the tea.’ (Daimai and Raguibou 2020:38)
29. *ape-(nūu)* *kātsa* *kám-thiú-jei*
 grandma-AGT tea CAUS-hot-DECL
 ‘Grandma made the tea hot/warmed the tea.’
30. *i-nūu* *lūṅsiliu-tu`* *kám-núu-jei* (transitive, causative clause)
 1SG-(AGT) Lungsilu-ACC CAUS-laugh-DECL
 ‘I made Lungsilu laugh.’
31. *i-nūu* *lūṅsiliu-tu`* *pí-tsə̀gàn* *laṅ-jēi* (ditransitive, causative clause)
 1SG AGT PN-ACC CAUS-curry cook-DECL
 ‘I make Lungsilu cook the curry.’

3.5 Inanimate marking

The marker *-nūu* also marks inanimate subjects. Such constructions with the agentive marker, as shown in (32) and (33), appear more natural against the unmarked inanimate subject.

32. *nàimík-nūu* *kə̀bún* *kám-núṅ-midēi*
 sun-AGT ice CAUS-melt-PERF-DECL
 ‘The sun caused the ice to melt.’
33. *kāu wāṅ-bāu* *sṅ-bāṅ-nūu* *pə̀-tū* *dāp-jēi*
 fall come-NOM wood-CL AGT 1S-ACC hit-DECL
 ‘The fallen tree has hit him’.

3.6 Contrast or comparison

The marker *-nūu* is also used to indicate contrast or comparison, as shown in (34) and (35). In addition, the usage or non-usage of the agentive marker *-nūu* seems to be suggesting a state of difference as shown in (36) and (37). The semantic difference between the two examples (36) and (37) cannot be distinguished easily by a native speaker. However, it seems to be indicating a difference where, (36) is in general warning someone to not venture out into the bushes because a snake or possibly any insect may bite him, while (37) is hinting that there is no other insect to harm him except the snake.

34. *hài-si-nūu* *rāṅkāj* *mədái*, *wúi-si-nūu* *rāṅkāj* *məṅjú*
 PROX-DET-CONTR money four PROX-DET-CONTR money five
 ‘This is four rupees, that is five rupees.’ (Literally- this costs Rs.4 and that costs Rs. 5)

35. *luŋthonbau-nīu pə̀-lùŋ kə̀wī-pūi nā, əsemliu-nīu pə̀-lùŋ*
 PN-CONTR 3SG-heart good-(GEN), child PN-CONTR 3SG-heart
kəsā-pūi nā-jei
 bad-(GEN) child-DECL
 ‘Lunthonbou is the child of the compassionate mother, and Asemliu is the child of the cruel mother.’ (Mataina 2018).
36. *mə̀nəm hāŋ-gā ŋət-tū-lāu, kənīu-Ø mə̀tsāu-né*
 bushes under-LOC enter-NEG-IMP snake-(AGT) bite-IRR
 ‘Do not enter into the bushes, the snake will bite (you).’
37. *mə̀nəm hāŋ-gā ŋət-tū-lāu, kənīu-nīu mə̀tsāu-né*
 bushes under-LOC enter-NEG-IMP snake-AGT bite-IRR
 ‘Do not enter into the bushes, the snake will bite (you).’

4 Conclusion

In the foregoing discussion, we attempted to identify the role of the case marker *-nīu* in Liangmai. The agentive marker in Liangmai is generally optional. It is basically used to indicate the doer. The mandatory marking happens only in pragmatic contexts especially a content question. This response statement to a content question can be either a transitive or an intransitive clause. It can be either a past or present or future statement. It also used to indicate a volition or a power. However, there is an exception to this namely, indicating volition and power as seen in (18). But this occurrence is very infrequent, and therefore we consider it as an exception. However, further investigation on the usage of the agentive marker for volition will be an important future direction.

In addition, the agentive marker *-nīu* is also used to indicate a contrast or a comparison. In regard to a subject experiencer (i.e., when an A argument is a subject experiencer), there are two contexts where agentive marking is totally absent and also where agentive marking can be optional. In a causativized construction, agentive marking appears more natural, but again it can be optional. The highly optional nature of the agentive marker or its multi-functional role makes it difficult to make its definite role besides what have been discussed in this paper. This paper will hopefully make some contribution to our endeavor to understand and study the core case marking system in the Trans-Himalayan languages.

Abbreviations

COMP	complimentizer
DEF	definite marker
POL	polite maker
PURP	purposive marker

References

- Bhat, D. N. Shankara and M. S. Ningomba. 1997. *Manipuri grammar*. München, Lincom Europa.
- Bradley, David. 1997. Tibeto-Burman languages and classification. *Papers in Southeast Asian Linguistics 14: Tibeto-Burman languages of the Himalayas*, ed. by David Bradley, 1-72. Canberra: Australian National University.
- Burling, Robbins. 2003. The Tibeto-Burman Languages of Northeastern India. *Sino-Tibetan Languages*, ed. by G. Thurgood and R J Lapolla, 169-192. London/ New York: Routledge.
- Chelliah, Shobhana L. 1997. *A Grammar of Meithei*. Mouton de Gruyter.
- Chelliah, Shobhana L and Hyslop Gwendolyn. 2011. Introduction to special issue on optional case marking in Tibeto-Burman. *Linguistics of Tibeto-Burman Area* 34.2
- Coupe, Alexander R. 2007. *A grammar of Mongsen Ao*. Berlin, Mouton de Gruyter.

- Coupe, Alexander R. 2011. On core case marking patterns of two Tibeto-Burman languages of Nagaland. *Linguistics of Tibeto-Burman Area* 34.2:21-47.
- Charengna, Widinibou. (2014). Case markers in Liangmai. *Language in India* 14.7:396–407.
<http://www.languageinindia.com/july2014/liangmaicasemarkers.html>
- Daimai, Kailadbou. 2018. Ergativity in Liangmai. *10th International Conference of the North East Indian Linguistics Society*. Assam University, Silchar.
- Daimai, Kailadbou & I.D. Raguibou. 2020. Causatives in Liangmai. *Himalayan Linguistics* 19.2:30-43. <http://dx.doi.org/10.5070/H91151000> Retrieved from
<https://escholarship.org/uc/item/0td2g3v4>
- DeLancey, Scott. 2011. Optional Ergativity in Tibeto-Burman Languages. *Linguistics of Tibeto-Burman Area* 34:2.
- Grierson, George A. 1903. *Linguistic Survey of India, Tibeto-Burman family, Bodo-Naga and Kachin Groups*. Vol.III, Part II. Calcutta (reprint Delhi: Low Price Publications; 2005).
- LaPolla, Randy J. 1995. ‘Ergative’ marking in Tibeto-Burman languages. *New Horizons in Tibeto-Burman Syntax* (ed) by Yoshio Nishi, James A. Matisoff and Yasuhiko Nagano. Osaka: National Museum of Ethnology.
- Mataina, Wichamdinbo. 2018. Case marking in Liangmai. *North East Indian Linguistics (NEIL)* 8, 17-31. Canberra, Australian National University: Asia-Pacific Linguistics Open Access. ISBN: 978-1-922185-41-9.
- Teo, Amos, 2012. Sumi agentive and topic markers: no and ye. *Linguistics Area of Tibeto-Burman Languages* 35.1:49-74.

studies of verb forms in Kapampangan (Bergaño 1916; Mirikitani 1971; Del Corro 1980) have similar positions, although they do not use the same term for the process. According to Mirikitani, the choice between <in> or the so-called vowel changes/difference in vowel quality depends on the phonological shape of the stem (1971:66), and Del Corro (1980:37-68) discussed in detail which affixes/processes are used in specific environments. Del Corro refers to the perfective morphemes in (2) and (3) as *replacives*.

(4) [u] → [i]
mugse ‘will throw’ → *migse* ‘threw’ (Del Corro 1980:40)

[a] → [ɛ]
danuman ‘will water’ → *de:numan* ‘have watered’ (Del Corro 1980:57)

(4) shows an example of the perfective replacives Del Corro mentioned. Similar to Forman’s description, replacives replace the V1 of the stem. For stems with [u] as its V1, the <i> perfective replacive is used, while <ɛ> is generally used for stems with [a] as V1. Aside from the transformations mentioned above, there are also instances wherein <i> replaces [a].

(5) *maldug* ‘is dropping’ → *mildug* ‘dropped’ (Del Corro 1980:43)

Bergaño also made the same observation but without identifying which vowel is used in replacing a specific vowel. (6) shows some examples of the conjugation of the perfective aspect found in Bergaño (1916:19, 44).

(6) *musn̄gi* ‘to open something’ → *misn̄gi* ‘have opened something’
isalicut ‘to hide (something/somewhere)’ → *selicut* ‘have hidden (something/somewhere)’

For stems that have [i] as V1, the infix <in> may be used (Mirikitani 1971:67). An example is given in (7). In some stems, [i] is simply lengthened (Mirikitani 1971:66; Del Corro 1990:51) as in (8).

(7) *diligen* ‘to water something’ → *dinilig* ‘to water something’ (Mirikitani 1971:67)

(8) *misakab* ‘fall down/trip’ → *mísakab* ‘fell down/tripped’ (Del Corro 1980:51)

To summarize, Kapampangan uses several strategies to mark the perfective aspect. This can be found in (9). The specific environment wherein these affixes are used can be found in Del Corro (1980:37-66).

- (9)
- a. infix <in>
 - b. replacives <i, ɛ>
 - c. lengthening of [i]

The analyses discussed above seem at odds with the fact that Philippine languages usually mark the perfective aspect through infixation. In this paper, we will present an alternative analysis for the perfective aspect marking specifically in (9b) and (9c), wherein the aforementioned replacives and vowel lengthening are unified into a single infix <i>, taking into account known sound changes in Kapampangan, namely, diphthong reduction.

2 Diphthong Reduction

Diphthong reduction, also known as monophthongization (Del Corro 1980, 1988), is a phonological process wherein diphthongs in Kapampangan occurring in a closed syllable or word-final position are simplified (Himes 2012:499). Himes traces the changes from Proto-Malayo-Polynesian (PMP) to Proto-Central Luzon (PLuzC) to Pre-Kapampangan (Pre-KPM).

- (10) *aj → [ɛ:]
 PLuzC *ʔeRbun → Pre-KPM *aybun → [ɛ:bun]
 PMP *beŋi → Pre-KPM *bayŋi → [bɛ:ŋi]
 PMP *baláy → Pre-KPM *baláy → [balɛ]
 (Himes 2012:499)
- (11) *uj and *iw → [iʔ]
 PMP *ikuR → Pre-KPM *i:kuy → [i:kiʔ] ‘tail’
 PMP *qapuR → Pre-KPM *a:puy → [ʔa:piʔ] ‘lime’
 PMP *sáliw → Pre-KPM *sa:liw → [sa:liʔ] ‘to buy’
 (Himes 2012:500)

(10) shows the change Pre-KPM *aj → [ɛ], while (11) shows the change Pre-KPM *uj, *iw → [iʔ]. The diphthongs *uj and *iw have the same pair of vowels ([u] and [i]), albeit in different sequences. This may suggest that the vowel pairs and not the vowel sequences determine the resulting forms. Although not directly relevant to the current study, the diphthong [aw] also underwent reduction to [ɔ] in Kapampangan (Himes 2012:500; Del Corro 1980:25), as shown in (12).

- (12) *aw → [ɔ]
 PMP *babaw → Pre-KPM *ba:baw → [ba:bɔ] ‘above’
 PMP *lá:ŋaw → Pre-KPM *la:ŋaw → [la:ŋɔ] ‘housefly’
 (Himes 2012:500)

Some of the Pre-Kapampangan forms Himes cited are actually attested in earlier documentation of Kapampangan such as those recorded by Bergaño (1860). (13) provides a few examples.

- (13) Pre-KPM *baláy ‘house’ cf. Bergaño *balay* ‘house’
 Pre-KPM *ba:baw ‘above’ cf. Bergaño *babao* ‘above’
 Pre-KPM *la:ŋaw cf. Bergaño *lanḡao*, ‘housefly’

Other words bearing word-final diphthongs not mentioned by Himes (2012) and found in Bergaño are *sablai* ‘to hang clothes’, *lacao* ‘remove,’ and *tacao* ‘gluttony’.

Although non-word-final diphthongs are unattested in Kapampangan, there is reason to believe that diphthongs in words like *aybun, *bayŋi (Himes 2012:499) did exist at some point. As evidence, words like *lelay* ‘edge’ and *tetay* ‘bridge’, which are found in Bergaño (1860), have cognates in other Philippine languages that possess medial diphthongs (cf. Tagalog and Botolan Sambal *laylay* ‘hem, edge’ and *taytay* ‘bridge’). This points to the possibility that diphthong reduction in Kapampangan came in two waves. The word-medial diphthongs were likely the first ones to undergo such change, followed by the second wave of diphthong reduction in word-final diphthongs. This initial wave may have been completed by the time Bergaño (1860) started his study of Kapampangan.

Meanwhile, the second wave of diphthong reduction affecting word-final diphthongs was ongoing during Bergaño’s study. In Bergaño (1860), citation forms are spelled with diphthongs, and the pronunciation guides that follow the lemmas also mark them as such “(Dipt.)”. Some of the entries also provide sample sentences, and they display the utterance-middle diphthong reduction.

- (14) (a) *BALAY* (Dipt) Noun. *House* ... *pibalebalay*, those that are presented in *pamibalebalay*, every dowry, e.g., cattle ... (Bergaño 1860:38)
- (b) *LACAO*. (Dipt.) Irregular verb. ... *Mamacao, mecao, macao*, to go. Mi, N. *Milaco ya ngeni* (Bergaño 1860:124)

(c) *TETAY* (Dipt.) Noun. Bridge. ... *tinetayan*, the bridge in which one passes, or the river in which does so as well (Bergaño 1860:252)

(d) *TACAO* (dipt.) Noun. Inappropriate desire to eat and drink; gluttony ...

The sample sentences from each entry are given in (15); (15E) however, is taken from Bergaño (1916).

- (15) (a) *Yian pibalebale co* (Bergaño 1860:38)
 yan pi-balay~balay=ku
 ABS.MED PL-REDUP-house=1SG.GEN
 ‘these are my dowry/trousseau’
- (b) *Milaco ya ngeni* (Bergaño 1860:124)
 mi-lakaw=ya ngeni
 INTR-going=3SG.ABS now
 ‘he left today’
- (c) *Mirugtungne tetay* (Bergaño 1860:252)
 mi-dugtuŋ=na=ya tetay
 INTR-linking=3SG.GEN=3SG.ABS bridge
 (metaph.) ‘he has come’
- (d) *Mirugtungne ing tetena* (Bergaño 1860:252)
 mi-dugtuŋ=na=ya iŋ=tetay=na
 INTR-linking=3SG.GEN=3SG.ABS SG.ABS=bridge=3SG.GEN
 (metaph.) ‘he has come’
- (e) *Dacpan me ing mapanacao* (Bergaño 1916:52)
 dakap-an=mu=ya iŋ=mapanakaw
 apprehending-TR=2SG.GEN=3SG.ABS SG.ABS=thief
 ‘catch the thief’

The appended forms of the roots in the sentences in (15b) and (15d) are found in medial position: they are followed by clitic pronouns. Notice, however, that the citation forms and other derivations and inflections have diphthongs, as in each entry: *pibalaybalay* ‘dowry, trousseau’, *panlacao* ‘was removed’, and *tinetayan* ‘bridge that was crossed’. Utterance-final diphthongs also remain as is. This is shown in sentences like (15c) and (15e). When they are found in the medial position, they are reduced. This is also discussed in Bergaño (1916).

“... when they [diphthongs] come in the middle, the vowel [pairs] are not clearly pronounced; rather, they merge, for example *pamanlacao*: one does not say *pamanlacao mo*, but *pamanlaco mo*, neither does one say *balay mo*, but rather *bale mo*; hence, *ay* sounds more like *e* than *a*, like *palay*, *pale mo*, and *ao* sounds more like *o*, like *galao*, *pamangalo mo*.” (Bergaño 1916:3). Translated from original Spanish)

The unreduced forms can still be seen in the sentence *Manlacao cayo* ‘you all go’ (1860), and *Ala con petay a tauo* ‘I did not kill anyone’ (1916).

While Himes (2012) and Del Corro (1980) noted that diphthong reduction only occurs on *aj and *aw, evidence suggests that this simplification process can also occur in [ia] and [ua] clusters, similar to how *uj and *iw were both reduced to [iʔ]. Examples (16) and (17) illustrate this.

(16) *paburian* ‘to disregard’ (Bergaño 1860:66) → *paburen*

(17) *manwas* ‘to wash’ (Del Corro 1988:17) → *manəs*

The form *manwas* is seen in some varieties of Kapampangan like Calaguiman-Mabatang. This particular variety has retained word-final diphthongs possibly due to the community’s relative isolation from other Kapampangan-speaking communities (Cruz, de la Rosa, Pelagio & Quizon 2020).

3 Kapampangan Perfective Infix <i>

In this section, we present an alternative to the past analyses of the seemingly complex and unpredictable formation of perfectives through replacives and vowel lengthening: these morphemes are products of phonological changes brought about by the infixation of another perfective affix <i>, which then underwent diphthong reduction as discussed in the previous section.

3.1. Stems with [u] as V1

When stems whose V1 is [u] (e.g., *kutang* ‘question’) is inflected with the perfective infix <i>, the diphthong formed is reduced to [iʔ] and further to [i] or [i:]. This process is illustrated in (18).

(18) [iu] → [i:]
kutaj → *kiutaj → *kiʔtaj → *ki:taj* ‘asked (a question)’
lukas → *liukas → *liʔkas → *li:kas* ‘took off (clothes)’

Del Corro (1980:42) noted that utterance-medial [ʔ], and in effect word-medial ones, are dropped and replaced instead with compensatory lengthening. (18) illustrates this point: the loss of the stop caused the lengthening of the preceding vowel [i]. This typically occurs in stems whose first syllable is open. For stems with a closed first syllable (i.e., CVC syllable structure), compensatory lengthening is suppressed. Examples are shown in (19).

(19) [iu] → [i]
buklat → *biuklat → *biʔklat → *biklat* ‘opened’
kumbiran → *kiumbiran → *kiʔmbiran → *kimbiran* ‘invited’

3.2 Stems with [i] as V1

For stems with [i] as V1, the perfective aspect is marked by the lengthening of [i] (Mirikitani 1971: 66; Del Corro 1990:51). This can also be explained by the infixation of <i> as is illustrated in (19).

- (20) [ii] → [i:]
miras → *miiras → *mi:ras* ‘arrived’
misabi → *miisabi → *mi:sabi* ‘talked’
minum → *miinum → *mi:num* ‘drank’
mipunta → *miipunta → *mi:punta* ‘went’

As we can see in (20), the lengthening of [i] is due to the juxtaposition of two consecutive [i] resulting from the infixation of <i>. The sequence of two [i] is then expressed as vowel lengthening or compensatory lengthening.

However, there are instances wherein the infixation of <i> does not produce length. This usually occurs in closed syllables similar to instances in (19).

- (21) *midlip* → *miidlip → *mi:dlip → *midlip* ‘napped’

3.3 Stems with [a] as V1

As previously illustrated in (13), the cluster [ia] produces the same change as the diphthong *aj in Himes (2012). The infixation of <i> triggers diphthong reduction in stems whose V1 is [a] and thus changes the V1 of the stem to [ɛ] in marking the perfective aspect.

- (22) [ia] → [ɛ:]
mate → *miate → *mɛ:te* ‘died’
paʔulyan → *piaʔulyan → *pɛ:ʔulyan* ‘sent home’
manabu → *mianabu → *mɛ:nabu* ‘fell’
mangan → *miangan → *mɛ:ngan* ‘ate’

Reduction to [ɛ:] usually occurs in stems with open first syllables. The vowel length in (22) is possibly caused by compensatory lengthening. However, in closed syllables, [ia] is reduced to [i], and compensatory lengthening is suppressed, as in (19) and (21).

- (23) [ia] → [i]
lakwan → *liakwan → *lɛ:kwan → *likwan* ‘left’
damdam → *diamdam → *dɛ:mdam → *dimdam* ‘heard’
makmul → *miakmul → *mɛ:kmul → *mikmul* ‘swallowed’
ʔakit → *ʔiakit → *ʔɛ:kit → *ikit* ‘saw’

In stems with the prefix *pag-*, the resultant monophthongs are either [i] or [ɛ]. In [ai] → [ɛ], notice that compensatory lengthening is suppressed as well.

- (24) [ia] → [i], [ɛ]
magobra (m-(p)ag-obra) → *miagobra → *mɛ:gobra → *mɛgɔbra*, *migɔbra* ‘worked’
magkwentu (m-(p)ag-kwentu) → *miagkwentu → *mɛ:gkwentu → *mɛgkwentu* ‘related’
pagsikapan → *piagsikapan → *pɛ:gsikapan → *pɛgsikapan* ‘worked hard on’
pagmulalan → *piagmulalan → *pɛ:gmulalan → *pɛgmulalan → *pigmulalan* ‘wondered about’
magaluk → *miagaluk → *migaluk*, *mɛ:galuk* ‘offeedr’
maglipat → *miaglipat → *miglipat*, *mɛ:glipat* ‘transferred’
magdala → *miagdala → *migdala*, *mɛ:gdala* ‘brought’

Alternations between [i] and [ɛ] can be observed in resulting perfective inflection from stems with /ia/. Forman (2019) noted this alternation between [i] and [ɛ], due to the recent changes in the language, has caused the semi-contrastive status of some vowel pairs. It was also observed that [i] and [ɛ] are less

clearly contrastive. In some words, such as *ante* ‘where’ and *anti* ‘like’, they are contrastive, but in words like *puru* and *pero* ‘but’, they are not. These recent changes are probably contact-induced.

As Himes (2012) and Del Corro (1980) discussed, [ɛ] is the product of the reduction or monophthongization of the diphthong *ay. This [ɛ] can be considered phonemic as evidenced by the minimal pairs provided by Forman:

- (25) /i/, /ɛ/
 [dɛ] > [da] ‘3PL.GEN’ + [ya] ‘3SG.ABS’
 [di] ‘plural core marker’

The [ɛ] in (25a) is a product of diphthong reduction that contrasts with [i] in *di* showing that /ɛ/ can be considered a phoneme, albeit one that is formed only recently. /ɛ/ however shares the same vowel space with the lowered variant of /i/, which occurs in the final syllable of a word, or in the final syllable of an utterance with high pitch, as in questions (Del Corro 1980:8). This overlap between /i/ and /ɛ/ can be explained by (1) the relatively recent formation of /ɛ/ as a phoneme and (2) Kapampangan distinguishing vowel phonemes with tongue position rather than tongue height (Del Corro 1980:8). Similarly, the overlap between these front vowels may also explain why [i] occurs instead of [ɛ] in stems with [a] as V₁ that are infixed with <i>, as can be seen in (24). This might also explain the apparent aberration in (23), wherein the preference to one alternant, in this case [i], has been fossilized.

4 Conclusion

The analysis presented above has taken into consideration the general diachronic patterns and morphophonemic processes present in Kapampangan and other Philippine-type languages. Since the infixation of the perfective affix <in> is established in Kapampangan, it follows that its counterpart <i> behaves similarly, meaning that the morpheme also should be an infix as well. The allomorphs when <i> is affixed to stems is probably due to diphthong reduction, a phonological change in Kapampangan that previous studies only associate with diphthongs in word-final position. This alternative is believed to be more satisfying than the so-called vowel change and use of replacive vowels used in previous analyses.

Based on these, we summarize the morphophonemic rules of infix <i> in (26). The infix is inserted between the first consonant and first vowel of the stem, with the resultant form varying depending on its V₁ and syllable structure.

- (26) when V₁ is [a], it becomes [ɛ:] but,
 when V₁ is [a] and the syllable has a CVC structure, it becomes [i]
 when V₁ is [u], it becomes [i:] but,
 when V₁ is [u] and the syllable has a CVC structure, it becomes [i]
 when V₁ is [i], it becomes [i:] but,
 when V₁ is [i] and the syllable has a CVC structure, it does not change

References

- Bergaño, Diego. 1916. *Arte de la lengua pampanga*. Tercera Edición. Manila: Tip. del Colegio de Santo Tomás.
- Bergaño, Diego. 1860. *Vocabulario de Pampango en Romance y Diccionario de Romance en Pampango*. Manila: Convento de Nuestra Señora de los Angeles.
- Blust, Robert. 2013. *The Austronesian languages*. Canberra: Asia-Pacific Linguistics Research School of Pacific and Asian Studies, Australian National University. Retrieved August 19, 2021, from <https://openresearch-repository.anu.edu.au/handle/1885/10191>.
- Cruz, Johans.; Pelagio, Earvin Christian.; Quizon, Rexie; & dela Rosa, Kriziel. 2020. *Revisiting Calaguiman-Mabatang as a relic area of Kapampangan*. Paper presented at the The Linguistics

Student Research Colloquium of University of the Philippines Department of Linguistics,
Quezon City.

Del Corro, Anicia. 1980. *Kapampangan Morphophonemics*. Cecilio Lopez Archives of Philippine Languages. University of the Philippines Diliman.

Del Corro, Anicia. 1988. *A Sequel to the Dialect Study of Kapampangan (1984): With Focus on Calaguiman and Mabatang Kapampangan*. University of the Philippines Diliman.

Forman, Michael L. 2019. *Kapampangan Grammar Notes*. Honolulu: University of Hawai'i Press.
<https://www.jstor.org/stable/j.ctv9hvsmc>.

Himes, Ronald. 2012. *The Central Luzon Group of Languages*. *Oceanic Linguistics* 51.2:490-537.
Retrieved August 7, 2021, from <http://www.jstor.org/stable/23321866>

Mirikitani, Leatrice. 1972. Kapampangan Syntax. *Oceanic Linguistics Special Publications* 10:I-263.
Retrieved August 30, 2021, from <http://www.jstor.org/stable/20019139>

NOMINAL ANCHORING FUNCTIONS OF POROHANON COMMON NOUN MARKERS¹

Vincent Christopher A. SANTIAGO
University of the Philippines Diliman
vasantiago@up.edu.ph

Abstract

Porohanon, spoken in the Municipality of Poro, Camotes, Cebu is a member of the Central Bisayan branch of the Bisayan complex (Zorc 1977). Previous descriptions of this speech variety (Wolff 1967, Zorc 1977, Ballo 2011) have tended to classify its common noun markers into NOMINATIVE, GENITIVE, and OBLIQUE case forms. These forms are also purported to encode distinctions of DEFINITE versus INDEFINITE and SPECIFIC versus NONSPECIFIC, notions which fall under what Balogh, Latrouite, and Van Valin (2020) call “nominal anchoring”. The current study re-evaluates the functions and present-day forms of these common noun markers using written and spoken data. An alternative classification is proposed in this paper’s conclusion. The syntactic alignment of Porohanon is also reassessed considering more contemporary research on ergativity in Philippine languages.

Keywords: Porohanon, nominal anchoring, definiteness, specificity
ISO 639-3 codes: prh

Introduction

Nominal anchoring (Balogh, Latrouite, and Van Valin 2020) is a vital component of any language. As a human system employed to refer to entities in the real world or to participants in a situation, various linguistic resources and structures are utilized to meet these needs. Notions traditionally associated with this such as **definiteness** and **specificity** have enjoyed wide coverage and exhaustive discussion in the philosophy of language and theoretical linguistics literature (cf. Balogh, Latrouite, and Van Valin 2020 for a cursory survey).

This paper examines the common noun markers of Porohanon spoken primarily in the Municipality of Poro, Camotes, Cebu, Philippines. The common noun markers are analyzed for their nominal anchoring functions in written and spoken data. The present study is intended to be another building block toward a more comprehensive grammatical description of Porohanon, which is relatively understudied and underdescribed compared to other varieties and Philippine languages of wider communication, such as Cebuano and Waray.

Zorc’s monumental *The Bisayan Dialects of the Philippines: Subgrouping and Reconstruction* (1977) proved to be an indispensable resource for the current study. One can even say that the work has already laid-out the main lines of inquiry and has documented the most significant aspects of the grammars of Bisayan varieties. Apart from lacunae in the Porohanon data acknowledged by the author

¹ List of Abbreviations:

1 – first person, 2 – second person, 3 – third person, A – the most agentive core argument of a transitive verb; APPL – applicative; ABS – absolutive; CAUS – causative; COMPL – completive; CONJ – conjunction; DEF – definite; DIST – distal; E – extended argument; ENUM - enumerative; ERG - ergative ; EXCL – exclusive; GEN – genitive; HAPP – happenstance; HES – hesitation pause; INDF – indefinite; INF – infinitive; INTR – intransitive; IPFV – imperfective, IRR – irrealis; LNK – linker; LOC – locative; NONSPEC – nonspecific; O – the most patientive core argument of a transitive verb; OBL – oblique; PL – plural; PFV – perfective, POS – postposed form; PRE – preposed form; PROX – proximal; PST - past; Q – question word; REAL – realis; S – the single core argument of an intransitive verb; SG – singular; SPEC – specific; STEM – stem-forming affix; STAT – stative; TR – transitive; = – clitic boundary; - - morpheme boundary

up-front (Zorc 1977:269, 276), another niche that the current study intends to occupy is presenting findings from **naturalistic speech data**. While traditional sentence elicitation may have already uncovered the basic structures at work, it is my belief that transcriptions of continuous speech recordings could yield valuable complementary observations. Several language-specific resources were also consulted and are cited throughout the rest of the paper.

Referential expressions in Philippine languages

McFarland (1978:151) operationalizes **reference** based on his definition of what a **referent** is, namely, a “non-linguistic entit[y] which [is] talked about”. Reference, therefore, is “the linguistic process whereby referents are identified.”

Referential expressions, or “reference expressions”², in Philippine languages tend to fall into four (4) types: (1) common referential expression, (2) personal referential expressions, (3) personal pronouns, and (4) deictics (McFarland 1978:141). In McFarland’s model of Philippine syntax, Table 1 shows the ways these types are distinguished. This paper only focuses on the first type, including the common referential expressions in Poroohanon.

Table 1: McFarland’s (1978:154) classification of referential expressions

Categories	Definitions
<i>Common Referential Expression</i>	“Names of objects and places are marked, and treated syntactically, as common reference expressions.”
<i>Personal Referential Expression</i>	“In Philippine languages, labels which are attached to persons and personified beings (personal names) are marked as personal referential expressions.”
<i>Personal Pronouns</i>	“Personal pronouns distinguish referents on the basis of the speaker-addressee relationship.”
<i>Deictics</i>	“Deictic pronouns distinguish referents on the basis of the spatial relationship (nearness or remoteness) to the speaker, and perhaps the addressee”

On the structure of common referential expressions, McFarland (1978:141) states that “A CRE [Common Referential Expression] (something like a common noun) in its most general form consists of a CRE article and a predicate phrase”.

Reid and Liao (2004:464), meanwhile, describe referential expressions as “strongly right-branching, with heads preceding modifiers”. This right-branching tendency dictates that “determiners” (as per Reid and Liao, 2004:464, roughly equivalent to “articles” in McFarland 1978 and Balogh, Latrouite, and Van Valin 2020) appear before the so-called “head noun”.³

Poroohanon common noun markers

Table 2 lists the forms of the common noun markers in Poroohanon. They are categorized according to the core cases ABSOLUTE (ABS), ERGATIVE (ERG), and GENITIVE (GEN) with the OBLIQUE (OBL) marker in the final column. The forms in boldface have already been identified by Wolff (1967:66) and Zorc (1977:85) and are, for the most part, corroborated by the data I have gathered. The rest of the alternant forms not in boldface are discussed in the following subsections.

² A terminological (and perhaps, conceptual) note: McFarland (1978) has elaborated on his choice of using “reference expressions” or “REs” rather than the more common (but also less theory-neutral) label “noun phrase” or “NP”. I tend to agree with McFarland’s choice and will be employing this label throughout the paper, even if more contemporary studies such as Reid and Liao (2004) still use the label “noun phrase/ NP”.

³ Reid (2002) has argued elsewhere that these monosyllabic forms prevalent in Philippine languages commonly analyzed as “determiners” or “articles” are better taken to be the head nouns in these constructions, and that the following verb or noun serves as its complement.

Table 2: Porohanon common noun markers

ABS	ERG	GEN	OBL
an [ʔan] / ang [ʔaŋ]	san [san] / sa [sa]	san [san] / sa [sa]	sa [sa]
in [ʔin] / =y [j]	sin [sm]	sin [sm]	

Ergativity in Porohanon

Firstly, a comment on syntactic alignment. In previous studies (Santiago 2018, 2019), I have labelled the case forms of Porohanon referential expressions as NOMINATIVE (NOM)-GENITIVE (GEN)-OBLIQUE (OBL) following the categories adopted by past studies.

This topic has been reviewed by a number of studies in the past. Wolff (1967) employs a NOMINATIVE-GENITIVE-LOCATIVE classification for the “construction markers” of Porohanon. Meanwhile, Zorc (1977:69) writes, “The case system of Bs [Bisayan] nominals includes three categories: **nominative**, **genitive**, and **oblique**. [emphasis added]”. He adds that NOM forms occur mainly as “topics of a clause”, his phrasing for describing the most privileged syntactic argument. Meanwhile, there is considerably more nuance to the types of nominals that are cast as GEN and OBL according to their “form, meaning, distribution, and use” (Zorc 1977:69). McFarland (1978:141), analyzing Bikol Legazpi and Tagalog, employs a similar case inventory, calling these “basic surface forms for each type of RE [reference expression]”. Ballo’s (2011) undergraduate capstone research project on Porohanon essentially adopts Wolff’s (1967) inventory of “construction markers”.

The classification in Reid and Liao (2004) is identical to that of Wolff (1967): “We choose to distinguish between case forms such as NOMINATIVE, GENITIVE, LOCATIVE, etc. marked either morphologically (i.e., by the actual form either of the nominal itself or one of its co-constituents), or syntactically (i.e., by word order), and case relations, namely PATIENT, AGENT, CORRESPONDENT, MEANS, and LOCUS.” (Reid & Liao 2004:434) Worth noting at this juncture is that these researchers do analyze Philippine languages as exhibiting an **ERGATIVE** syntactic alignment, a claim absent in the studies cited previously. However, they prefer to stick to using the term NOMINATIVE over the term ABSOLUTIVE to refer to the “least indispensable complement of a basic predication.” (Reid & Liao 2004:435) They also describe NOMINATIVE as the “typologically more general term”.

The present stage of my research on Porohanon has enabled me to put forward a more informed stance on the issue. The variables “S, A, O”⁴ are used to categorize the various referential expressions and their relation to the predicate of a clause. Slightly modifying Dixon’s (1968, 1972) original formulation, the variables are defined as follows:

- S – the single core argument⁵ of an intransitive verb
- A – the most agentive core argument of a transitive verb
- O – the most patientive core argument of a transitive verb

Following this set of variables, one of the parameters identified for prototypical transitive constructions is a distinct source of action (A) apart from the most affected entity (O) (Nolasco 2009:9). An intransitive construction, on the other hand, involves the convergence of the “source of action/ most agentive core argument” and the “most affected entity/ most patientive core argument” on a single argument of the clause, thus labelled (S). Non-core arguments—those least “immediately-involved”⁶ arguments in a clause which cannot be considered, S, A, or O—are encoded as obliques or OBL. Typically, **spatial**, **temporal** expressions are encoded as obliques.

⁴ This system was first introduced by Dixon (1968, 1972) as a heuristic for demonstrating the different types of syntactic alignment exhibited in the marking of arguments in Dyrbal in contrast to Indo-European languages.

⁵ This modification of Dixon’s original formulation is done to acknowledge Mithun and Chafe’s (1999) reservations with the commensurability of the notion of ‘subject’ for typologically diverse languages.

⁶ This characterization of “immediacy of involvement” in the action/event expressed by the predicate follows Mithun (1994:255) in her description of the ABSOLUTIVE as the “participant that is the most immediately or directly involved in an event or state”.

Regarding the predicate *giputol* ‘chopped’, indeed, this action of chopping a piece of wood with an instrument such as an axe involves much effort and intention on the part of its source, a person. Also, the entity being chopped, a piece of wood, undergoes a transformation in its physical state, what Nolasco (2009:8) in his revision of Hopper and Thompson’s (1980) transitivity parameters categorizes as “total affectedness”.

Nolasco (2009:22) adds that “most work subscribing to the ergative analysis has tended to focus on the formal aspects of the phenomenon, downplaying its semantic, pragmatic and discourse motivations... The meaning-based and formal evidence points to the robust manifestation of the ergative-absolutive relation in Philippine-type languages...” Thus, this excursus into transitivity in Porohanon has been essential in demonstrating its ERG-ABS syntactic alignment.

On the distinction between the grammatical relations ERG and GEN

Something needs to be said about distinguishing the grammatical relations ERG from GEN, even if their common noun marker forms are identical.⁸ I posit that the source of the action in a transitive construction, the ERG, is a grammatical relation distinct from the possessor (Kroeger 2005:104), the GEN. The homophony of the forms for these two grammatical relations can be observed in example (4).

- (4) *Gisuwat* *sa taw* *an pangan* *sa daga*
 Gi-suwat **sa**=taw an=pangan **sa**=daga
 REAL.PFV.TR-write **ERG**=young.man ABS=name **GEN**=young.lady
 ‘The young man wrote the young lady’s name.’
 (#354, UP Dept. of Linguistics 775-sentence list)

The ERG-marked argument *taw* ‘young man’ is the source of the action *gisuwat* ‘wrote’ which affects the argument *pangan* ‘name’, marked ABS. The action of the young man gives rise to the name on the written page. Meanwhile, the one who possesses that name, the *daga* ‘young lady’ is in a different case, GEN.

On the apparent homophony among the ERG/GEN and OBL common noun markers

Not only can the markers for ERG/GEN fall together, but also the common noun marker for OBL, as seen in the referential expression *sa bay* ‘the house’ in example (2). The linguist who wishes to describe the common noun marker system of Porohanon must now contend with three distinct grammatical relations—the ERG, GEN, and OBL—converging on one form: *sa* [sa].

Wolff (1967), later cited by Zorc (1977), did not record this phenomenon in Porohanon. This homophony was probably not yet apparent in the late 60s in the speech of John Wolff’s language consultants (Wolff 1967:78). Zorc, however, makes the important observation that “the Ceb [Cebuano] oblique and definite genitive markers are homophonous (sa).” (1977:97) Ballo (2011:67) also notes that the marker *san* is already falling into disuse among speakers of Porohanon.

Once again, like the case of the ABS common noun marker *an* [ʔan]/ *ang* [ʔaŋ] discussed in footnote 7, one can no longer hold that this is simply a process of apocope, or the loss of the final segment of a morpheme (Crowley & Bowerman 2010). Could this be another indication, a sociolinguistic variable, that Porohanon is becoming more and more like Cebuano?⁹

⁸ McFarland (1978:140) rightly observes that “The various cases in Philippine languages are not so clearly or discretely marked as in many languages.”

⁹ Notably, I was able to interview Atty. Lourdito Borlasa, one of Wolff’s original language consultants, when I had conducted fieldwork in Poro, Camotes, Cebu last 2018. He is of the opinion that Porohanon is already a dialect of Cebuano, just as Wolff argued in his 1967 paper. My other language consultants, on the other hand, stress the unintelligibility of Porohanon to Cebuano speakers. They can understand and can easily switch to Cebuano, but Cebuano speakers from the “mainland” have a difficult time understanding their *tinaga-Poro* ‘(lit.) from Poro’ speech. (E. Marquez, personal communication, 12 December 2018). The following excerpt from Lobel (2006:915) suggests an extended history of contact between Porohanon and

Definiteness and specificity in Philippine languages

Among the notions of nominal anchoring, definiteness seems to be the one that has received the most coverage in key works of Philippine morphosyntax. Despite the frequent reference to this notion, different scholars seem to have different versions of it and acknowledge different manifestations of it.

Constantino, for example, considers the “simple, predicative, definite sentence” (1965:108, 1971:1) to be the kernel sentence from which other construction types are derived through transformations. Constantino (1971:2) later elaborated on this notion, stating that “they [simple, predicative, definite sentences] are definite in that both their subject and predicate are ‘marked’, that is, each is preceded or followed by an article or affix. The “Definiteness Hypothesis” of Constantino, then, seems to be hinged on the mere presence or absence of an article or affix in the subject and predicate of a sentence. This is so that Tagalog sentence pairs such as (5) and (6) are treated as definite and indefinite, respectively, based on the presence of *ang* in the first constituent (English glosses are Constantino’s own):

- (5) *Ang bata ang kumain sa mangga.*
 Ang=bata ang=k<um>ain sa=mangga
 ABS=child ABS=<INTR.PFV>eat OBL=mango
 ‘It was the child who ate the mango.’ (Constantino 1971:66)
- (6) *Bata ang kumain sa mangga.*
 Bata ang=k<um>ain sa=mangga
 Child ABS=<INTR.PFV>eat OBL=mango
 ‘It was a child who ate the mango.’ (Constantino 1971:66)

Cubar ([1975] 2019) wrote an extended critique of this analysis, calling him out, among other things, on his decision to label the second *ang* in sentences such as those above a “predicate marker”.

“It is obvious that we have here a different notion of what a *noun phrase* is. For Constantino, *ang tumakbo* ‘the one who ran’, *ang kinain ng bata* ‘that which was eaten by the child’, *ang maganda* ‘the one that is pretty’, and *ang nasa kahon* ‘that which is in the box’ are not noun phrases because their heads or centers are not nouns. He would call the first two phrases **verb phrases**, the third **adjective phrase**, and the last **particulate phrase**. However, these phrases have unmistakable nominal readings. They are what linguistic philosophers call **definite reference expressions** – expressions which are used for naming objects.” (Cubar [1975] 2019:73)

Cubar still anchors the notion on the presence of article/affix-marking in the sentence constituents. However, he offers a more detailed discussion of referential expressions based on “degrees of definiteness”. A definite common noun, according to him, “derives its definiteness either from the presence of its referent in the common immediate environment of the speaker and the hearer, or from linguistic anaphora, including the use of definitizing attached relative clause.” (Cubar [1975] 2019:84) Addressing another point of contention in Constantino’s analysis, Cubar writes that “A speaker uses the definite form of a noun phrase when he assumes that the existence of a referent has been registered in the consciousness of the hearer, or when he believes that the referent has been sufficiently described such that it has a determined identification for the hearer.” (Cubar [1975] 2019:91)

McFarland problematized the same issue three years later in his article *Definite objects and subject selection in Philippine languages* (1978) beginning with his conception of reference already cited earlier in this paper. “Definite reference”, according to McFarland (1978:153), “indicates that the referent is specific and known to the speaker and known to the addressee.” “Indefinite reference”, on the other hand, “indicates that the referent is non-specific or unknown to the speaker or unknown to the addressee.” Reid and Liao (2004:469) write that “In all Philippine languages, Nominative phrases

Cebuano: “It is also interesting that the dialects of the oldest settlements in Baybay, Leyte, (C. Rubino, personal communication), and the Camotes Islands (Wolff, 1967) show a Warayan substratum, indicating that Waray-Waray was much more widespread in previous centuries before the expansion of the Cebuanos in the mid-1800s (Larkin, 1982).”

typically have a definite interpretation, that is, the speaker assumes that the addressee knows the general reference of the actant which is the head of the phrase.” An “indefinite actant”, on the other hand, “is typically expressed by a phrase carrying the Correspondent case relation in an intransitive clause and is marked with a Locative, Genitive, or Oblique Determiner...”

Let us now turn to the notion of specificity. McFarland (1978:151) had already discussed specificity in relation to his notion of definiteness. According to him, “A referent may be specific, non-specific, or generic. In the first case, the speaker is saying something about a particular, i.e., specific individual.” A speaker refers to something generic when s/he refers to “a whole class of entities” or uses a form “intended to apply to all members of that class.” Therefore, “[i]f the referent is specific, it has an identity. That is, it has an existence which is distinct from all other referents, even those which may be very similar to it.” (McFarland 1978:152)

McFarland’s crucial contribution lies in his disambiguation of specificity from definiteness: “The identity of a specific referent may or may not be known to the speaker. When I say that a speaker knows the identity of a referent, I mean that (1) he is in possession of a file of prior knowledge about that referent and that (2) he is aware that the referent currently being discussed is the same as the one to whom this prior knowledge relates.” (McFarland 1978:152)

McFarland’s keenness to the negotiatory nature of discourse is evident in the following statement:

“If the identity of the referent is known to the speaker, he must make a judgment as to whether the addressee also knows [sic] the identity. If he judges that the addressee already possesses a file of information about the referent, the speaker must choose a linguistic expression which will enable the addressee to locate that file and to add the new information to what is already there. If he judges that the addressee does not have such a file, he may choose a linguistic expression which indicates to the addressee that he is not expected to have a file on the particular referent, and that he may or may not choose to open such a file. Such information is carried by ‘indefinite noun phrases’ and existential sentences.” (McFarland 1978:152)

Reid and Liao (2004) maintain the distinction between definiteness and specificity by writing that:

“Knowing the general reference of an actant does not imply that the addressee knows the specific actant being referred to. Although Nominative phrases are typically definite, they may or may not be specific. The degree of specificity often depends on the presence of a demonstrative, either as the head of the noun phrase, or as a post-head modifier, or on the presence of some other post-head modifier such as a genitively marked noun phrase, or a relative clause. A number of languages mark a distinction between specific and non-specific phrases, with the specific phrase being invariably marked by a form which is either a demonstrative, or can be shown to have been a demonstrative at some earlier stage of the language.” (Reid and Liao 2004:471)

Table 3 summarizes the key points pertaining to definiteness and specificity across the cited literature.

Table 3: Summary of notions of definiteness and specificity in Constantino (1965, 1971), Cubar ([1975] 2019), McFarland (1978), Reid and Liao (2004)

Category	A	B	C	D	E
Notion	Constantino 1965	Constantino 1971	Cubar [1975] 2019	McFarland 1978	Reid and Liao 2004
Definiteness	-	FORMAL CRITERION: Presence of subject/ “predicate” marking thru article or affix	FORMAL CRITERIA: Presence of constituent marking thru article or affix, Linguistic anaphora, Definitizing attached relative clause SEMANTICO-PRAGMATIC CRITERIA: Presence of referent in the common immediate environment of the speaker and hearer, Registration in the consciousness of the hearer, Determined identification for the hearer	FORMAL CRITERION: Reference expression marking SEMANTICO-PRAGMATIC CRITERION: Specificity and knownness to the hearer and addressee	FORMAL CRITERION: Case-marking (Nominative, typically for definite, Locative, Genitive, or Oblique for indefinite) SEMANTICO-PRAGMATIC CRITERION: Assumption of the speaker that the addressee knows the general reference of the actant
Specificity	-	-	-	FORMAL CRITERION: Reference expression marking SEMANTICO-PRAGMATIC CRITERIA: Particularization of an individual, Existence which is distinct from all other referents	FORMAL CRITERION: Presence of a demonstrative, Presence of some other post-head modifier such as a genitively marked noun phrase or relative clause

Definiteness and specificity in Porohanon

We return to the discussion of the common noun marker forms of Porohanon. Wolff (1967) distinguishes the first row of markers as “definite” from the second row which are “indefinite”. Much like Constantino, however, he takes as a given the reader’s notion of definiteness and does not elaborate on it anywhere else in the paper:

Table 4: Porohanon common noun markers (revised according to the Wolff (1967) classification)

	ABS	ERG	GEN	OBL
Definite	an [ʔan]/ ang [ʔaŋ]	san [san]/ sa [sa]	san [san]/ sa [sa]	sa [sa]
Indefinite	in [ʔin] / =y [j]	sin [sin]	sin [sin]	

Zorc (1977:84) stipulates that for the entire Bisayan complex, “All dialects that have two genitive markers **can make a distinction between definite and indefinite.**” He recognizes the “differences in formation” of these common noun markers, stating that the vowel [a] almost always occurs in “general,

definite, or past” markers and the vowel [i] almost always occurs in “indefinite or nonpast” markers. (Zorc 1977:86)

How does definiteness, purported to be the feature that distinguishes the two forms in the ABS, operate in naturalistic speech? The following excerpt is from a Pear Film narrative (Chafe (Ed.) 1980)¹⁰ recorded with Mr. Joseph Andriano, 33 years old (at the time of recording), from Brgy. Teguis, Poro, Camotes, Cebu. After watching the video and thinking of what to say for a few moments, he begins by introducing his narrative, as in (6).

- (6) *Ara nakoy istorya* *nimo,* *Sir Vincci*
 ara=na=ko=y=istorya (ka)nimo Sir Vincci
 EXIST=COMPL=1SG.GEN.POS=ABS.INDF=story 2SG.GEN.PRE Sir Vincci
 ‘I already have a story for you, Sir Vincci.’ (Joseph Andriano – Pear Story, 0:00)

Wolff (1967:66) observed that “The indefinite Cebuano marker *y* refers to an indefinite subject only in certain set of expressions.” Tanangkingsing (2011:146) later wrote that “...the referent is indefinite and is marked by the neutral marker =*y* that phonologically attaches to the preceding unit”

Wolff states that Porohanon makes a distinction that Cebuano does not make, in that it still has two unreduced forms for the “subject marker” *an* versus *in*. This is reflected in Table 4. It seems, though, that contemporary speakers of Porohanon like Sir Joseph tend to reduce this indefinite ABS marker, just like in Cebuano.

The phonological change is not yet absolute, however, since there are still instances where =*y* retains its full form *in* [ʔin]. See sentence (7) from the article *PESO*¹¹ *MANG REHESTRO* from the Porohanon Newsletter (September-October, 2012:1).

- (7) *Inin* *pagpang rehistro nga gibuhāt* *sa PESO*
 inin pag-pang-rehistro=nga=gi-buhāt sa=PESO
 ABS.DEM.PROX INF-DISTR-register=LNK=REAL.PFV.TR-do OBL=PESO

wa ini in bazad.
 wa ini in=bazad
 NEG LOC.PROXABS.INDF-payment
 ‘This registration to PESO, this has no payment.’

Now, let us return to Sir Joseph’s Pear Story narrative to see how the supposed definite counterpart of the ABS common noun marker functions. Our young male protagonist has gone off after his fall from the bicycle he was riding. However, Sir Joseph said that the young boy had forgotten his hat:

- (8) *Nan sara, nalimtan na niza,* *uh,*
 nan sara n(k)a-lim(o)t-an=na=niza uh
 CONJ now REAL.PFV.(STEM).HAPP-forget-APPL=COMPL=3SG.GEN.POS HES

an iza kalo.
 an=iza=kalo
 ABS.DEF=3SG.GEN.PRE=hat
 ‘And now, he had already forgotten his, uh, his hat.’ (Joseph Andriano – Pear Story, 02:13.01)

In this example, the common noun marker *an* is selected for *iza kalo* ‘his hat’. Designating *an* as definite within the larger context of discourse would be peculiar because **this is the first time this referent is**

¹⁰ The video is available on Youtube: <https://www.youtube.com/watch?v=bRNSTxTpG7U&t=154s>.

¹¹ Public Employment and Service Office

introduced. Without the aid of the video, it is only through sentence (7) that our young male protagonist is specified as actually wearing a hat. If this is the very first time this referent *iza kalo* ‘his hat’ is “pushed into the scene” in this narrative, why is it being marked with a supposedly “definite” common noun marker when there is no prior knowledge of this referent on the addressee’s part?

Specificity might be the more appropriate notion to associate with this form, especially since Reid and Liao (2004) consider the presence of genitive marking an indicator that this referent is specific. Indeed, with the third-person, genitive, preposed pronoun *iza*, the speaker’s proposition is that the *kalo* ‘hat’ is only the young male protagonist’s and no one else’s.

Let us try to examine the ERG and GEN common noun markers this time. In this part of the story, the three children met by the young male protagonist have returned the hat which he had dropped. In return, the boy who owns the hat gave them each a pear. So the children part ways and the gang of three who helped the boy on the bicycle go along the path toward the farmer picking the pears from a tree, as in (9).

- (9) *An tulo ka puza, padung didto sa nangipo sin peras.*
 an=tulo ka-puza pa-dung didto sa=n(p)ang-ipo **sin**=peras
 ABS=three-ENUM-child CAUS-go LOC.DIST OBL=REAL.PFV.DISTR-pick **GEN.INDF?**=pear
 ‘The three children went back to the one who was picking pears.’
 (Joseph Andriano – Pear Story, 02:35.01)

Analyzing the common noun marker *sin* as indefinite in the referential expression *sin peras* does not seem to work because the very beginning of the Pear Film is a shot of a middle-aged man up in a ladder on a tree picking pears. More importantly, it has already been established early on in Sir Joseph’s narrative that this man in the video is, indeed, a man who picks pears. To consider *sin* as indefinite would be ignoring this knowledge of the referent by the speaker and addressee.

Analyzing it as **nonspecific**, however, would yield a more accurate reading. *sin* could be classified as nonspecific since it was shown that the farmer had already picked multiple baskets of pears. What Sir Joseph’s proposition is here is that the farmer is a person **who simply picks pears**, not one, particular, individuated pear, but rather just entities that would be considered pears.

Can we, therefore, consider *san/ sa* the **specific** counterpart? Consider sentence (10). Sentence (10) comes from the shared experiences of Mr. Abel Garciano, a municipal official at the Local Government Unit of Poro, when Super Typhoon Yolanda hit Eastern Visayas in 2013. In sentence (9), we see the ERG marker *sa* marking the common noun *bagzo* ‘storm’. *Bagzo* is then being modified by a relative clause *nga Yolanda* ‘which is Yolanda’. Following McFarland (1978) and Reid and Liao’s (2004) criteria, the particular, individual referent of *bagzo* is being specified not only through the use of the marker *sa*, but also through the relative clause *nga Yolanda*.

- (10) *Diin ang Cebu giagian sa bagzo nga Yolanda.*
 diin ang=Cebu gi-agi-an **sa**=bagzo=nga=Yolanda
 LOC.Q.PST ABS=Cebu TR.REAL.PFV-pass-APPL **ERG.SPEC**=storm=LNK=Yolanda
 ‘Where the storm which is Yolanda passed by Cebu.’
 (Abel Garciano – Unforgettable experience, 14:24)

Conclusion

Evidence from recordings of continuous discourse has shown that in terms of nominal anchoring, the common noun markers of Porohanon seem to encode a SPECIFIC versus NONSPECIFIC reading of the referent, rather than a DEFINITE versus INDEFINITE distinction as earlier assumed by Wolff (1967) and Zorc (1977). Thus, I propose the common noun marking system for Porohanon in Table 5.

Table 5: Revised Porohanon common noun markers

	ABS	ERG	GEN	OBL
Specific	an [ʔan]/ ang [ʔaŋ]	san [san]/ sa [sa]	san [san]/ sa [sa]	sa [sa]
Nonspecific	in [ʔin] / =y [j]	sin [sin]	sin [sin]	

Does this mean that definiteness as a nominal anchoring strategy is totally absent from the system of Porohanon? The possibility is that it may be encoded in other forms of referential expression marking such as in deictics, pronouns, and suffixes.

References

- Ballo, Krystel. 2011. The Porohanon Language of the Camotes Islands. Undergraduate Research Paper. Department of Linguistics, University of the Philippines Diliman.
- Balogh, Kata, Latrouite, Anja, & Van Valin, Robert D., Jr. 2020. *Nominal anchoring: Specificity, definiteness, and article systems across languages*. Language Science Press.
- Chafe, Wallace (Ed.). 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex.
- Constantino, Ernesto. 1965. The sentence patterns of twenty-six Philippine languages. *Lingua* 15:71 – 124. [https://doi.org/10.1016/0024-3841\(65\)90009-4](https://doi.org/10.1016/0024-3841(65)90009-4).
- Constantino, Ernesto. 1972. The deep structures of the Philippine languages. Pre-publication copy. Quezon City: University of the Philippines Diliman.
- Crowley, Terry & Bower, Claire. 2010. *An introduction to historical linguistics (4th Ed.)*. Oxford University Press.
- Cubar, Ernesto. [1975] 2019. *The Archive Classics: Topicalization and Some Related Processes in Philippine Languages*. Department of Linguistics, University of the Philippines Diliman.
- Dixon, Robert M. W. 1968. The Dyirbal language of North Queensland. Dissertation. University of London
- Dixon, Robert M. W. 1972. *The Dyirbal language of North Queensland*. Cambridge University Press.
- Eberhard, David M., Simons, Gary F. & Fennig, Charles D. 2020. Porohanon. *Ethnologue: Languages of the World*, Twenty-third Edition. <https://www.ethnologue.com/language/prh>.
- Hopper, Paul & Thompson, Sandra. 1980. “Transitivity in grammar and discourse.” *Language* 56, 2:251-299.
- Kroeger, Paul R. 2005. *Analyzing grammar: An introduction*. Cambridge University Press.
- Lobel, Jason William. 2006. Samar-Leyte. In K. Brown & S. Ogilvie (Eds.), *Concise Encyclopedia of the Languages of the World* (pp. 914-916). Elsevier Ltd.
- McFarland, Curtis D. 1978. Definite objects and subject selection in Philippine languages. *Studies in Philippine Linguistics* 2.1:139-182.
- Mithun, Marianne. 1994. Implications of Ergativity for a Philippine Voice System. *Voice: Form and Function (Typological Studies in Language 27)*. John Benjamins Publishing Company.
- Mithun, Marianne & Chafe, Wallace. 1999. What are S, A, and O? *Studies in Language* 23.3:569-596.
- Municipality of Poro. 2012, September/October. *Porohanon Newsletter*. Retrieved from <https://www.facebook.com/PorohanonNewsletter/>.
- Nolasco, Ricardo Maria D. 2009. What Philippine Ergativity Really Means. Manuscript. University of the Philippines Diliman.
- Reid, Lawrence A. 2002. Determiners, Nouns, or What? Problems in the Analysis of Some Commonly Occurring Forms in Philippine Languages. *Oceanic Linguistics* 41(2). University of Hawai'i Press.
- Reid, Lawrence A. and Liao, Hsiu-chuan. 2004. A Brief Syntactic Typology of Philippine Languages. *Language and Linguistics* 5.2:433-490, DOI: 2004-0-005-002-000081-1.

- Rosero, Michael Wilson. 2011. A Grammatical Sketch of Masbatenyo. Undergraduate Research Paper. Department of Linguistics, University of the Philippines Diliman.
- Santiago, Vincent Christopher. 14 November, 2018. Porohanon Discourse Markers: Forms & Functions. Paper read at the 13th Philippine Linguistics Congress. NISMED Auditorium, University of the Philippines Diliman.
- Santiago, Vincent Christopher. 29 March, 2019. Existential, possessive, and locative in Porohanon. 2019 Philippine Indigenous Languages Lecture Series. In celebration of the International Year for Indigenous Languages. CSSP Faculty Lounge, Rm. 2134, Palma Hall Pavilion 2, University of the Philippines Diliman.
- UP Department of Linguistics. n.d. 775-sentence list. Department of Linguistics, College of Social Sciences and Philosophy, University of the Philippines Diliman.
- Wolff, John U. 1967. History of the Dialect of the Camotes Islands, Philippines, and the Spread of Cebuano Bisayan. *Oceanic Linguistics* 6.2:63-79.
- Zorc, R. David. 1977. The Bisayan dialects of the Philippines: Subgrouping and reconstruction. *Pacific Linguistics* C(44). Department of Linguistics, School of Pacific Studies, The Australian National University.

NONCANONICAL PRONOUNS IN VIETNAMESE AND CHINESE

Chenchen SONG Li NGUYEN
Zhejiang University University of Cambridge
cjs021@zju.edu.cn nhbn2@cl.cam.ac.uk

Abstract

In this paper, we study a new type of pronominal item emerging on the Internet in Vietnamese and Chinese. First, we demonstrate that pronominal items of this new type, which we dub “noncanonical”, are a separate category from both textbook default pronouns and imposters (Collins and Postal 2012). Then, we illustrate their real-life usage in detail. Our investigation shows that noncanonical pronouns in the two Asian languages are similar not only in syntactic behavior but also in lexical sources, based on which we propose three subtypes for them. Finally, we account for the half-grammatical-half-lexical status of noncanonical pronouns in the theory of generalized root syntax (Song 2019), a recent version of syntactic root theory. We also suggest a link between the propensity for noncanonical grammatical elements and high analyticity.

Keywords: Vietnamese, Chinese, pronoun, Internet, root syntax, analyticity
ISO 639-3 codes: vie, zho

1. Introduction

1.1 *Pronominal items in previous research*

Previous research has documented two types of personal pronominal items in human language. The first type is the default personal pronoun (henceforth default pronoun)¹—namely, the kind of pronoun typically seen in language textbooks and reference grammars, as exemplified in (1).

- (1) a. English: I, you, he, she, it...
b. German: *ich, du, er, sie, es*...
c. Vietnamese: *tôi, bạn* (lit. ‘friend’), *anh ta, cô ta*...
d. Mandarin: *wǒ, nǐ, tā* (‘he/she/it’)...

Such pronouns are not necessarily the most often used, especially in languages like Vietnamese, but they are deemed textbook standards and are also the most widely studied type of pronominal item in linguistics. Textbook pronouns can be viewed as exponents of formal features, especially person, number, and gender (i.e., phi) features and occasionally also honorific features, as exemplified in (2).

- (2) a. **I** am a student. (I = [1SG])
b. *Möchten Sie etwas zu trinken?* [German]
would.like.3PL you.HON something to drink
“Would you like something to drink?” (Sie = [2SG/PL, HON])

¹ Since we are only concerned with personal pronouns in this study, for simplicity’s sake we use “pronoun” to mean “personal pronoun” throughout. By “default” we mean to contrast the kind of pronominal item in (1) with the imposters and especially the noncanonical pronouns to be introduced below.

In (2a), the English pronoun *I* is the exponent of the featural specification [1SG]. In (2b), the German pronoun *Sie* is the exponent of [2SG/PL, HON]. A hallmark of textbook pronouns is that their grammatical behavior can be explained solely by their formal features.

The second type of pronominal item documented in previous research is the “imposter” (Collins and Postal 2012). An imposter looks like an ordinary referring expression (R-expression) and is subject to third-person agreement in languages where syntactic agreement is required, but it semantically refers to the speaker or the addressee instead of a real third person. See (3) for an illustration.

- (3) a. **Daddy** (= I) is going to get you an ice-cream cone.
 b. Is **the general** (= you) going to dine in his suite. (Collins & Postal 2012:1–3)

Imposters carry more pragmatic content than textbook pronouns. Thus, *daddy* in (3a) sounds more affectionate than *I*, and *the general* in (3b) sounds more formal than *you*. Descriptively we can say that imposters are R-expressions employed to refer to the speaker/addressee. There are many more imposters in English, such as *yours truly* (= I), *this reviewer* (= I), *Madame* (= you), *sweetie* (= you), etc. Imposters are also widely attested in other languages (see the articles in Collins 2014). In fact, they are used so frequently and naturally in some languages, including Vietnamese, that they may be viewed as the de facto default pronominal items there, taking over the role of textbook standard pronouns. Imposters have not taken over the role of textbook pronouns in Chinese but are also highly common. See (4) to (6) for a comparative illustration of Vietnamese and Chinese² imposters, which fall in three subtypes: kin terms, career titles, and personal names.

- (4) a. *Hôm-nay* **mẹ** *sẽ* *ngủ* *làm*. [Viet.]
 today mom.1SG FUT stop work
 “Mom (= I) is having a day off today.” (a mom talking to her child)
- b. *Jīntiān* **māma** *zuò* *yú-tāng*. [Mandarin]
 today mom.1SG make fish-soup
 “Today mom (= I) will cook fish soup.” (a mom talking to her child)
- (5) a. **Thầy** *về-hưu* *rồi*. [Viet.]
 teacher.1SG retire PRF
 “Teacher (= I) has retired.” (a teacher talking to their student)
- b. **Lǎoshī** *yě* *bù* *zhīdào* *dá’àn*. [Mandarin]
 teacher.1SG also not know answer
 “Teacher (= I) don’t know the answer either.” (a teacher talking to their student)
- (6) a. **Linh** *hiểu* *chưa?* [Viet.]
 Linh.2SG understand IMPERF
 “Has Linh (= you) understood yet?” (the addressee’s name is Linh)
- b. **Línglíng** *bù* *chī* *qiǎokèlì* *ma?* [Mandarin]
 Lingling.2SG not eat chocolate Q
 “Lingling (= you) don’t eat chocolate?” (the addressee’s name is Lingling)

Although previous studies on Vietnamese rarely (if ever) use the term “imposter”, that the boldfaced items in (4a) to (6a) qualify as imposters is clear: they are R-expressions employed to refer to the speaker/addressee.

² Unless otherwise specified, all our Chinese data are from Standard Mandarin. We also use the term Common Mandarin (i.e., the Mandarin variety commonly spoken in daily life) when discussing Internet language phenomena that have not yet been officially recognized as part of Standard Mandarin.

Overall, (4) to (6) reveal that imposters are used a lot more freely in Vietnamese³ and Chinese than in English and that as such they are not a crosslinguistically homogeneous phenomenon.⁴ A most noticeable point of variation concerns agreement patterns. Since Vietnamese and Chinese both lack formal agreement, their imposters do not manifest the kind of mismatch between grammatical and notional person discussed at length in Collins and Postal (2012). Note that while the English translations all manifest third-person agreement (e.g., *is*, *has*), there is no syntactic agreement in the Vietnamese sentences. The same is true for the Chinese examples in (4b) to (6b). For this reason, Wang (2014) calls Chinese-style imposters “pseudo-imposters”. For expository convenience we will keep using the umbrella term “imposter” when there is no risk of ambiguity (as Wang himself does).

Quite often, the referents of imposters are contextually determined. In (4a) and (5a) the kin terms meaning “mom” are glossed as 1SG because they are uttered by mothers to their children, but the same sentences can be uttered by children to their mothers, and in that case the same kin terms will be glossed as 2SG. There may be crosslinguistic variation as to how much imposter interpretation depends on context, but it is clear that imposters can have context-dependent person indexing (or “floating reference”, see Alves 2007 for more on Chinese influence on Vietnamese pronouns), unlike textbook pronouns, whose person indices are lexically fixed (e.g., *I* can never mean “you”).

1.2 A new type of pronominal item

Textbook pronouns and imposters, however, are not the only types of “pronouns” out there. Observe the sentences in (7).

- (7) a. *Mị* *thề* *là* *mị* *không* *soi* *cái* *bụng*. [Viet.]
 Mị.1SG swear COP Mị.1SG NEG zoom.into CLF belly
 “Mị (= I) swear Mị (= I) didn’t zoom into (your) belly.”
- b. *Mị* *không* *hiểu*. *Các* *chị* *hiểu* *hông?*
 Mị.1SG NEG understand PL sister understand NEG
 “Mị (= I) don’t understand. Do sisters (= you) understand?”

In (7) the Vietnamese word *Mị* is used in place of a first-person pronoun. *Mị* is originally the name of a character in an old literary work (*Vợ chồng A Phủ* ‘Couple A Phủ’), who suffered a lot of injustice in the old days but stood up for herself and fought for her own happiness. Perhaps inspired by her story, contemporary netizens (mainly females, but occasionally also males) sometimes use her name as a term of self-address with a joking tone. In (7a), for instance, the netizen says “*Mị* swear *Mị* don’t feel upset” instead of “I swear I don’t feel upset” when commenting on her growing belly size; and similarly, in (7b), the netizen says “*Mị* don’t understand” instead of “I don’t understand” to deliberately sound naïve.

There are a number of reasons why items like *Mị* differ from imposters. First, although *Mị* is originally a personal name, its pronominal usage in (7) is clearly different from that of the personal name *Linh* in (6a). Specifically, *Linh* can refer to either the speaker or the addressee whereas *Mị* can only refer to the speaker. Second, when *Linh* is used pronominally, there is actually someone named *Linh* in the discourse, whereas *Mị* could refer to any speaker, similarly to “I” in English. *Mị* thus behaves more like a textbook pronoun than an imposter, except that it carries idiosyncratic extragrammatical and pragmatic effects and is mainly restricted to online usage.

³ As Paul Sidwell pointed out, it is important to note that much of the standard Vietnamese pronoun system was historically replaced by imposters.

⁴ Between Vietnamese and Chinese, imposters are used even more freely in Vietnamese.

This phenomenon is not exclusive to Vietnamese but also observed in Chinese. See (8) for instance.

- (8) a. *Jiù bù gěi āijiā xīnlǐ jiànshè* [Mandarin]
 just not give mourner.1SG psychological construction
shíjiān. Zhè jiù qīn-shàng le. Méi-rén gàosu āijiā
 time now just kiss-up CRS no-person tell mourner.1SG
háiyǒu qīn dì'èr-huí a.
 still kiss second-round EXCL
 “They just don’t give mourner (= me) any time to be psychologically ready. They just begin to kiss right away. No one told mourner (= me) that they were going to kiss again.” (Sina Weibo⁵)
- b. *Guǒrán shùxué de shìjiè shì méi-yǒu sècǎi de.*
 as.expected math POSS world is not-have color NMLZ
Āijiā fá le.
 mourner.1SG tired CRS
 “The world of math is colorless just as expected. Mourner (= I) is tired.”

Here the item of interest is *āijiā*, an ancient term of self-address that literally means “mourner” and was originally used by empress dowagers (i.e., emperors’ mothers). Contemporary netizens (mostly young females, occasionally also males) often use it in a jocularly arrogant tone. Thus, the two speakers in (8) respectively complain about an unexpected kissing scene on TV and the difficulty level of math, both sounding assertive and much more fun than if the default 1SG pronoun *wǒ* is used.⁶ Note that there is an ongoing debate as to whether empress dowagers in Chinese history had really called themselves *āijiā* or this was just a coinage of ancient playwrights, who then passed it on to modern scriptwriters (see Chen 2009). In spite of this, however, there is no doubt that the online term *āijiā* is borrowed from historical contexts and that it synchronically behaves like a 1SG pronoun with special pragmatic effects, similarly to Vietnamese *Mị*.

In sum, pronominal items like Vietnamese *Mị* and Chinese *āijiā*, which represent a fashionable linguistic phenomenon in the Internet era, constitute a unique category (see more examples in §2). Unlike imposters, they do not have flexible, context-dependent person indexing or common R-expression usage. Unlike textbook pronouns, they are not exponents of formal features but carry idiosyncratic extragrammatical effects. We dub them “noncanonical pronominal items” as a working term. Furthermore, noncanonical pronominal items also differ from imposters in terms of their history, typology, and lexical materials. We summarize these distinctions in Table 1.

Table 1: Differences between imposters and noncanonical pronominal items

	Imposters	Noncanonical pronominal items
Usage	wide in real life	limited to certain registers
History	long	emerging (mainly online)
Typology	prevalent in many languages	available in far fewer languages
Reference	flexible (contextual)	fixed (lexical)
R-expression usage	yes	no
Lexical material	nouns in contemporary use	miscellaneous

⁵ Sina Weibo is the Chinese counterpart of Twitter.

⁶ Also notice the speaker’s choice of verb in (7b) in the vicinity of *āijiā*. Here, for “tired”, the ancient-sounding *fá* is used instead of the synchronically more common *lèi*. This sort of stylistic or register-based agreement is commonly observed in Chinese (see Feng 2010 et seq.).

As Table 1 shows, imposters have been in use in both Vietnamese and Chinese for a long time, whereas noncanonical pronominal items have only recently emerged in the Internet era. Imposters are attested in many languages (see Collins and Postal 2012 and Collins 2014), whereas noncanonical pronominal items to our knowledge are less prevalent. Finally, imposters always have contemporary nominal counterparts, which allows them to be used as ordinary R-expressions, whereas noncanonical pronominal items have miscellaneous lexical sources (e.g., *Mị* recycles a fictional character name, and *āijiā* recycles an ancient term of self-address). These distinctions set noncanonical pronominal items apart from imposters.

Despite their particular characteristics, however, noncanonical pronominal items have not been well documented. The aim of our article is thus to present a preliminary investigation of these items in contemporary usage and situate them in modern syntactic theory. To do so, we will first present noncanonical pronominal items in Vietnamese and Chinese in more detail (§2), then refine the crosslinguistic pronominal item taxonomy (§3), and eventually incorporate noncanonical pronominal items in the syntactic theory of pronouns (§4). Finally, we conclude with a few points for future research (§5).

2. Subtypes of noncanonical pronominal items

Noncanonical pronominal items in Vietnamese and Chinese fall in three subtypes based on their lexical sources: revived ancient terms (§2.1), dialectal terms (§2.2), and creative online coinages (§2.3).

2.1 Subtype I: Revived ancient terms

The first subtype of noncanonical pronominal item we have identified in both Vietnamese and Chinese is that of revived ancient terms. Such terms can be either literary, like Vietnamese *Mị* ‘female character name.1SG’, or (quasi) royal, like Chinese *āijiā* ‘mourner.1SG’. Table 2 contains more examples belonging to this subtype. We provide literal glosses in single quotes and original/historical usage restrictions in parentheses.

Table 2: Subtype-I noncanonical pronominal items in Vietnamese and Chinese

	<i>Mị</i> ‘female character name.1SG’
Vietnamese	<i>trẫm</i> ‘emperor.1SG’ (by emperors)
	<i>ái-phi</i> ‘beloved-concubine.2SG’ (by emperors or princes to their concubines)
	<i>āijiā</i> ‘mourner.1SG’ (by empress dowagers)
	<i>zhèn</i> ‘1SG’ (by emperors)
Chinese	<i>guǎ-rén</i> ‘lacking-person.1SG’ (by pre-Qin state rulers)
	<i>běn-gōng</i> ‘this-palace.1SG’ (by emperors’ sons and wives/concubines to inferiors)
	<i>chénqìè</i> ‘slave.1SG’ (by emperors’ wives/concubines to superiors)
	<i>qīng</i> ‘2SG’ (by emperors to royal officials or between husbands and wives)

Vietnamese *trẫm* and *ái-phi* have respectively been borrowed from Chinese *zhèn* and *ài-fēi*,⁷ and have become increasingly popular via hugely successful TV series such as *My Fair Princess*. *Zhèn* had originally been a default 1SG pronoun in Old Chinese (9a) but got reserved for emperors in Qin dynasty

⁷ However, these cognates have clearly developed different uses in the two languages. First, Chinese *zhèn* has stayed a purely pronominal item even after being reserved for emperors, whereas Vietnamese *trẫm* seems to be in the process of further lexicalization (hence our different glosses for them). An informal survey reveals that Vietnamese speakers tend to think *trẫm* means “emperor”, even though the term has no common R-expression usage (otherwise it would be an imposter like *thầy* ‘teacher’). Second, while Vietnamese *ái-phi* is a noncanonical pronominal item, Chinese *ài-fēi* is an imposter (hence its absence from Table 2), for it has common R-expression usage, as in *shénmì wángyē de ài-fēi* ‘the beloved concubine of the mysterious prince’ (novel title).

(221 B.C.E.) (9b). In the pre-Qin era, vassal state rulers humbly referred to themselves as *guǎ-rén* (9c), which literally means “a person who lacks virtue”.⁸

- (9) a. **Zhèn** huángkǎo yuē bóyōng. [Old Chinese]⁹
 1SG ancestor is.called Bóyōng
 “My ancestor’s name is Bóyōng.” (*The Lament*, 3rd century B.C.E.)
- b. *Tiān-zǐ* zì-chēng yuē zhèn.
 heaven-son self-refer is.called 1SG
 “The Heaven’s Son calls himself zhèn.” (*Records of the Grand Historian*, 1st century B.C.E.)
- c. **Guǎ-rén** suī sǐ, yì wú huǐ yān.
 lacking-person.1SG even.if die also not.have regret in.it
 “Even if lacking-person (= I) dies, I will have no regret.” (*Commentary of Zuo*, late 4th century B.C.E.)

As we mentioned earlier, although there are debates over whether some of the ancient terms that are being revived online had really been used in history,¹⁰ the historical origin/usage of a revived term is orthogonal to its synchronic categorial identification. It thus suffices to identify a term as a Subtype-I noncanonical pronominal item based on just two criteria: (i) the term has been borrowed from historical contexts (either real-life or fictional), and (ii) it synchronically qualifies as a noncanonical pronominal item. It is based on these criteria that we have identified the items in Table 2. See (10) and (11) for some real-life examples. Unless otherwise specified, all our Vietnamese examples are taken from Facebook or Twitter, and our Chinese examples, from Sina Weibo.

- (10) a. *Hôm-nay* có ai muốn rủ **trẫm** đi uống [Viet.]
 today have who want invite emperor.1SG go drink
cà-phê không?
 coffee NEG
 “Does anyone want to invite emperor (= me) out for a coffee today?”
- b. **Trẫm** tha tội haha.
 emperor.1SG forgive wrongdoing haha
 “Emperor (= I) forgives (you) haha.”

⁸ It was common practice for ancient Chinese rulers to use humble terms of self-address, so in this regard *zhèn* is an exception, since it sounds authoritative and ruler-like even in archaic Chinese contexts (see, e.g., *Oracle Bone Script Dictionary* by Zhongshu Xu). The first emperor of Qin was responsible for the official royalization of the term according to official historical records (e.g., *Records of the Grand Historian*) and authoritative dictionaries (e.g., *Kangxi Dictionary*, *Xinhua Dictionary*, *Big Dictionary of Chinese Characters*).

⁹ We present historical Chinese examples with Mandarin pronunciation for expository convenience.

¹⁰ A quick search in the Chinese Text Project database (the largest online database of premodern Chinese texts) returns no results for *āijiā* (while *zhèn* and *guǎ-rén* both occur many times), and some modern dictionaries (e.g., *Revised Mandarin Chinese Dictionary*) explicitly mark *āijiā* as a term from traditional Chinese opera.

- (11) a. *Jīntiān fāxiàn-le yì-gēn bái húzi, zhèn hěn yōushāng.* [Mandarin]
today discover-PRF one-CLF white beard zhen.1SG very sad
“Noticed a gray one in the beard today. Zhèn (= I) is very sad.”
- b. *Zìcóng bèi diào-dào xīn bùmén... zhèn jiù méi zhǔndiǎn xiàbān-guò.*
since PASS transfer-to new department... zhèn.1SG still not.have on.time
knock.off-EXP
“Ever since being put in the new department, zhèn (= I) has never been able to knock off on time.”
- c. *Suīrán guǎ-rén shì dānshēn yìzú, dànshì gū juéde, nǚshēng zuì xūyào de shì péibàn.*
although lacking-person.1SG is single community but lone.1SG
think girl most need NMLZ is company
“Although lacking-person (= I) is single by choice, lone (= I) thinks what girls desire is company.”
- d. *Guǎ-rén chí-zǎo huì sǐ-zài shèyǒu yǒngyuǎn chǎo-bù-xǐng zìjǐ de nàozhōng xià.*
lacking-person.1SG late-early will die-at roommate forever wake-not-awake
self REL alarm.clock underneath
“Lacking-person (= I) will sooner or later die from my roommate’s alarm clock, which can never wake herself up.”

In (10), the term *trǎm* in both sentences is used to convey a jokingly arrogant tone. Similarly, in (11) *zhèn* and *guǎ-rén* sound funnily bossy and a lot less sad/mad than if the default 1SG *wǒ* were used. Note that although Vietnamese *trǎm* and Chinese *zhèn* and *guǎ-rén* were all once state rulers’ terms of self-address, they have a key difference: while Vietnamese *trǎm* is only used by male speakers, the two Chinese terms are used by both male and female speakers more or less equally frequently.¹¹ For instance, the netizens in (11a, c) and (11b, d) are respectively male and female Weibo users, and they all sound jokingly pretentious.¹²

There are also predominantly feminine terms in this subtype. See (12) for a Vietnamese example and (13) for two Chinese examples.

- (12) a. *Ái-phi hôm-nay đẹp quá!* [Viet.]
beloved-concubine today pretty INTS
“Beloved-concubine (= you) look gorgeous today!”
- b. *Chào ái-phi, nhớ ái-phi quá.*
Greet beloved-concubine miss beloved-concubine INTS
“Hello beloved-concubine (= you), (I) miss beloved-concubine (= you) a lot.”

¹¹ In particular, *zhèn* has evidently been gender-neutral throughout history, which may have to do with its original status in Old Chinese as a default 1SG pronoun. Thus, the Tang-dynasty empress Wu Zetian also referred to herself as *zhèn*, as is recorded in official historical documents like *New History of Tang* (compiled in the 11th century).

¹² Interestingly, in (11c) the speaker mixes *guǎ-rén* and *gū*. The latter, literally “lone”, is an alternative to *guǎ-rén* and had also been frequently used by rulers in pre-Qin China. This suggests that the revival of ancient terms of address is a quite general trend online.

- (13) a. *Jiù-mìng! Kuài gěi běn-gōng lái yì-píng* [Mandarin]
 save-life quickly give this-palace.1SG bring one-bottle
sù-xiào-jiù-xīn-wán!
 fast-effect-save-heart-ball
 “Help! Quickly bring this-palace (= me) a bottle of instant cardio-reliever pills!”
- b. *Dǎ-qiú dǎ-de běn-gōng yí-gè gēbo cū yí-gè gēbo xì.*
 hit-ball hit-RES this-palace.1SG one-CLF arm thick one-CLF arm thin
Zěnmē pò?
 how break
 “Since this-palace (= I) played too much badminton, one of my arms has become much thicker than the other. How can I get rid of this?”
- c. *Wǒ yào bǎ táobao xiè-le, kànjiàn xīn yuèjì jiù*
 1SG will DISP Taobao uninstall-PRF see new Chinese.rose just
xiǎng mǎi... chéngqiè rěn-bú-zhù wa!
 want buy chéngqiè.1SG endure-not-TEL EXCL
 “I’ll uninstall Taobao, as I want to buy every new Chinese rose I see... chéngqiè (= I) can’t help it!”
- d. *Xiū-wán liǎng-tiān jià, yòu yào huíqù bān-zhuān le.*
 rest-finish two-day break again must return carry-brick CRS
Chéngqiè bù xiǎng shàngbān le.
 chéngqiè.1SG not want work CRS
 “After a two-day break I must return to carry bricks (an idiom for ‘work’) again. Chéngqiè (= I) don’t want to work anymore.”

In (12), the Vietnamese term *ái-phi* is used in a funnily flirtatious way to refer to a female addressee. Like *trăm*, *ái-phi* synchronically retains its original gender (FEM) and is used only by male speakers to female addressees. This shows virtually no deviation from the term’s historical usage. The situation in Chinese, by contrast, is much more complex. Specifically, as (13) illustrates, there are at least two different feminine terms in Subtype I, *běn-gōng* (13a–b) and *chéngqiè* (13c–d). The pragmatic effect of *běn-gōng* is similar to that of *āijiā*, though it sounds slightly less bossy due to the lower ranking of emperors’ wives/concubines than their mothers. *Chéngqiè*, on the other hand, sounds humbler and even a bit miserable due to its historical status as a term of self-address used by low-status females to their (royal) superiors.¹³ Netizens are well aware of the difference between *běn-gōng* and *chéngqiè*, which is reflected in the different contexts of usage in (13a–b) and (13c–d). While *běn-gōng* is used to jokingly give orders to imaginary servants (13a) or to express “worries” about one’s imperfect appearance (13b) (as royal concubines typically did), *chéngqiè* is used for more miserable scenarios, such as overspending (13c) or overworking (13d).

Moreover, the historical usages of *běn-gōng* and *chéngqiè* were much broader than their revived usages. Historically *běn-gōng* could be used by anyone possessing a (royal) palace, including emperors’ wives, high-ranking concubines, and crown princes. But its modern revival is exclusively based on the wife/concubine sense, probably due to the omnipresence of this usage in TV series. Similarly, *chéngqiè* could be used in history by any low-status female when they spoke to royal superiors, including actresses, prostitutes, plebeians, emperors’ wives/concubines, and even empress dowagers when they needed to sound humble (Xia 2018). But its online usage is only based on the wife/concubine sense too, again due to its omnipresence in TV series. In addition, while these terms are predominately used by females, they are occasionally also used by males (mostly gay). For instance, a flamboyant gay character Yu Hao in a TV series *Stand by Me* constantly refers to himself as *āijiā*, and a gay vlogger and cosmetics expert Benny Dong on Bilibili (the Chinese YouTube) regularly calls himself *běn-gōng* in his video

¹³ The ultimate origin of *chéngqiè* was an Old Chinese compound meaning “slaves” (lit. male slave [*chén*] and female slave [*qiè*]). However, this original sense had long become obsolete, and *chéngqiè* shifted to its feminine usage in as early as Eastern Han dynasty (25–220 C.E.). See Xia (2018) for a detailed discussion.

titles. As for *chénqiè*, the catchphrase *Chénqiè zuò-bú-dào a!* ‘I really can’t do it!’¹⁴ is trendy among netizens of all genders and sexual orientations.

A word of caution is in order on the category of *běn-gōng*. Although its literal meaning “this-palace” makes it resemble English *this reporter*, *the present author*, etc., which are imposters à la Collins and Postal, we must note that this Chinese term has no R-expression usage or flexible reference. Thus, it cannot be used in third-person cases like (14a), unlike English *this*-terms, as in (14b).

- (14) a. **Běn-gōng* *bǐ* *bié-gōng* *měi*.
 this-palace.3SG compared.to other-palace beautiful
 “Intended: This palace (= she) is more beautiful than other palaces (= other concubines).”
 b. **This reporter**_{3SG} is nicer than that one.

In other words, Chinese *běn-gōng* is a lexically fixed, idiomatic term of self-address, which makes it qualify as a noncanonical pronominal item in our criteria.¹⁵

2.2 Subtype II: Dialectal terms

The second subtype of noncanonical pronominal item we have identified in Vietnamese and Chinese involves dialectal terms that have made their way into the common language via mass media (e.g., TV programs) or the Internet. See Table 3 for some examples.

Table 3: Subtype-II noncanonical pronominal items in Vietnamese and Chinese

Vietnamese	<i>hǎn</i> ‘3SG’ (from Central and Southern dialects) ¹⁶
	<i>y</i> ‘3SG’ (from Northern dialects)
Chinese	<i>ǒu</i> ‘1SG’ (from Min/Yue Chinese)
	<i>é</i> ‘1SG’ (from Shaanxi Mandarin Chinese)
	<i>ǎn</i> ‘1SG’ (from Northern/Central Mandarin Chinese)
	<i>nóng</i> ‘2SG’ (from Shanghai Wu Chinese)
	<i>yā</i> ‘3SG’ (from Beijing Mandarin Chinese)

There are several distinctions between dialectal terms and revived ancient terms. First, while revived ancient terms are restricted to the first and the second person, dialectal terms also involve third-person items. For Vietnamese in particular, these terms are strictly 3SG (see Alves 2017 for more detail on the etymology of these). This is not surprising because dialectal terms are simply default pronouns in their original dialects. Second, unlike revived ancient terms, dialectal terms may not have gender restrictions at all, and in some cases not even preferences. Third, unlike many revived ancient terms (especially the royal ones), dialectal terms generally do not bear arrogant or bossy tones, so their pragmatic effects are of a different sort. Consider (15) for Vietnamese.

¹⁴ This is a quote from the highly popular TV series *Empresses in the Palace* and has gone viral via memes.

¹⁵ In this sense *běn-gōng* patterns more like *yours truly* and *muggins here* in English, which have no R-expression usage or flexible reference either, even though they are classified as imposters in Collins & Postal (2012).

¹⁶ Although *hǎn* has been documented in various parts of Vietnam, including the North (Cao 2014), the term was originally from the Southern dialect (see, e.g., Hoang 1989). It should also be noted that while *hǎn* is mostly used as a neutral 3SG in the South, it is often used with pragmatic effects in the North (addressed later in this section).

- (15) a. *Đi rồi không biết bao-giờ hấn mới về.* [Vietnamese, Southern speaker]
 go PRF NEG know when 3SG PRT return
 “(He)’s gone. (I) don’t know when s/he would return.”
- b. *Không biết thì hỏi hấn thử.* [Southern speaker]
 NEG know then ask 3SG try
 “If (you) don’t know then try asking him/her.”
- c. *Chính hấn cục vàng của tui.* [Northern speaker]
 precisely 3SG CLF gold POSS 1SG
 “It’s precisely him, my piece of gold.”
- d. *Lại nhớ hấn à?* [Northern speaker]
 again miss 3SG Q
 “Are (you) missing him again?”
- e. *Mèo nhà tao đã bắt y sáng nay rồi.* [standard variety, Internet language]
 cat home 1SG PST catch 3SG morning DEM PRF
 “My cat caught it this morning.”
- f. *Tôi nói lời yêu y, nhưng sao y không hiểu?* [standard variety, Internet language]
 1SG say word love 3SG but why 3SG NEG understand
 “I said loving words to him, but why hasn’t he understood?”

First, the term *hấn* is specific to Central and Southern dialects of Vietnamese, is a regional variant of the standard 3SG *nó*, and has entered the standard variety due to dialect contact. Note that although *hấn* is a neutral term in the original dialect, as in (15a–b), it is often used to sound cute/funny by speakers of other varieties. In (15c), for example, the speaker of Northern Vietnamese uses *hấn* as an endearing term to refer to her baby, who she considers “a piece of gold” in her possession.¹⁷ Similarly, another Northern Vietnamese speaker in (15d) uses *hấn* to refer to their interlocutor’s boyfriend, who is presumably being missed by the interlocutor. The use of *hấn* in both contexts sounds more fun and cuter than the standard variant *nó*.

The next Vietnamese item in this subtype is *y*, which is originally and mainly used in Northern dialects to refer to a male criminal. The term is therefore formal, but due to crossdialectal contact it has now become more widely used online as a jokingly serious pronominal form. In (15e), for example, *y* is used to refer to a mouse (who is in this sense cast as a criminal), which makes the sentence much funnier. Similarly, the speaker in (15f) uses *y* to complain about her crush, who has not returned her affection. Like *hấn*, the use of *y* in these contexts brings about some dramatic comic effects. Unlike *hấn*, however, there is a strong preference for a masculine interpretation particularly when the referent of *y* is human. When the referent is nonhuman, *y* can in principle be used neutrally (e.g., in (15e), we do not know whether the mouse is male or female).

The Chinese inventory for this subtype is again more diverse. Due to space limitations, we restrict our detailed description to only three of the Chinese terms from Table 3. All examples in (16) are from Common Mandarin produced by Weibo users.

¹⁷ *Cục vàng* ‘a piece of gold’ is an idiomatic expression in Vietnamese which is most often used by parents to refer to their precious children.

- (16) a. *Ǒu zhè-ge yǎn-zhū-zi hǎo-kàn.* [Mandarin]
 1SG this-CLF eye-ball-DIM good-look
 “My eyes are good-looking.”
- b. *Wèishénme Ǒu de xīn-dòng nánshēng hái bù chūxiàn?*
 why 1SG POSS heart-move guy still not appear
 “Why hasn’t the guy I will fall in love with shown up yet?”
- c. *É dìyī-cì pèngdào chǎo-wán jià jǔbào jiā gào-lǎoshī de.*
 1SG first-time encounter argue-finish quarrel report plus tell-teacher NMLZ
 “This is the first time I have ever encountered someone who reports the other person to teachers after quarreling with them.”
- d. *É dì guīmì jìngrán bǎ é zuì-ài de*
 1SG POSS best.female.friend go.so.far.as.to DISP 1SG most-love REL
xiāngshuǐ yòng-zuò kōngqì-qīngxīn-jì!
 perfume use-as air-freshen-agent
 “My bestie outrageously used my favorite perfume as air freshener!”
- e. *Nǐ hé zhǔrén shuō yí-jù ràng tā shuān shéng, yā zhāng-kǒu*
 2SG with owner say one-CLF let 3SG tie rope 3SG.OFF open-mouth
jiù xiàoxī xiǎo hé nǐ shuō: “wǒ jiā gǒu bù yǎo rén.”
 just giggle with 2SG say 1SG house dog not bite person
 “You ask the dog owner to tie their dog up, and they_{3SG} (= that asshole) just giggle and tell you: ‘My dog does not bite.’”
- f. *Nǐ yuè gēn yā shuō hǎo-tīng-de yā yuè lái-jìn.*
 2SG the.more with 3SG.OFF say good-listen-NMLZ 3SG.OFF the.more come-strength
Yíshànglai bǎ yā mà-xiaqu... hái néng yǒu diǎn yòng.
 right.at.the.beginning DISP 3SG.OFF scold-down still can have some use
 “The more kind words you say to them_{3SG} (= that asshole), the more shameless they_{3SG} (= that asshole) become. It would be more effective if you simply swear back and tell them_{3SG} (= that asshole) to get out of the car right away.”

The three terms in (16) have respectively been borrowed from Min/Yue Chinese (*Ǒu*), Shaanxi Mandarin Chinese (*é*), and Beijing Mandarin Chinese (*yā*). The first item, *Ǒu*, is a regional variant of Mandarin *wǒ* and (re)entered Common Mandarin due to netizens’ mocking of the dialectal pronunciation. According to Chen (2009:215), it was the most popular mutant personal pronoun online in the noughties. Perhaps due to its initial role as a mocking term, *Ǒu* sounds funny and cute and is often used by netizens who want to appear jolly and adorable. The current usage of *Ǒu* is no longer for mocking purposes. Thus, the netizen in (16a) happily posts about her satisfaction with the look of her eyes without the intention to mock anyone, and the netizen in (16b) laments her single status with a puppy-face-like tone. In both sentences, the use of *Ǒu* instead of the default 1SG *wǒ* makes the speakers sound more likable and less boastful or whiny.

The second item, *é* (sometimes rendered as *è* or *ngè*) became widely known in the noughties via the popular TV series *My Own Swordsman*, in which the leading actress spoke Shaanxi Mandarin throughout the eighty episodes. Due to the comedic nature of that show and the fussy personality of its main character, the term has subsequently gained a jokingly fussy tone in Internet language. Thus, the netizen in (16c) is making a fuss about the base behavior of a tattletale student, and that in (16d), about her best friend’s inadequate use of her perfume. No such fussy tone would be present if the default 1SG *wǒ* were used instead. In addition, in (16d) the possessive marker *dì* is also borrowed from Shaanxi Mandarin, whose Standard Mandarin counterpart is *de*. This again reflects the stylistic agreement mentioned in footnote 6.

A special note is in order concerning the Beijing Mandarin term *yā*, which is originally a highly vulgar expression meaning “child of a girl with no recognized marital status”. It is etymologically short for *yātou-yǎng-de* ‘low.status.girl-raise-NMLZ’, but nowadays this literal meaning is obsolete, and *yā* is

mainly used as an offensive suffix attached to pronouns and demonstratives (e.g., *nĩ-yā* ‘2SG-OFF’, *nèi-yā* ‘that.person-OFF’).¹⁸ That said, it has developed a stand-alone pronominal usage as well. Crucially, in this usage, it can only be interpreted as 3SG (gender neutral). Thus, the netizens in (16e–f) respectively complain about the bad behavior of a dog owner and that of a taxi customer. Since *yā* has no synchronic R-expression usage related to its pronominal usage¹⁹ and has a lexically fixed person index, we treat it as a noncanonical pronominal item instead of an imposter.

2.3 Subtype III: Creative online coinages

The third subtype of noncanonical pronominal item we have identified in Vietnamese and Chinese involves items that do not fall in the previous two subtypes. These are mainly creative coinages on the Internet and so in a sense “native” to Internet language. Since such online coinages are not based on any particular type of source, their lexical materials are miscellaneous or even totally novel. See Table 4 for some examples.

Note that all three Vietnamese examples in Table 4 are second-person terms, which carry different pragmatic effects due to the lexical materials they recycle. See (17) for some real-life examples.

Table 4: Subtype-III noncanonical pronominal items in Vietnamese and Chinese

	<i>cung</i> ‘dear.2SG’
Vietnamese	<i>con-quỷ</i> ‘devil.2SG’
	<i>người-đẹp</i> ‘beautiful.person.2SG’
	<i>qīn</i> ‘dear.2SG’
Chinese	<i>běn-lū/běn-lú/běn-lǔ</i> ‘this-loser.1SG’
	<i>lúnjiā</i> ‘others.1SG’

- (17) a. ***Cung*** *muốn gì từ anh nào?* [Viet.]
 dear.2SG want what from 1SG.MASC AFFECT
 “What do you (= dear) want from me?”
- b. ***Con-quỷ*** *đang làm gì đó?*
 devil.2SG PROG do what DM
 “What are you (= devil) doing?”
- c. *Cám-on người-đẹp đã mở hàng.*
 thank beautiful.person.2SG PST open shop
 “Thank you (= beautiful person) for being the first customer today!”

All three boldfaced terms in (17) are creative coinages by market sellers, which are now widely used thanks to online marketing. Similar to noncanonical pronominal items in the other two subtypes, those in Subtype III encode special pragmatic effects too. For example, the uses of *cung* and *con-quỷ* as 2SG terms in (17a–b) sound deliberately cute and friendly (and possibly a little flirtatious), while the use of *người-đẹp* in (17c) is flattery/fashionable. These terms are very creative and have no fixed lexical sources. Similarly, see (18) for some real-life examples of Chinese terms in this subtype.

¹⁸ A *de* can be optionally added to these terms (e.g., *nĩ-yā-de*), which is a residue of the nominalizer in the full form.

¹⁹ The character for *yā* (丫) does record other meanings too, such as “branch, twig” or more generally any Y-shaped object like the front part of a foot, but those are irrelevant to the pronominal *yā*. Thus, such polysemy is qualitatively different from that in cases like “teacher” as an R-expression and “teacher” as a term of address.

- (18) a. *Qīn* kěyǐ shōucáng wǒ-men de diànpù hé liànjiē... [Mandarin]
 dear.2SG can save 1SG-PL POSS store and link
xūyào de shíhòu zài liánxì ne.
 need REL time again contact EMP
 “Dear (= you) can save our store and link and contact us again when you have need.”
- b. *Suīrán běn-lú méi qián méi míng, dàn wǒ ānquán*
 though this-loser not.have money not.have fame but 1SG safety
yìshì yīliú.
 awareness first-class
 “Although this loser (= I) has no money or fame, I have first-class safety awareness.”
- c. *Lúnjiā zhēndeshì huīcháng huīcháng xǐhuān máo bùyì o.*
 others.1SG really very very like Mao Buyi EXCL
 “Others (= I) really like Mao Buyi very very much!”

Qīn originated on the shopping website Taobao as a friendly term of address between sellers and customers. Thus, in (18a) the Taobao seller says “dear can...” instead of “you can...” to encourage the customer to save their online store. *Běn-lú* and its tonal variants are all coined by combining the deictic *běn* ‘this’ and the first syllable of English *loser*. It has the pragmatic effect of self-mocking.²⁰ Thus, in (18b) the netizen self-mockingly takes pride in his safety awareness despite his poor status. *Lúnjiā* is a deliberately distorted variant of *rénjiā* ‘others’ and sounds cute and jocular when used as a 1SG term. Thus, in (18c) the netizen expresses her obsession with the pop singer Mao Buyi. Also note the form *huīcháng* in (18c), which is a distorted variant of the degree adverb *fēicháng* ‘very’ and adds to the cuteness of the utterance. This is another instance of the aforementioned stylistic agreement (see footnote 6).

Since this is the first systematic documentation of such creative pronominal coinages to our knowledge, we want to give a bit more detail on the above terms to justify our identification of them as noncanonical pronominal items. First, we are aware that *qīn* has R-expression usages, partly due to the versatility of its lexical root, which can mean “parent, kin (n.)”, “intimate, dear (adj.)”, “kiss (v.)”, etc. Moreover, one of its R-expression usages is closely related to its pronominal usage. Thus, one can friendly refer to someone (a third person) as *qīn*, as in (19).

- (19) *Wēibó hái yǒu qīn zài ma?* [Mandarin]
 Weibo still have dear.N be.at Q
 “Are there still anyone_{friendly} on Weibo?”

However, we do not treat *qīn* as an imposter because its pronominal usage has evidently developed from its term-of-address usage, which is also its predominant usage on Taobao (see, e.g., Deng 2012 and Liu 2012). In fact, previous studies rarely mention the usage exemplified in (19), which suggests that it might be a more recent development from either the pronominal or the term-of-address usage. In any event, the 2SG *qīn* is not a pronominally used R-expression in nature and thus does not fit the canonical situation of imposters.

Second, just like the Subtype-I term *běn-gōng* ‘this palace.1SG’, *běn-lū* also contains a deictic *běn*, but we treat it as a noncanonical pronominal item instead of an imposter because it too can only refer to the speaker (but not a third person) and has no R-expression usage, as exemplified for *běn-gōng* in (14).

Third, the phonological distortion that has created *lúnjiā* (*rén*→*lún*) has brought along some interesting change to its syntactic status. On the one hand, while both *rénjiā* and *lúnjiā* can be used as 1SG terms, only *rénjiā* has a separate 3SG usage (i.e., “others”), hence the ambiguity of (20a). By contrast, *lúnjiā* can only refer to the speaker, hence the ungrammaticality of the 3PL reading in (20b).

²⁰ This term, especially its *běn-lū* variant, is mainly used by males, since the Chinese character usually adopted to represent *lū* (撻) also means “(male) masturbate”, which further adds to the self-mocking effect of the term.

- (20) a. *Bié zhuāng le! Rénjiā yòu bú shì shāzi.* [Mandarin]
 don't pretend CRS others.3PL/1SG EMP not COP idiot
 “Stop pretending! Others (= they/I) are not idiots.”
- b. *Bié zhuāng le! Lúnjiā yòu bú shì shāzi.*
 don't pretend CRS others.*3PL/1SG EMP not COP idiot
 “Stop pretending! Others (= *they/I) are not idiots.”

On the other hand, while *rénjiā* has a shy or embarrassed tone when used as a term of self-address, *lúnjiā* furthermore sounds adorable and cartoon-like. Thus, while (20a) sounds like real blaming (with an embarrassed tone in the 1SG reading), (20b) sounds like the speaker is just teasing the addressee.

2.4 Interim summary

Based on Vietnamese and Chinese data, we have identified and exemplified a new type of pronominal item that is emerging in the Internet era, which we have dubbed noncanonical. As we have shown, noncanonical pronominal items differ from both textbook default pronouns and imposters in nontrivial ways. Crucially, unlike imposters, they have lexically fixed referents and no common R-expression usage, and unlike default pronouns they have various pragmatic effects. Overall, though, they pattern almost identically to default pronouns in syntax except for their extragrammatical effects. So, we tentatively rename noncanonical pronominal items “noncanonical pronouns” and give them the following working definition:

(21) **Noncanonical pronouns** are syntactically well-behaved pronouns with extragrammatical effects.

Moreover, the extragrammatical effects in noncanonical pronouns are not associated with conventional sociolinguistic factors like the relationship or relative hierarchical status between speakers and addressees.²¹ In fact, most examples we have given are not even from interpersonal communication but from online posts. Rather, the extragrammatical (e.g., register, tone) effects we have observed are more typically associated with the mind-sets and personalities of individual netizens themselves. To illustrate, adjectives we have used to describe the special effects of Vietnamese and Chinese noncanonical pronouns include the following:

- (22) joking, jocularly arrogant, fun, funnily bossy, jokingly pretentious, funnily flirtatious, miserable, cute, jokingly serious, dramatic, jolly, adorable, puppy-face-like, jokingly fussy, offensive, deliberately cute, friendly, flattery, fashionable, self-mocking, cartoon-like, teasing

These descriptions are highly compatible with the Internet register. For instance, while it would sound bizarre or even off-putting if an adult keeps talking in a dramatic, cute, or cartoon-like tone in reality, this is totally fine and acceptable on the Internet. In a similar vein, while gender does play a role in regulating the use of noncanonical pronouns, at least in Chinese, it is apparently one's gender identity (or sexual orientation) rather than their biological sex that guides their choices of noncanonical terms. This is another state of affairs increasingly normal in contemporary Chinese society, especially on the Internet.

Perhaps due to the unique features of the Internet as a modality of communication and the somewhat similar technological context it has endowed netizens around the world with, we have observed striking similarities in Vietnamese and Chinese noncanonical pronouns not only in their usage but also in their lexical sources. Specifically, for both languages we have identified three major types of noncanonical pronouns based on their evolution pathways: revived ancient terms, dialectal terms, and creative online coinages. Also, for certain terms (e.g., *trâm*, *ái-phi*), we have even observed mass-media-based crosslinguistic borrowing.

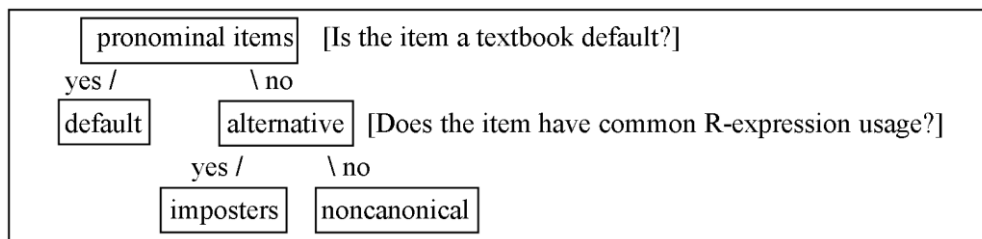
²¹ So noncanonical pronouns are different from the kind of interpersonal-relationship- or social-hierarchy-based pronominal items typically seen in East and Southeast Asian languages like Japanese, Korean, and Thai.

On the other hand, Vietnamese and Chinese show two main differences in their noncanonical pronouns. First, they differ in the sizes of their respective inventories, with Chinese having more actively used terms in almost every subtype. Second, they differ in the person/gender/number propensities of certain terms or even entire subtypes. For instance, Vietnamese terms like *trăm* (male) and *ái-phi* (male to female) show rather strict gender-based usage, and both Subtype-II and Subtype-III terms in Vietnamese show person/number restrictions (to 3SG and 2SG respectively). Whether these restrictions are categorical or due to the limited size of our data set requires further investigation, but our observation so far suggests that while Vietnamese has a more developed and less restricted imposter system than Chinese (see footnote 4), Chinese has a more developed and less restricted noncanonical pronoun system than Vietnamese. What this contrast means is an intriguing point of future research, but in the rest of this article we focus on the formal-grammatical status and syntactic representation of noncanonical pronouns.

3. Taxonomy

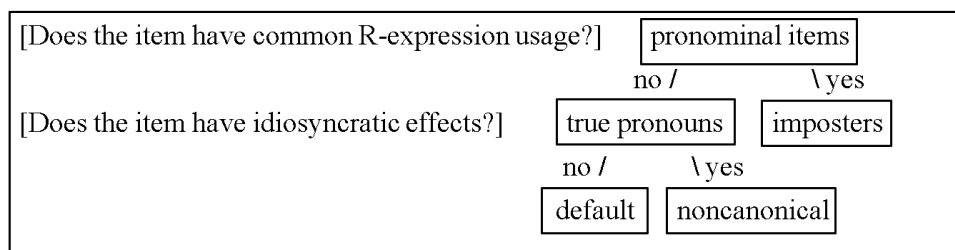
Before presenting our formal syntactic analysis, we first make a brief remark on the taxonomy of pronominal items. At first sight, an intuitive way of classifying pronominal items is to build on the conventional textbook system. Since textbook pronouns are deemed standard or default, nontextbook pronominal items could be termed alternative or nondefault. Then, under the alternative class, we further set two subclasses: imposters and noncanonical pronouns. We illustrate this taxonomy with the diagram in Figure 1 and call it the usage-based taxonomy, since properties like conventional/nonconventional and canonical/noncanonical are from the perspective of language use. Also, to give the taxonomy a bit more systematicity, we recast the two binary-branching nodes in terms of two yes-no questions.

Figure 1: A usage-based taxonomy of pronominal items



However, from a grammatical perspective the taxonomy in Figure 1 is obviously flawed, because even though imposters and noncanonical pronouns are both nondefault pronominal items, grammatically speaking noncanonical pronouns resemble default pronouns to a much greater extent than they resemble imposters. So, in a grammatically more precise taxonomy default and noncanonical pronouns should be grouped under the same class at some level. To this end, we propose the syntactically based taxonomy in Figure 2, which is also accompanied by two yes-no questions defining the two binary-branching nodes.

Figure 2: A syntactically based taxonomy of pronominal items



Incidentally, the syntactically based taxonomy is also more desirable from an acquisitional perspective, since the two questions in Figure 2 can both be easily answered based on the primary linguistic data a child has direct access to, whereas the first question in Figure 1 hinges on more sophisticated knowledge about the world, more exactly about language textbooks and reference grammars. According to the

principles and parameters approach to linguistic variation and diversity (Chomsky 1981), especially its neo-emergentist incarnation (Biberauer 2017), children acquire grammatical knowledge based on positive evidence in the significant acquisitional input (e.g., high-frequency recurring forms and collocations in everyday adult speech). So, a certain grammatical phenomenon is acquired early (or at least is acquirable) if it regularly and saliently exists in the primary linguistic data and therefore can be easily detected by the acquirer. Building on this theoretical background, a further prediction Figure 2 makes is that in languages where imposters are a regular part of the grammar, such as Vietnamese, imposters are acquired earlier than default pronouns.²² We leave the verification of this claim to future research.

4. A formal syntactic analysis

After laying out the comparative data (§2) and the syntactically based taxonomy (§3), in this section we propose a formal syntactic analysis of noncanonical pronouns within the minimalist program (Chomsky 1995 et seq.), more exactly within the generalized root syntax theory put forth in Song (2019). The purpose of this analysis is threefold: (i) to explain why noncanonical pronouns behave the way they do in a formally explicit way, (ii) to improve current syntactic theory of pronouns and make it empirically more adequate, and (iii) to tentatively explain why noncanonical pronouns have restricted crosslinguistic distribution. Before presenting our particular analysis (§4.2), we first introduce the theoretical background of pronominal syntax (§4.1).

4.1 Pronouns in generative syntax

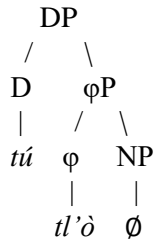
As we mentioned in §1, previous syntactic studies of pronouns have mainly focused on default pronouns, especially those that constitute exponents of phi features. Within the Chomskyan school, since the proposal of the DP hypothesis (Abney 1987), default pronouns have been associated with the category D (or its more elaborate equivalents) in one way or another. Thus, Abney (building on Postal 1966) puts pronouns at the D head position and argues that they project DPs on their own (i.e., without specifier/complement elements). Then, with the popularization of the split-functional-projection idea, which was initiated by Pollock's (1989) split-IP hypothesis and Rizzi's (1997) split-CP hypothesis in the verbal domain, many authors have proposed elaborate hierarchical structures for the nominal domain as well (e.g., Ritter 1995, Cardinaletti and Starke 1999, Neeleman and Weerman 1999, Borer 2005, Ritter and Wiltschko 2019). In relation to pronouns, quite a few authors have put forth the idea that they may correspond to different parts or zones in the nominal tree.

For instance, Ritter (1995) proposes two kinds of pronouns, which respectively occupy D and Num. Déchaine and Wiltschko (2002) propose three kinds of pronouns, which they name pro-DP, pro-φP, and pro-NP. Like Ritter, Déchaine and Wiltschko also let pronouns occupy head positions. But that is not the only solution. There are also researchers who propose that pronouns realize entire nominal tree(let)s (e.g., Weerman and Evers-Vermeul 2002, Neeleman and Szendrői 2007). As Neeleman and Szendrői suggest, this approach is more natural in late spell-out frameworks like distributed morphology (Halle and Marantz 1993 et seq.), in which syntactic computation operates on formal features, whose phonological realization is only dealt with at the syntax-phonology interface. Since such a computation-before-pronunciation view is also more generally endorsed in the minimalist program, we accept it as a background assumption without further discussion. That is, we assume that both terminal and nonterminal nodes may be spelled out by phonological units. This means that what looks like a single word or even morpheme on the surface may be a complex hierarchical structure in the underlying syntax and that generative syntax can well handle this type of phenomenon (by means of nonterminal spell-out).

Special attention needs to be paid to two recurring issues in previous generative analyses of pronouns. The first issue is the division of labor in the elaborate structure of pronouns. Take Déchaine and Wiltschko's pro-DP structure for example.

²² However, Figure 2 does not predict that default and noncanonical pronouns are acquired simultaneously, because the latter are a novel phenomenon in Internet language and not yet part of the grammatical knowledge relevant for first language acquisition or parameter setting.

- (23) Pro-DP (illustrated by Halkomelem *tú-tl'ò* ‘DET-3SG’; Déchaine and Wiltschko 2002:412)

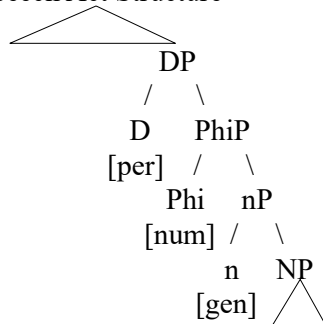


In this structure, both D (for definiteness) and φ (for person/number) are overt, but NP is empty. Déchaine and Wiltschko hypothesize that the NP position may be either overt or null, and that when it is filled with lexical material we get a normal [determiner noun] phrase, such as (24).

- (24) *Tl'ó-cha-l-su qwemcíwe-t [thú-tl'ó q'ami]_{ARG.}* [Halkomelem]
 then-FUT-1SG-so hug-TRANS DET.FEM-3SG girl
 “Then I’m going to hug that girl.” (Galloway 1993:174, via Déchaine and Wiltschko 2002:412)

However, Déchaine and Wiltschko do not specify what the semantic contribution of NP is when it is null as in (23). A similar scenario occurs in Ritter and Wiltschko (2019), where an even more elaborate nominal domain is proposed, which has three zones—a lexical, a functional, and an interactional zone—each subsuming a number of categories. Take their analysis of German *du* ‘2SG’ in (25) for example.

- (25) German pronoun *du* ‘2SG’ (Ritter and Wiltschko 2019:3)
 Speech Act Structure



The authors do not discuss the spell-out procedure, but their theory is compatible with a late spell-out approach (Wiltschko, p.c.), so we can assume that the phonological unit *du* somehow realizes the whole tree in (25). Again, we see concrete semantic contributions of each syntactic category except NP, though it presumably has to be there since otherwise the tree is not lexically grounded.

Lexical grounding is precisely the second issue in previous syntactic studies that we would like to invite readers to pay attention to. While both (23) and (25) have a functional-above-lexical scaffolding, as is standardly assumed in current generative syntax, this is not always the case in earlier studies of pronouns. For example, the following trees from Abney (1987) and Ritter (1995) have no lexical bases (i.e., NP).

- (26) a. Abney (1987:284) b. Ritter (1995:419) c. Ritter (1995:421)
- | | | |
|-------------------------|---|--|
| DP

D

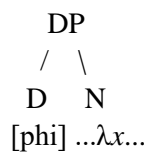
we | DP

D
[person]
[number]
[gender] | DP
/ \
D NumP
[definite]
[person] Num
hu/hi/hem/... |
| | (1 st /2 nd -person pronouns in Hebrew) | (3 rd -person pronouns in Hebrew) |

It is unclear how such trees can be generated in minimalism, in which functional categories by definition build on or extend lexical categories. In fact, in some minimalist theories, such as distributed morphology, the lexical grounding even provides the categorial feature for the entire tree (Roberts 2019 has a similar idea).

The reason why we want to draw readers’ attention to the above two issues is because they are prevalent in the literature and deserve reflection if we want to guarantee the implementability of particular theories in the minimalist program (e.g., in the fashion of Collins and Stabler 2016). That said, we do not think the theories cited above are inherently problematic. Rather, they are just of a lower granularity level, where certain details are glossed over. As a granularity-raising move, we find an idea in Harbour (2016) helpful. Abstracting away from the technical details, Harbour’s core idea is that the individual variable (a semantic primitive), which any further nominal semantic function (be it person, number, or whatever else) relies on, is introduced *before* the further formal features are introduced. He lets the individual variable be introduced at the N level and the further features be introduced at the D (or split-D) level, as in (27).

- (27) Semantic division of labor in the nominal structure (adapted from Harbour 2016:77)



The overall effect of (27) is that the phi features in D together yield a particular set of individuals, which the individual variable in N ranges over.²³ Eventually, some other features in the nominal structure pin down the referent and assign it to the individual variable *x*. Acquaviva (2019) independently proposes a similar idea in an even more fine-grained theory, where the N part is further decomposed into a nominalizer *n* and a root in the sense of distributed morphology, and it is the nominalizer that introduces the individual variable that serves as the semantic grounding of the whole nominal phrase.

Assuming the above high-granularity details in the background, we can now safely take any of the aforementioned theories of default pronouns as a point of departure without worrying about shaky foundations. In fact, which specific DP structure we adopt is immaterial to our own analysis as long as it does not suffer from foundational problems and has the necessary components to derive a syntactically well-behaved default pronoun. Therefore, when the DP-internal structure is inconsequential, we simply use the very-low-granularity label DP_{pro} to indicate a pronominal DP (as Aldridge 2021 does in her study on Old Chinese pronouns).²⁴

Having seen how default pronouns are typically analyzed in generative syntax, we now build our analysis of noncanonical pronouns on top of that, because as we mentioned in §2.4, noncanonical pronouns are just default pronouns equipped with idiosyncratic extragrammatical effects. Since they

²³ Harbour’s tree is highly abbreviatory, which cannot be taken at face value for issues like labeling (Chomsky 2013 et seq.). We assume that the tree in (27) automatically gets the label DP once the omitted details are filled back.

²⁴ We use DP as an umbrella label for the entire (pro)nominal structure, which abbreviates not only the lexical and functional zones but also the Wiltschkovian interactional zone if that is present.

behave like default pronouns in syntax, they must involve the same sort of syntactic structure as default pronouns, whatever that structure is. What our analysis mainly serves to explain, therefore, is how the idiosyncratic effects are formally associated with that structure.

4.2 Generalized root syntax

To begin with, note that the special effects of noncanonical pronouns come from the particular terms themselves, more exactly from the idiosyncratic content associated with their lexical materials, rather than from the context. This is because they show the same effects in all examples. For instance, Vietnamese *trẫm* and Chinese *zhèn* (both 1SG) sound jokingly arrogant wherever they appear, and they have this effect precisely because they were once used by emperors. Similarly, Vietnamese *cung* and Chinese *qīn* (both 2SG) sound deliberately friendly, either in commercial discourses or not, and this effect has to do with the terms' original lexical meaning “dear”. An even more interesting case is Chinese *lúnjiā* ‘others.1SG’, which not only takes some lexical material (*rénjiā* ‘others.1SG/3SG/3PL’) but also distorts it (*rén*→*lún*), and the distortion in turn serves to distinguish the new noncanonical usage from the old imposter usage. The key mechanism involved in all these cases of noncanonical pronouns is thus the recycling of existing lexical material for new grammatical purposes (i.e., a kind of grammaticalization). This process is evident in all three subtypes of noncanonical pronouns in §2. It is also a manifestation of a fundamental strategy in human language and cognition, which Biberauer (2017) terms “maximize minimal means” (MMM). In Biberauer’s (2017:41) words, MMM is both “a generally applicable learning bias harnessed by the acquirer during acquisition” and “a principle of structure building” that “facilitat[es] the kind of efficient computation and ... the self-diversifying property that allows human language to be the powerful tool that it is” (see Biberauer 2011 et seq. for more background on this line of thought).

Song (2019) develops a “generalized root syntax” to tackle half-grammatical-half-lexical vocabulary items, of which the noncanonical pronouns studied here are a specific instance. Song’s theory is an extension of the root theory in distributed morphology (hence its name), where content words like nouns and verbs are decomposed into a functional part, called the categorizer, and a purely lexical part (which does not have a syntactic category), called the root. Thus, the noun *dog* is analyzed as [*n* √DOG], and the verb *run* is analyzed as [*v* √RUN]. The idea is that all that participates in formal computation is essentially functional-categorial when the representation is fine-grained enough, while idiosyncratic information like lexical sound/meaning and encyclopedic knowledge is sealed in a syntactically inert capsule that is only opened when the syntactic representation is sent to the phonological/semantic interfaces for interpretation. This lexical decompositional practice pushes syntactic methods to the traditional morphological arena, so distributed morphology is also known as a “syntax all the way down” approach.²⁵

While accepting the lexical decompositional approach of distributed morphology, Song points out that its particular view on root categorization is flawed, for it stipulates that only traditional lexical categories can serve as categorizers, but that assumption leads to theory-internal contradiction under close scrutiny (see Song 2019:102 for details). Since from a formal perspective the categorization procedure just serves to equip the otherwise inert root with a syntactically active shell, logically speaking any functional category can do the job, and the specialness of traditional lexical categories (i.e., the little *x* categorizers in distributed morphology) merely lies in their bare-predicate-making/typing semantics. That is, they introduce typed individual variables in the sense of Harbour (2016) and Acquaviva (2019):

²⁵ A caveat here is that distributed morphology does *not* predict that any root-categorizer merger can yield a legitimate vocabulary item (which is a common straw man in criticism of the framework). Rather, the interpretability of particular root-categorizer combinations is a matter of language-specific lexicalization (in a broad sense of the term), and root syntax merely offers a tool to structurally represent and analyze such lexically stored information. A biggest achievement of root syntax, in hindsight, is that it has pushed syntactic theory to a higher granularity level and thereby formalized a further aspect of regularity in human language (i.e., the very basic phenomenon of categorization). We deem this a significant step forward.

- (28) a. N-categorizer (“nominalizer”): introduces an entity-type individual variable
 b. V-categorizer (“verbalizer”): introduces an eventuality-type individual variable
 c. X-categorizer: introduces an x-type function

Thus, root categorization is more generally root support; namely, the enrichment of a functional category with some idiosyncratic information encapsulated in a root. For instance:

- (29) a. *dog* = [$n \sqrt{\text{DOG}}$] = an entity-type individual that is called /dɔg/, has four legs, can bark, etc.
 b. *run* = [$v \sqrt{\text{RUN}}$] = an eventuality-type individual that is called /rʌn/, involves leg-moving, etc.

From the perspective of syntacticosemantic computation, only the underlined information in (29) is relevant, while the rest merely supports this skeletal information and makes it suitable for postsyntactic purposes (describing the world, communication, etc.). In other words, the extragrammatical effects of content words is precisely their idiosyncratic root content. Extending this mechanism to X-categorizers, what we obtain is a functional category equipped with idiosyncratic extragrammatical effects; namely, a half-grammatical-half-lexical item. Song (2019) gives miscellaneous examples from Chinese to illustrate this, such as those in (30).

- (30) a. passive auxiliaries: *bèi* ‘lit. cover, suffer (neutral)’, *gěi* ‘lit. give (colloquial, negative)’
 b. classifiers: *jiàn* ‘lit. item (for clothes, etc.)’, *duǒ* ‘lit. flower (for flowers, clouds, etc.)’
 c. conjunctions: *hé* ‘lit. union (general purpose)’, *yǔ* ‘lit. accompany (formal, literary)’

These items can all be analyzed as a functional category supported by a root (e.g., [$\text{Voice}_{\text{PASS}} \sqrt{\text{BÈI}}$]). The categorial part determines their syntactic functionality, while the root part determines their extragrammatical effects and thereby conditions their real-life usage (e.g., the “and” in *Harry Potter and ...* is *yǔ* instead of *hé*).

Returning to noncanonical pronouns, the same analysis applies. Since we have established that the syntactic behavior of noncanonical pronouns is the same as that of default pronouns, we treat them as root-supported D_{pro} items. However, given the elaborate DP structures in §4.1, we next clarify three technical details to show how our analysis fits into the big picture of pronominal syntax.

4.3 Deriving noncanonical pronouns

First, following Nunes and Uriagereka (2000), Johnson (2003), and especially Zwart (2007 et seq.), we assume syntactic derivation to be multilayered. That is, derivational products of an earlier cycle/workspace can be used in a subsequent cycle, probably in an “atomized” fashion (Fowlie 2013). This is an inevitable state of affairs if we look closely at the assembling of syntactic trees. At each step of Merge, the object that is newly selected from the lexical (sub)array into a workspace by definition has not undergone any Merge step in that workspace (so it is a “minimal” category in a relative sense). While this is natural in the merger of a primitive category with an existing phrase, a question arises in the merger of a specifier or adjunct with a phrase: Where has the specifier/adjunct (which is also phrasal by definition) been derived? The only logical possibility is that it has been derived in another workspace before being selected and merged in the current workspace. The upshot is that “layered derivation” (Zwart’s term) must be a standard mechanism in minimalist syntax. We contend that this is also what happens in the derivation of noncanonical pronouns.

- (31) a. Workspace I b. Workspace II c. Workspace III
- | | | |
|---|---|---|
| $\begin{array}{c} \text{DP}_{\text{pro}} \\ / \quad \backslash \\ \text{D} \quad \text{NumP} \\ / \quad \backslash \\ \text{Num} \quad \text{NP} \\ \text{(atomize DP}_{\text{pro}}) \end{array}$ | $\begin{array}{c} \text{DP}_{\text{pro}} \\ / \quad \backslash \\ \text{DP}_{\text{pro}} \quad \checkmark \\ \text{(support DP}_{\text{pro}} \text{ with a root)} \\ \text{(atomize the root-supported DP}_{\text{pro}}) \end{array}$ | $\begin{array}{c} \text{VoiceP} \\ / \quad \backslash \\ \text{DP}_{\text{pro-}\checkmark} \quad \text{VoiceP} \\ / \quad \backslash \\ \text{Voice} \quad \text{VP} \\ \text{(merge the root-supported DP}_{\text{pro}} \text{ as a specifier)} \end{array}$ |
|---|---|---|

The above derivation involves three consecutive layers, each defined by a workspace. In (31a), a default pronominal phrase is built up in Workspace I. We use the DP-NumP-NP structure as an example, but as mentioned in §4.1, any workable DP structure is fine for current purposes. In (31b), the assembled and atomized pronominal DP is selected into Workspace II and merged with a root. This is the categorization step. As a result, the root is assigned the category D and the D category is associated with the idiosyncratic content in the root. Finally, in (31c) the root-supported pronominal DP is selected into Workspace III and merged as a specifier. The particular scenario here is Spec-VoiceP; namely, the subject of a transitive verb.

Second, the derivation in (31) does not hinge on the concrete content in the supporting root, since that is syntactically inert anyway. This means that miscellaneous lexical materials can be (re)used at the root slot, as long as it is properly lexicalized, with its original formal features (if any) reanalyzed as lexical features. As such, not only simple roots like $\checkmark\text{ZHÈN}$ but also derived roots like $\checkmark\text{ĀI-JĪĀ}$ can be used to support DP_{pro} . The key prediction here is that this recategorized *āijiā* can only be used as a pronoun but not as a common R-expression anymore, even though its original meaning clearly *is* R-expressional (i.e., “mourner”). As we observed in §2, this is generally true for noncanonical pronouns. Third, the tree in (31a) is just that of a default pronoun, which in principle can have its own overt form. For instance, if the pronoun is 1SG, then it in principle can be spelled out as the default *wǒ* in Chinese. Yet this never happens with noncanonical pronouns. That is, noncanonical pronouns do not allow appositive default pronouns, which sharply contrasts them with imposters, as in (32).²⁶

- (32) a. *Lǎoshī (wǒ) kuài yào shīqù wǒ de nàixìng le.* [Mandarin, imposter]
 teacher.1SG 1SG quick going.to lose 1SG POSS patience CRS
 “Teacher_{1SG} (I) is going to lose my patience.” (adapted from Want 2014:185)
- b. *Zhèn (*wǒ) kuài yào shīqù wǒ de nàixìng le.* [noncanonical pronoun]
 zhèn.1SG 1SG quick going.to lose 1SG POSS patience CRS
 “Zhèn_{1SG} (*I) is going to lose my patience.”

Intuitively, *zhèn* is just *wǒ* with an alternative pronunciation and some idiosyncratic effects, so saying *zhèn wǒ* is like saying “I I” in English, which is clearly ungrammatical. How can the syntactic derivation in (31) bear this out, though? The observation, in conditional terms, is the following:

- (33) i. If a functional category is root-supported, it assumes the root’s exponent (possibly distorted).
 ii. If a functional category is non-root-supported, it assumes its default exponent (if any).

There are two ways to explain (33). One way is to view it as a concomitant of the categorization procedure. Recall that the classic case of categorization in distributed morphology is that of content words. In this case the uncategorized root has no fixed pronunciation and only gets one when it is assigned a category. For instance, the English root $\checkmark\text{PERMIT}$ may be pronounced as /pəˈmɪt/ or /ˈpəˌmɪt/ depending on its category. This is even more evident in languages like Hebrew, where uncategorized roots cannot be vocalized at all (e.g., $\checkmark\text{K-T-B}$ ‘related to writing’, $\checkmark\text{L-M-D}$ ‘related to learning’). The

²⁶ While imposters may co-occur with appositive default pronouns in Chinese, this is impossible in Vietnamese.

theory for this in distributed morphology is that each step of categorization in syntax corresponds to a step of retrieving stored phonological/semantic information at the interfaces. And crucially, it is the [categorizer root] unit as a whole, not its subparts, that gets assigned that information. As such, it is plausible that each noncanonical pronoun has its own lexical entry, and when a root-supported DP_{pro} like (31b) is interpreted, it is assigned the stored pronunciation, which in most cases is the same as that of the recycled lexical material (again a manifestation of MMM) but may also involve certain distortion as in *lúnjiā*.

Another way to explain (33) is to invoke Kiparsky's (1973) elsewhere principle, which basically says that when a more general and a more specific rule are adjacent, the more specific rule applies. For instance, the past tense form of *go* is *went* instead of **goed* because the irregular rule is more specific. Neeleman and Szendrői (2007) use the elsewhere principle to explain radical pro drop. We leave out the details of their application for space limitations but merely cite a well-known implication of the principle that they list:

(34) All else being equal, the phonological realization of syntactic structures favors spell-out of a category C over spell-out of the categories contained in C. (Neeleman and Szendrői 2007:685)

This immediately explains why a root-supported functional category gets pronounced differently from a non-root-supported one—because in a tree like (31b) the spell-out rule that targets $[DP_{pro} \sqrt{\quad}]$ as a whole blocks the rule that targets DP_{pro} itself. We can reformulate (33) as follows to better reflect the elsewhere principle.

(35) i. If a functional category is root-supported, it assumes the root's exponent (possibly distorted).
ii. Elsewhere it assumes its default exponent (if any).

In any event, the different pronunciations of default pronouns and noncanonical pronouns, in spite of their partially identical underlying structures, well conform to independently motivated rules in generative syntax.

4.4 Crosslinguistic availability

We mentioned in §1 that noncanonical pronouns have a much more limited crosslinguistic distribution than imposters. Our foregoing analysis may explain why this should be the case. Song (2019:136–137) points out that root-supported heads are “analytic heads” because they have a very low category/morpheme-per-word ratio (respectively 1:1 and 2:1), where a word is understood as a morphophonologically freestanding unit. And Chinese-style semifunctional items (like those in (30)) are furthermore “analytic heads par excellence” because their category-per-word ratio and morpheme-per-word ratio are both 1:1, where a morpheme is understood in the traditional sense as a minimal sound-meaning pair. As such, by analyzing noncanonical pronouns in generalized root syntax, we automatically get the following prediction:

(36) Noncanonical pronouns are more common in highly analytic languages.

This may explain why we can easily find noncanonical pronouns in Vietnamese and Chinese but not in familiar European languages—both Vietnamese and Chinese are highly analytic languages, which have the right grammatical setting for root-supported heads to stably exist. Note that we are not claiming that high analyticity is solely defined by root support. In fact, if there is a high analyticity parameter at all, that is very likely to be codefined by a cluster of smaller parameters or grammatical settings (see Huang 2015 for a discussion). Here we are merely stating that a highly analytic language has a root-support-friendly setting. For instance, among others it may have scarce head movement, which means it has more standalone words than affixes, and root support just provides a convenient way to create standalone grammatical words. Of course, we need to check more highly analytic languages to verify (36), which we leave to future research.

5 Conclusion

In this paper, we studied a new type of pronominal item emerging on the Internet in Vietnamese and Chinese. First, we demonstrated that pronominal items of this new type, which we call noncanonical pronouns, are a separate category from both textbook default pronouns and imposters (§1). Then, we illustrated their real-life usage in detail (§2). Our investigation showed that noncanonical pronouns in the two Asian languages are similar not only in syntactic behavior but also in lexical sources and evolution pathways. Syntactically, both Vietnamese and Chinese noncanonical pronouns behave like default pronouns except that they have various extragrammatical effects, mostly reflecting the speaker's mind-set or personality. Lexically, both languages subsume three subtypes of noncanonical pronouns based on their evolution pathways: revived ancient terms, dialectal terms, and creative online coinages. After presenting the empirical facts, we briefly discussed the taxonomy of pronominal items and specified why we prefer a syntactically based taxonomy (§3).

In the theoretical part of the paper, we analyzed noncanonical pronouns in the theory of generalized root syntax (§4), which is an extension of the root theory in distributed morphology (a branch of the minimalist program). Specifically, we analyzed noncanonical pronouns in the schema $[DP_{\text{pro}} \checkmark]$, where DP_{pro} is a separately derived and atomized default pronoun structure and \checkmark is a purely lexical root supporting that structure. The root part may be constituted by terms (re)lexicalized from the three sources above. After the (re)categorization, the root-supported structure $DP_{\text{pro-}\checkmark}$ is selected into the main workspace and merged onto the main tree, where it behaves like a default pronoun in syntax but triggers idiosyncratic phonological and semantic properties at the interfaces, just as we have observed in Vietnamese and Chinese. Alongside our analysis, we clarified a number of theoretical and technical issues, such as how to use previous theories of pronominal syntax in a well-founded way, why cross-workspace or layered derivation must be allowed in generative syntax, and why default and noncanonical pronouns with the same pronominal structure can have totally different overt forms. Finally, we also tentatively explained why noncanonical pronouns have limited crosslinguistic availability in terms of the correlation between root support and high analyticity.

Due to limited scope, we have had to leave some interesting questions to future research, including but not limited to the acquisitional order predicted by our taxonomy in §3 and the availability of noncanonical pronouns in other highly analytic languages as predicted in §4. We also observed a quasi-complementary contrast between Vietnamese and Chinese in §2 concerning the distribution of noncanonical pronouns and imposters. So, as a future plan we would also like to further compare Vietnamese and Chinese imposters and look into questions such as why imposters are used more freely in Vietnamese (footnote 4), why they may co-occur with appositive default pronouns in Chinese but not in Vietnamese (footnote 26), and so on.

Abbreviations

1/2/3 = first/second/third-person

AFFECT = affectionate

ARG = argument

CLF = classifier

COP = copula

CRS = currently relevant state

DEM = demonstrative

DET = determiner

DIM = diminutive

DISP = disposal

DM = discourse marker

EMP = emphatic

EXCL = exclamative

EXP = experiential

FEM = feminine

FUT = future

HON = honorific

IMPERF = imperfective
INTS = intensifier
MASC = masculine
N = noun
NEG = negation
NMLZ = nominalizer
Num = number
OFF = offensive
PL = plural
POSS = possessive
PRF = perfective
PROG = progressive
PRT = particle
Q = question marker
REL = relative clause marker
RES = resultative
SG = singular
TEL = telic marker
TRANS = transitive

References

- Abney, Steven. 1987. The English noun phrase in its sentential aspect. MIT dissertation.
- Alves, Mark. 2007. Categories of Sino-Vietnamese grammatical vocabulary. *Mon-Khmer Studies* 37:217–229.
- Alves, Mark. 2017. Chinese influence on Vietnamese pronouns and terms of address and reference. Presentation at NACCL 29, June 16–18.
- Acquaviva, Paolo. 2019. Categorization as noun construction: Gender, number, and entity types. *Gender and noun classification*, ed. by Eric Mathieu, Miriam Dali and Gita Zareikar, 41–63. Oxford: Oxford University Press.
- Aldridge, Edith. 2021. The syntax of pronoun fronting in Late Archaic Chinese negated clauses. *Journal of East Asian Linguistics* 30.1:39–79.
- Biberauer, Theresa. 2011. In defence of lexico-centric parametric variation: Two 3rd factor-constrained case studies. Paper presented at Workshop on Formal Grammar and Syntactic Variation, Oct 28, Madrid.
- Biberauer, Theresa. 2017. Factors 2 and 3: A principled approach. *Cambridge occasional papers in linguistics*, vol. 10, ed. by Chechens Song and James Baker, 38–65.
- Borer, Hagit. 2005. *In name only. Structuring sense*, vol. 1. Oxford: Oxford University Press.
- Cao, Xuan Hai. 2014. Đại từ nhân xưng đích thực trong tiếng Thanh Hoá [De facto personal pronouns in the dialect of Thanh Hoa]. *Ngôn ngữ và Đời sống* 9.27:17–22.
- Cardinaletti, Anna; and Michal Starke. 1999. The typology of structural deficiency. *Clitics in the languages of Europe*, ed. by Henk van Riemsdijk, 145–233. Berlin: Mouton de Gruyter.
- Chen, Cuizhu. 2009. Hanyu rencheng daici kaolun [An examination of Chinese pronouns]. Central China Normal University dissertation.
- Chomsky, Noam. 1981. *Lectures on government and binding*. Berlin: Mouton de Gruyter.
- Chomsky, Noam. 1995. *The minimalist program*. Cambridge MA: MIT Press.
- Chomsky, Noam. 2013. *Problems of projection*. *Lingua* 130:33–49.
- Collins, Chris. (ed.) 2014. *Cross-linguistic studies of imposters and pronominal agreement*. Oxford: Oxford University Press.

- Collins, Chris; and Paul Postal. 2012. *Imposters: A study of pronominal agreement*. Cambridge MA: MIT Press.
- Collins, Chris; and Edward Stabler. 2016. A formalization of minimalist syntax. *Syntax* 19.1:43–78.
- Déchaine, Rose-Marie; and Martina Wiltschko. 2002. Decomposing pronouns. *Linguistic Inquiry*, 33.3:409–442.
- Deng, Da. 2021. Wangluo chengwei qin de yanjiu [A study of the online appellation qin]. Xiangtan University MA thesis.
- Feng, Shengli. 2010. Lun yuti de jizhi ji qi yufa shuxing [On mechanisms of register system and its grammatical property]. *Studies of the Chinese Language* 5:400–412.
- Fowlie, Meaghan. 2013. Multiple multiple spellout. *Challenges to linearization*, ed. by Theresa Biberauer, 129–169. Berlin: Mouton de Gruyter.
- Galloway, Brent. 1993. *A grammar of Upriver Halkomelem*. Berkeley: University of California Press.
- Halle, Morris; and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. *Essays in linguistics in honor of Sylvain Bromberger*, ed. by Kenneth Hale and Samuel Jay Keyser, 111–176. Cambridge MA: MIT Press.
- Harbour, Daniel. 2016. *Impossible persons*. Cambridge MA: MIT Press.
- Hoang, Thi Chau. 1989. *Tiếng Việt trên các miền đất nước: phương ngữ học* [Vietnamese in the various areas of the motherland: a dialectological study]. Hà Nội: NXB Khoa học Xã hội.
- Huang, C-T. James. 2015. On syntactic analyticity and parametric theory. In *Chinese syntax in a cross-linguistic perspective*, ed. by Audrey Li, Andrew Simpson and Wei-Tien Tsai, 1–48. Oxford: Oxford University Press.
- Johnson, Kyle. 2003. Towards an etiology of adjunct islands. *Nordlyd* 31.1:187–215.
- Kiparsky, Paul. 1973. “Elsewhere” in phonology. *A festschrift for Morris Halle*, ed. by Stephen Anderson and Paul Kiparsky, 93–106. New York: Holt, Rinehart and Winston.
- Liu, Yang. 2012. Taobaoti zhong qin de yanjiu [A study of qin in Taobaospeak]. Hebei Normal University MA thesis.
- Neeleman, Ad; and Kriszta Szendrői. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry*, 38.4:671–714.
- Neeleman, Ad; and Fred Weerman. 1999. *Flexible syntax: A theory of case and arguments*. Dordrecht: Kluwer.
- Nunes, Jairo; and Juan Uriagereka. 2000. Cyclicity and extraction domains. *Syntax* 3.1:20–43.
- Pollock, Jean-Yves. 1989. Verb movement, universal grammar, and the structure of IP. *Linguistic Inquiry* 20.3:365–424.
- Postal, Paul. 1966. On so-called pronouns in English. *Modern studies in English*, ed. by David Reibel and Sanford Schane, 201–223. Englewood Cliffs, N.J.: Prentice-Hall.
- Ritter, Elizabeth; and Martina Wiltschko. 2019. Nominal speech act structure: Evidence from the structural deficiency of impersonal pronouns. *Canadian Journal of Linguistics*, 64.4:709–729.
- Ritter, Elizabeth. 1995. On the syntactic category of pronouns and agreement. *Natural Language & Linguistic Theory* 13:405–443.
- Rizzi, Luigi. 1997. The fine structure of the left periphery. *Elements of grammar: A handbook of generative syntax*, ed. by Liliane Haegeman, 281–337. Dordrecht: Kluwer.
- Roberts, Ian. 2019. *Parameter hierarchies and universal grammar*. Oxford: Oxford University Press.
- Song, Chenchen. 2019. On the formal flexibility of syntactic categories. University of Cambridge dissertation.
- Wang, Chyan-An. 2014. Mandarin pseudo-imposters. *Cross-linguistic studies of imposters and pronominal agreement*, ed. by Chris Collins, 173–195. Oxford: Oxford University Press.
- Weerman, Fred; and Jacqueline Evers-Vermeul. 2002. Pronouns and case. *Lingua* 112:301–338.

Xia, Xianpei. 2018. Nüzi zicheng chenqie de yuanliu [On the origin of the female self-appellation chenqie]. *Journal of Changsha University of Science and Technology (Social Science)*, 33.5:118–122.

Zwart, Jan-Wouter. 2007. Layered derivations. ACTL lecture. London: University College London.

THE GRAMMATICAL NATURE OF SINO-VIETNAMESE “VERB-OBJECT COMPOUNDS”¹

Yasunori TAKAHASHI
Kobe University
ytakahashi@port.kobe-u.ac.jp

Abstract

In this study, I address the grammatical nature of Sino-Vietnamese words borrowed from Chinese verb-object compounds (or *lǐhécí* in Chinese). In Chinese, this type of compound has several idiosyncratic characteristics, and previous studies of Chinese grammar suggest that they have dual status as words and phrases. In Vietnamese, similar to other lexical items, a number of verb-object compounds have been borrowed from Chinese. I conducted two experiments to investigate whether the Sino-Vietnamese “verb-object compounds” retain separability and object restriction. The results show that unlike the original forms, they cannot be separated by other morphemes, and they sometimes became transitive verbs during or after the borrowing, which are different from the original intransitive forms in Chinese. These findings clearly indicate that the Sino-Vietnamese “verb-object compounds” have almost completely lost their phrasal status attested to in their original forms and only retain the word status.

Keywords: Vietnamese, Chinese, verb-object compounds, grammatical borrowing
ISO 639-3 codes: vie, cmn, yue, jpn

1 Introduction

It is well known that Vietnamese, belonging to the Vietic branch of the Austroasiatic family, has largely been influenced by Chinese through long-term contact between the two languages. Previous studies such as Alves (2017) and Shimizu (2017) demonstrate that the impact of Chinese is remarkable at the phonological and lexical levels of Vietnamese, and many studies on Sino-Vietnamese words (i.e., loanwords from Chinese) have paid attention to their phonological and/or lexicosemantic characteristics for decades. In contrast, a smaller number of studies have focused on the grammatical (i.e., morpho-syntactic) nature of Sino-Vietnamese words because it is believed that the grammatical influence of Chinese is more subtle than its phonological and lexicosemantic effects. In exploring lexical borrowing from Chinese to Vietnamese in the early and pre-modern eras (i.e., spoken borrowing during the first millennium CE and borrowing through Chinese writing after the Tang dynasty), Alves (2007a:343) argued that many of the grammatical characteristics typical of varieties of Chinese are not part of Vietnamese grammar, and that the grammatical influence of Chinese is primarily lexical rather than structural. In contrast, Alves (2007b) suggested that a number of Sino-Vietnamese words acquired new grammatical meaning or functions after borrowing, which means that internal grammaticalization of them occurred in Vietnamese. Along these lines of research, Washizawa (2019) investigated the internal grammaticalization processes of some Sino-Vietnamese words from the 16th to the 19th centuries.

As for the grammatical borrowing from Chinese into Vietnamese, I have noticed that few studies have examined the synchronic nature of Sino-Vietnamese words and compared it with that of the original forms in Chinese. Therefore, for this study, I focused on the grammatical characteristics of Sino-Vietnamese “verb-object compounds” (VO compounds) which belong to ‘Sino-neologisms’ largely borrowed at the beginning of the 20th century for translating modern Western concepts (Vinh 1993, Alves 2007a, 2017). According to reference grammars of Chinese (Chao 1968, and Li and

¹ I would like to express my great appreciation to Mark Alves and Sho Yamaoka for their helpful comments and suggestions on earlier drafts of this article. Any remaining errors are my own. This research was supported by JSPS KAKENHI Grant Number 21K00501.

Thompson 1981, etc.), the original VO compounds have several idiosyncratic features; thus, they are suitable for considering the issues related to the nature of grammatical borrowing from Chinese into Vietnamese.

2 VO compounds in Chinese

In Chinese, there is a special lexical compound class called “VO compounds” (or *lǐhécí* 离合词). Li and Thompson (1981:73) suggested that they consist of two constituents having the syntactic relation of a verb and its direct object, but which are different from verb-object phrases in the following aspects. First, VO compounds consist of one or two bound morphemes. For example, the second morphemes of *shàng-shì* 上市 (ascend-city) ‘come to market’ and *lǐ-fà* 理发 (arrange-hair) ‘have a haircut’ are bound forms because they are not normally used alone in modern Chinese and must be combined with another morph, such as *chéng-shì* 城市 (city-city) ‘city’ and *tóu-fa* 头发 (head-hair) ‘hair’. These facts signal that these VO expressions are not verb phrases, but compounds.

Second, these Chinese-style VO compounds have the idiomaticity of the meaning of the entire unit (Li and Thompson 1981:73). In other words, their entire meaning is not derived from that of the constituents. For example, *shàng-shì* 上市 can be used to suggest that some fruits or vegetables are sold in the market because they are in season, which is not predictable from *shàng* ‘ascend’ and *shì* ‘city’. Similarly, *lǐ-fà* 理发 refers to cutting one’s hair, which is slightly different from the combined meaning of the morphemes *lǐ* ‘arrange’ and *fà* ‘hair’.

Third, and most significant to this study, even though they have a compound-like nature, as stated above, the vast majority of VO compounds allow their constituents to be separated by other morphemes (Li and Thompson 1981:73). In *shàng-shì* 上市 and *lǐ-fà* 理发 examples, they can be separated by various morphemes, such as the perfective marker *le* 了 and the experiential marker *guo* 过, as depicted in (1a) and (1b), respectively.

(1) Separation of VO compounds in Chinese

a. Separation by the perfect aspect marker *le* 了 (Packard 2016:76)

上	了	市
<i>shàng</i>	<i>le</i>	<i>shì</i>
ascend	PERF	city
‘came to market’		

b. Separation by the experiential aspect marker *guo* 过 (Li and Thompson 1981:75)

她/他	还	没	理	过	发
<i>Tā</i>	<i>hái</i>	<i>méi</i>	<i>lǐ</i>	<i>guo</i>	<i>fà</i>
3SG	still	not	arrange	EXP	hair
‘S/He still hasn’t ever had a haircut’					

Previous studies of Chinese grammar have paid much attention to this nature and discussed the grammatical status of VO compounds for decades (Lu 1957, Chao 1968, Lü 1979, Packard 2016, etc.). Packard (2016:76) summarized the current consensus very nicely: “*Li2he2ci2* are best viewed as a kind of word (词 *ci2*—i.e., a morphologically complex verb) with one or two bound constituents, but a word whose constituents are subject to syntactic reanalysis as free elements in certain limited contexts”. This clearly shows that VO compounds have dual status as words and phrases according to the situations in which they occur.

Li and Thompson (1981:76) pointed out another important feature of VO compounds in Chinese: the vast majority of VO compounds do not take a direct object after these VO lexical compounds. For example, in (2), *fēn-lèi* ‘classify’ (2a) cannot be followed by a direct object like *dòngwù* ‘animals’ (2b), the latter of which has to appear as an object of the preceding prepositional phrase headed by *gěi* 给 ‘to’, as portrayed in (2c).

(2) Object restriction in the VO compound *fēn-lèi* 分类

- a. 分类
fēn-lèi
divide-class
'classify'
- b. *分类 动物
**fēn-lèi* *dòngwù*
divide-class animal
'classify animals'
- c. 给 动物 分类
gěi *dòngwù* *fēn-lèi*
to animal classify
'classify animals'

Some previous studies have found that not all Chinese VO compounds are subject to this constraint. For example, Maruo and Han (2018) demonstrated that some words, such as *liú-xué* 留学 'study abroad' and *chū-xí* 出席 'participate in,' are allowed to take an object after them.

VO compounds are found not only in Mandarin, but also in other Chinese dialects as well. For example, Cheung (1972/2007) and Matthews and Yip (1994/2011) described VO compounds in Cantonese, which have grammatical features similar to those of Mandarin. In (3a), a VO compound *duhk-syū* 读书 'study' can be separated by the progressive aspectual marker *gán* 紧. In (3b), an object-like word *ngóh* 我 'first person singular pronoun' cannot occur after the VO compounds *bōng-mòhng* 帮忙 'help'; instead, the pronoun must be used as the possessor of the second morpheme *mòhng* 'busy'.

(3) VO compounds in Cantonese (Cheung 2007:91; Matthews and Yip 2011:58-59)

- a. 读书 读 紧 书
duhk-syū *duhk* *gán* *syū*
study-book study PROG book
'study' '(be) studying'
- b. 帮忙 帮 我 嘅 忙
bōng-mòhng *bōng* *ngóh* *ge* *mòhng*
help-busy help 1SG LP busy
'help' 'help me'

3 Sino-Vietnamese “VO compounds”

In addition to many other lexical items, a large number of VO compounds have been borrowed from Chinese to Vietnamese. Although the grammatical nature of Sino-Vietnamese “VO compounds” has not been well studied, several studies in Chinese as Second Language (CSL) partially refer to this issue. For example, Ngô (2007) showed that Vietnamese-translated words from Chinese VO compounds cannot be separated by other morphemes in general, and that their transitivity is not identical to that of VO compounds in Chinese. Nguyễn (2019) also investigated the transitivity of Vietnamese verbs having the same meaning as Chinese VO compounds, and suggested that some of them are transitive verbs, which leads to grammatical errors when Vietnamese learners use VO compounds in Chinese. These CSL studies provide important insights into the grammatical nature of Sino-Vietnamese “VO compounds”. However, we cannot identify the systematic characteristics of the compounds since both Ngô (2007) and Nguyễn (2019) analyzed mixed data, including both Sino-Vietnamese and Vietic words. It is necessary to focus only on Sino-Vietnamese words to grasp the nature of grammatical borrowing from Chinese to Vietnamese. Therefore, in this study, I only deal with Sino-Vietnamese “VO compounds”, which were largely borrowed in the early 20th century, and compare their grammatical aspects to those of the original forms in Chinese.

4 Methods and results

I conducted two types of experiments: separation and object tests. The former explores whether the constituents of Sino-Vietnamese “VO compounds” can be separated by other function words and the latter investigates whether they can take an object word behind them.

4.1. The separation test: Can Sino-Vietnamese “VO compounds” be separated by other elements?

Before conducting the experiments, I selected the target words in the following manner. First, I collected 268 VO compounds in Chinese from a *línhệ* dictionary by Zhou (2011) and determined that 91 compounds were listed as Sino-Vietnamese words in Kawamoto (2011), which is the largest Vietnamese-Japanese dictionary with Chinese characters. I then showed these 91 words to a female native speaker of Vietnamese (hereafter, S1) and asked her to comment on her familiarity with each word. She regarded 47 words as “familiar and commonly used” for her; thus, I selected them as the targets of the experiments, which are shown in Appendix 1.

The female speaker S1 participated in the separation and object tests. She was born in 1985 and had lived in Hanoi since childhood except for six years in Japan. I conducted the experiments in Hanoi in August of 2019.

The separation test allowed us to determine whether the constituents of the target words could be separated by three function words: *được* ‘fortunately, successfully, with a good result’, *lại* ‘repeating, doing over’, and *hết* ‘be completely finished, all used up, all gone, etc.’, which can be inserted between a verb and an object in verb phrases (Thompson 1965), as displayed in (4),

(4) Insertion of function words *được*, *lại*, and *hết* between a verb and an object (Thompson 1965:268, 345, 347)

a. *được* : fortunately, successfully, with a good result

Ông ấy mua **được** một ngôi nhà to.
he buy GR one CL house big
‘He purchased a large villa.’

b. *lại* : repeating, doing over

Lúc bạn tôi ra Hà Nội, tôi ở **lại** Sài Gòn làm việc.
when friend I go Hanoi I in (verb) again Saigon do thing
‘At the time my friend went to Hanoi, I remained in Saigon working.’

c. *hết* : be completely finished, all used up, all gone; no longer; completely, to the very end

Anh ấy tiêu **hết** cả tiền rồi.
he consume finish all money PERF
‘He has spent all the money already’

I asked S1 whether it was possible to insert the three function words between the first and second constituents of each target word. The results indicated that no target words could be separated by them, with one exception: *phát được điện* ‘to be able to generate electricity’ from *phát điện* (generate-electricity) ‘to generate electricity’. S1 created an example sentence, as in (5).

(5) *Cái máy này phát được điện.*
CL machine this generate GR electricity
‘This machine can generate electricity.’

It is noted that, in Vietnamese, the constituents *phát* and *điện* are full words rather than bound morphemes, as suggested to me by Mark Alves (p.c.) and Sho Yamaoka (p.c.). However, S1 said other two function words *lại* and *hết* were not allowed to be inserted in the same position.

4.2. The object test: Can Sino-Vietnamese “VO compounds” take an object?

After the insertion test, I conducted an object test with S1 simultaneously. In this experiment, I asked her whether each target word could take an object behind it. To test this sense of grammaticality, I often formed example phrases or sentences in which target words were followed by related object words, and S1 judged their grammaticality. If the expressions were grammatical to her, I asked her to create another example phrase or sentence with an object. After the experiment, I found that S1’s answers were sometimes inconsistent and seemed to differ slightly from the predictions inferred from reference grammars and dictionaries of Vietnamese. For this reason, I thought it necessary to collect additional data from another native Vietnamese speaker to reach a valid conclusion. However, I could not conduct the full experiments due to the Covid-19 pandemic as of early 2020; thus, I only performed a partial test with the second native speaker of Vietnamese (hereafter, S2) in January of 2021 via Zoom. The second speaker, S2, was born in 1993 and lived in Vung Tau in Southern Vietnam for 18 years, and then studied at a university in Ho Chi Minh City for four years. After graduating from the university, she stayed in Japan as a graduate student; at the time of the experiment, she was still studying at a graduate school in Japan.

Before conducting the second experiment with S2, I presented her with 47 original target words (see Appendix 1) and asked her to comment on her familiarity with each word. She regarded 32 words as “familiar and commonly used” for her; therefore, I selected them as the new target words for the object test. In the experiment, S2 judged whether each target word could take an object in the same manner as in the experiment with S1.

The combined results collected from S1 and S2 can be classified into five groups. First, S1 and S2 agreed that seven words could take an object, as seen in Table 1.

Table 1: Group 1 (target words for which the speakers accepted object noun phrases)

Word	Meaning	Example (Words in bold are object words)
<i>bảo mật</i>	keep-secret ‘keep a secret’	<i>Bảo mật thông tin</i> ‘keep information ’
<i>chiêu sinh</i>	invite-student ‘recruit students’	<i>Chiêu sinh các lớp học</i> ‘recruit students for each class ’
<i>gia công</i>	add-work ‘process’ (verb)	<i>Gia công máy móc</i> ‘process a machine ’
<i>phát bệnh</i>	generate-disease ‘become ill’	<i>Phát bệnh tim</i> ‘get heart disease’
<i>phân công</i>	divide-work ‘divide the work’	<i>Phân công công việc này</i> ‘divide this task ’
<i>phân loại</i>	divide-kind ‘classify’	<i>Phân loại động vật</i> ‘classify animals ’
<i>tốt nghiệp</i> ²	finish-job ‘graduate’	<i>Tốt nghiệp trường đại học</i> ‘graduate from university ’

According to Zhou (2011:107), in Chinese, only *jia-gong* 加工 (add-work) ‘to process’ corresponding with the Sino-Vietnamese *gia công* can take an object, while the remaining six VO compounds cannot. These results imply that some Sino-Vietnamese “VO compounds” changed their transitivity during or after the borrowing.

The second group included seven words where both native speakers made different judgments: S1 argued that all of them could take an object, while S2 believed that none of them could. I searched for additional information about the words from Vietnamese dictionaries and the SEALang Library Vietnamese Text Corpus, and found that at least four of them (i.e., *điểm danh* ‘call the roll’, *nhập học*

² *Tốt nghiệp* is a common Sino-Vietnamese word borrowed from *zú-yè* 卒業 (finish-job) ‘to graduate’. This word is not included in Zhou (2011); instead, *bì-yè* 毕业 (finish-job) ‘to graduate’ is listed there. However, in this study, we adopted *tốt nghiệp* as a target word because *zú* 卒 and *bì* 毕 have the same meaning ‘to finish’.

‘enter a school’, *thi công* ‘build’, and *từ chức* ‘resign’) have examples with an object, as presented in Table 2.

Table 2: Group 2 (target words for which the speakers had different judgements; SEA: SEALang Library Vietnamese Text Corpus)

Word	Meaning	Example from dictionaries and a corpus (Words in bold are object words)
<i>báo danh</i>	report-name ‘enroll’	
<i>bế mạc</i>	close-curtain ‘close (a ceremony, etc.)’	
<i>điểm danh</i>	call-name ‘call the roll’	Điểm danh học sinh (Chu et al. 2015) ‘call the students ’
<i>khai mạc</i>	open-curtain ‘open (a ceremony, etc.)’	
<i>nhập học</i>	enter-study ‘enter a school’	khi tôi nhập học đại học ... (SEA) ‘When I entered university ...’
<i>thi công</i>	carry out-work ‘build’	Thi công khù nhà ở cao tầng (Nguyễn and Phú 2013) ‘build a high-rise housing area ’
<i>từ chức</i>	resign-job ‘resign’	Từ chức hiệu trưởng (Nguyễn and Phú 2013) ‘resign as principal ’

The third group includes two words that can be followed by a verb phrase, rather than an object noun phrase. The two native speakers agreed that *phạm tội* can be followed by a verb phrase involving a criminal behavior such as *buôn bán ma túy* ‘to sell drugs’, and *tuyên thệ* can take a verb phrase representing a detailed content of the declaration. In the examples in Table 3, the verb phrases seem to be complementary clauses, rather than a part of serial verb constructions.

Table 3: Group 3 (target words for which the speakers accepted verb phrase adjuncts)

Word	Meaning	Example (Words in bold represent verb phrases)
<i>phạm tội</i>	violate-crime ‘commit a crime’	<i>Phạm tội</i> buôn bán ma túy . ‘commit drug sales ’
<i>tuyên thệ</i>	declare-notify ‘declare’	<i>Tuyên thệ</i> sẽ làm gì đó . ‘declare that (I/you/he/she) will do something ’

The fourth group included five words that could be followed by a prepositional phrase. In Table 4, prepositions *về* ‘about, on’, *đến* ‘about, on, over’, and *với* ‘to, together with, against’ (Nguyễn 1997: 162) are used between the target verbs and their related nouns. This group has shared grammatical characteristics with the corresponding VO compounds in Chinese: As noted in (2), VO compounds in Chinese generally use a prepositional phrase in place of an object noun phrase.

Table 4: Group 4 (target words for which the speakers accepted prepositional phrase adjuncts)

Word	Meaning	Example (Words in bold represent prepositional phrases)
<i>an tâm</i>	be satisfied-heart 'be relieved'	<i>an tâm về tương lai</i> 'be relieved about the future '
<i>hành quân</i>	go-army '(troops) march'	<i>hành quân đến [tên địa điểm]</i> 'march to [a place name] '
<i>kết hôn</i>	connect-marriage 'marry'	<i>kết hôn với anh ấy</i> 'marry him '
<i>ly hôn</i>	leave-marriage 'divorce'	<i>ly hôn với Nam</i> 'divorce Nam '
<i>nhập cảnh</i>	enter-border 'enter into a country'	<i>nhập cảnh vào Việt Nam</i> 'enter into Vietnam '

Finally, the research showed that 11 words that cannot be followed by any related elements, such as object nouns, verb phrases, or prepositional phrases. The intransitive nature of these 11 words is consistent with the corresponding VO compounds in Chinese.

Table 5: Group 5 (target words for which the speakers accepted no adjuncts)

Word	Meaning	Word	Meaning
<i>bãi công</i>	cease-work 'go on strike'	<i>sinh bệnh</i>	get-disease 'get ill'
<i>biến chất</i>	change-quality 'go bad'	<i>thất học</i>	lose-study 'be deprived of education'
<i>nhập viện</i> ³	enter-hospital 'be in the hospital'	<i>thất nghiệp</i>	lose-job 'be unemployed'
<i>phá sản</i>	break-property 'go bankrupt'	<i>trực ban</i>	be on duty-shift 'be on duty'
<i>phạm pháp</i>	violate-law 'break the law'	<i>xuất cảnh</i>	go out-border 'leave the country'
<i>phát điện</i>	generate-electricity 'generate electricity'		

Table 6 portrays the overall results of the object test: the members of Group 2 are excluded because we need to further investigate their transitivity. In Table 6, although a majority of the target words remain intransitive, at least eight Sino-Vietnamese words (i.e., six words in Group 1, except for *gia công* 'to process', and two words in Group 2) are different from their original Chinese forms in transitivity.

Table 6: Summary of the object test

May take object noun phrases	May take verb phrase adjuncts	May take prepositional phrase adjuncts	May not take any adjuncts
7	2	5	11
Different from Chinese (except for <i>gia công</i> 'to process')		In common with Chinese	

5 Discussion and conclusion

In this study, I conducted two experiments to investigate the separability and transitivity of Sino-Vietnamese "VO compounds". In this section, I discuss their grammatical features and compare them with those of the original VO compounds in Chinese,

³ In the experiments, I adopted *nhập viện* borrowed from *rù-yuàn* 入院 (enter-hospital) as a target word, instead of *zhù-yuàn* 住院 (live-hospital) listed in Zhou (2011), because *nhập viện* is a common word in Vietnamese.

In the separation test, almost none of the target words could be separated by functional morphemes, which clearly shows that they have almost all lost their phrasal status attested to in the VO compounds in Chinese. As observed in Section 2, the dual status of Chinese VO compounds is one of the idiosyncratic characteristics in the grammatical system of the language. However, their borrowed forms in Vietnamese only have status as compound-like words.

Meanwhile, in the object test, a number of target words have different transitivity from their original forms in Chinese. As demonstrated by Li and Thompson (1981), in Chinese, the majority of VO compounds behave like intransitive verbs and cannot take an object. In contrast, a number of target Sino-Vietnamese words behave like transitive verbs, which means they changed their transitivity in the process or following being borrowed. Although it is not clear what factors determine the transitivity of each Sino-Vietnamese “VO compound”, it may be the case that some cross-linguistic semantic factors play an important role because, as shown in (6), the transitivity of Sino-Vietnamese “VO compounds” is basically the same as that of Sino-Japanese ones. In (6a), the Sino-Japanese *bunrui*, borrowed from *fēn-lèi* 分类 ‘classify’, can take an object noun phrase (i.e., noun + the accusative particle *o*) and, in (6b), *hasan*, borrowed from *pò-chǎn* 破产 ‘bankrupt’, is only available as intransitive. These Sino-Japanese words are entirely consistent with the corresponding Sino-Vietnamese words *phân loạ*i ‘classify’ and *phá sản* ‘bankrupt’ in transitivity. We will evaluate the importance of this semantic factor in the future.

(6) The transitivity of Sino-Japanese “VO compounds” *bunrui* and *hasan*

- | | | | | | |
|----|-------------------------------|---------------|---------------|--------------|---------------|
| a. | <i>dōbutsu</i> | -o | <i>bunrui</i> | -suru | |
| | animal | ACC | classify | do | |
| | ‘classify animals’ | | | | |
| | | | | | |
| b. | <i>kono-</i> | <i>kaisha</i> | -wa | <i>hasan</i> | -shita |
| | this | company | TOP | bankrupt | did |
| | ‘This company went bankrupt.’ | | | | |

Finally, let us consider the historical process of borrowing VO compounds. It is worth noting that the majority of target words in my experiments belong to ‘Sino-neologisms’ borrowed into Vietnamese at the beginning of the 20th century. Referring to Vinh (1993), Alves (2007a, 2017) pointed out that the borrowed items in this period contained a mixture from both Japan and China because the Japanese had been translating Western concepts by utilizing Chinese lexical material in the late 1800s and early 1900s. In this study, it is necessary to confirm whether the target words were borrowed from China or Japan since several target words have the same characteristics as the corresponding Sino-Japanese “VO compounds” in separability and transitivity. According to the word list 越南语非纯汉双音节汉越词表 (Luo 2018), including disyllabic Sino-Vietnamese words originating from countries other than China, only *thi công* ‘build’ was created outside of China (i.e., from Japan) in my target words. Chen (2019) showed that in his database—which includes 1,028 common verbs in Chinese—only 47 (4.57%) were borrowed from Japanese, and that the borrowing rate of verbs was much lower than that of nouns (11.39%) and adjectives (7.7%). Based on the findings of the previous studies, it may be the case that verbs created in the Japanese language were not the main resources of Sino-neologisms, and that the Vietnamese borrowed very few verbs directly from Japanese at the beginning of the 20th century. We still have to investigate the borrowing route of Sino-neologisms in more detail; however, at this point, there is no clear evidence of influence from Japanese, at least in the target words of my experiments.

References

- Alves, Mark. 2007a. Sino-Vietnamese grammatical borrowing: An overview. In *Grammatical borrowing in cross-linguistic perspective*, ed. by Yaron Matras and Jeanette Sakel, 343–361. De Gruyter Mouton.
- Alves, Mark. 2007b. Categories of grammatical Sino-Vietnamese vocabulary. In *The Mon-Khmer Studies* 37:217–229.

- Alves, Mark. 2017. Chinese loanwords in Vietnamese. In *Encyclopedia of Chinese language and linguistics*, volume 1, ed. by Rint Sybesma, Wolfgang Behr, Yueguo Gu, Zev Handel, C.-T. James Huang, and James Myers, 585–592. Brill.
- Chao, Yuen-Ren. 1968. *A grammar of spoken Chinese*. University of California Press.
- Chen, Li Wei. 2019. *Kindaichi no honyaku to denpa: Kango wo baikai ni* (近代知の翻訳と伝播 漢語を媒介に) [Translation and propagation of modern knowledge: Through the medium of Sino-Japanese words]. Sanseido (三省堂).
- Cheung, Hung-nin (张洪年). 1972/2007. *A grammar of Cantonese as spoken in Hong Kong* (香港粤语语法的研究). The Chinese University Press, The Chinese University of Hong Kong.
- Chu Bích Thu, Nguyễn Ngọc Trâm, Nguyễn Thị Thanh Nga, Nguyễn Thuỳ Khanh, and Phạm Hùng Việt. 2015. *Từ điển tiếng Việt phổ thông* (tái bản có sửa chữa & bổ sung) [The common dictionary of Vietnamese (modified and supplemented reprint)]. Nhà Xuất Bản Phương Đông.
- Kawamoto, Kunie. 2011. *Shōkai Betonamugo jiten* (詳解ベトナム語辞典) [The detailed dictionary of Vietnamese-Japanese]. Taishukan (大修館).
- Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. University of California Press.
- Lu, Zhiwei (陆志韦). 1957. *Hanyu de goucifa* (汉语的构词法) [Chinese morphology]. Kexue Chubanshe (科学出版社).
- Luo, Wenqing (罗文青). 2018. *Dangdai Yuenanyu hanzi cihui shiyong xianzhuang yanjiu* (当代越南语汉字词汇使用现状研究) [A study of usage of Chinese words in modern Vietnamese]. Shijie Tushu Chubanshe Guangdong (世界图书出版广东).
- Lü, Shuxiang (吕叔湘). 1979. *Hanyu yufa fenxi wenti* (汉语语法分析问题) [Problems in analyzing Chinese grammar]. The Commercial Press.
- Maruo, Makoto and Tao Han. 2018. Separable word usage in Chinese with particular reference to the verb *liuxue*. *Journal of Humanities Nagoya University* 1:129–145.
- Matthews, Stephen and Virginia Yip. 1994/2011. *Cantonese: A comprehensive grammar*. Routledge.
- Ngô Thị Lư Hải (吴氏流海). 2007. *Yuenan xuesheng Hanyu dongbinshi liheci xide yanjiu yu jiaoxue duice* (越南学生汉语动宾式离合词习得研究与教学对策) [An acquisition study of Chinese verb-object compounds by Vietnamese students and its learning strategies]. Master's thesis, Beijing Language and Culture University.
- Nguyễn Đình-Hoà. 1997. *Vietnamese*. John Benjamins.
- Nguyễn Lý Uy Hân (阮李威欣). 2019. *Dui Yue Hanyu jiaoxue zhong de liheci yanjiu* (对越汉语教学中的离合词研究) [Research on separable words in Vietnamese-Chinese teaching]. Ph.D. dissertation, Central China Normal University.
- Nguyễn Tôn Nhan and Phú Văn Hân. 2013. *Từ điển tiếng Việt*. [The Vietnamese dictionary]. Nhà Xuất Bản Từ Điển Bách Khoa.
- Packard, Jerome. 2016. Lexical word formation. In *A reference grammar of Chinese*. ed. by Chu-Ren Huang and Dingxu Shi, 67–80. Cambridge University Press.
- Shimizu, Masaaki. 2017. Chinese and Vietnamese. In *Encyclopedia of Chinese language and linguistics*, volume 1, ed. by Rint Sybesma, Wolfgang Behr, Yueguo Gu, Zev Handel, C.-T. James Huang, and James Myers, 389–392. Brill.
- Thompson, Laurence C. 1965. *A Vietnamese grammar*. University of Washington Press. reprinted 1987 as *A Vietnamese reference grammar*. University of Hawaii Press.
- Vinh, Sinh. 1993. Chinese characters as the medium for transmitting the vocabulary of modernization from Japan to Vietnam in early twentieth century. *Asian Pacific Quarterly* 25.1:1–16.
- Washizawa, Takuya. 2019. Process of grammaticalization of function words *cửa, sự, không, and bị* in Vietnamese. *Tokyo University Linguistic Papers (TULIP)* 41:367–392.

Zhou, Shangzhi (周上之). 2011. *Hanyu changyong liheci yongfa cidian* (汉语常用离合词用法词典) [A dictionary of the usage of common Chinese separable words]. Beijing Language and Culture University Press.

Corpus

SEALang Library Vietnamese Text Corpus, online corpus: <http://sealang.net/vietnamese/corpus.htm> (Last accessed 5 July 2021).

Appendix 1: Target words of the separation test

Word in Quốc ngữ	Meaning of		Meaning of the entire word	Chinese characters
	1st morpheme	2nd morpheme		
<i>an tâm</i>	be satisfied	heart	‘be relieved’	安心
<i>bãi công</i>	cease	work	‘go on strike’	罢工
<i>báo danh</i>	report	name	‘enroll’	报名
<i>bảo mật</i>	keep	secret	‘keep a secret’	保密
<i>bế mạc</i>	close	curtain	‘close (a ceremony, etc.)’	闭幕
<i>biến chất</i>	change	quality	‘go bad’	变质
<i>biến hình</i>	change	shape	‘be out of shape’	变形
<i>cáo trạng</i>	tell	complaint	‘charge, indictment’ (noun)	告状
<i>chiêu sinh</i>	invite	student	‘recruit students’	招生
<i>dụng công</i>	use	work	‘try hard’	用功
<i>điểm danh</i>	call	name	‘call the roll’	点名
<i>đính hôn</i>	agree on	marriage	‘be engaged to’	订婚
<i>gia công</i>	add	work	‘process’ (verb)	加工
<i>hành quân</i>	go	army	‘(troops) march’	行军
<i>kết hôn</i>	connect	marriage	‘marry’	结婚
<i>khai khẩu</i>	open	mouth	‘open one’s mouth and say something’	开口
<i>khai mạc</i>	open	curtain	‘open (a ceremony, etc.)’	开幕
<i>lộ diện</i>	show	face	‘show up’	露面
<i>luyện binh</i>	practice	solder	‘train troops’	练兵
<i>lưu tâm</i>	keep	heart	‘take care’	留心
<i>lưu ý</i>	keep	intention	‘pay attention to’	留意
<i>ly hôn</i>	leave	marriage	‘divorce’	离婚
<i>mệnh danh</i>	order	name	‘name (a baby etc.)’	命名
<i>nhập cảnh</i>	enter	border	‘enter into a country’	入境
<i>nhập học</i>	enter	study	‘enter a school’	入学
<i>nhập viện</i>	enter	hospital	‘be in the hospital’	入院
<i>nhượng bộ</i>	give way	step	‘make concessions’	让步
<i>phá sản</i>	break	property	‘go bankrupt’	破产
<i>phạm pháp</i>	violate	law	‘break the law’	犯法
<i>phạm tội</i>	violate	crime	‘commit a crime’	犯罪
<i>phát bệnh</i>	generate	disease	‘become ill’	发病
<i>phát điện</i>	generate	electricity	‘generate electricity’	发电
<i>phát hỏa</i>	generate	fire	‘catch fire; be angry’	发火
<i>phát tài</i>	generate	wealth	‘get rich’	发财
<i>phân công</i>	divide	work	‘divide the work’	分工
<i>phân loại</i>	divide	kind	‘classify’	分类
<i>sinh bệnh</i>	get	disease	‘get ill’	生病

<i>tạo phản</i>	make	rebellion	'rebel'	造反
<i>thất học</i>	lose	study	'be deprived of education'	失学
<i>thất nghiệp</i>	lose	job	'be unemployed'	失业
<i>thi công</i>	carry out	work	'build'	施工
<i>thương tâm</i>	hurt	heart	'be sorrowful'	伤心
<i>tốt nghiệp</i>	finish	job	'graduate'	毕业
<i>trực ban</i>	be on duty	shift	'be on duty'	值班
<i>tuyên thệ</i>	declare	notify	'declare'	宣誓
<i>từ chức</i>	resign	job	'resign'	辞职
<i>xuất cảnh</i>	go out	border	'leave the country'	出境

Abbreviations

ACC = accusative

CL = classifier

EXP = experiential

GR = good result

LP = linking particle

PERF = perfective

PROG = progressive

SG = singular

TOP = topic

1, 3 = first, third person

BAGOBO-KLATA PHONOLOGY

Edward G. ESTRERA
University of the Philippines-Diliman
egestrera@up.edu.ph

Abstract

This paper provides a phonological description of Bagobo-Klata, a Southern Philippine language spoken in the eastern slopes of Mt. Apo, Davao City, the Philippines. Based on a 1,026-word list, the description is mostly concerned with the synchronic phonology of Bagobo-Klata but notes selected diachronic sound changes. The synchronic part discusses its phonemic inventory, syllable structure, segmental stress, phonotactics, and phonological processes, while the diachronic part lists and explains salient Bagobo-Klata reflexes of Proto-Malayo-Polynesian (hereafter, PMP) phonemes. Among the salient phonological features of Bagobo-Klata are its five-vowel system /a, ε, ɔ, i, u/, consonant clusters, geminates, several phonological processes triggered by affixation, and the historically deleted word-final *n (e.g., PMP *bulan > *bula* ‘moon, month’, PMP *dahun > *daʔu* leaf, PMP *ipen > *ʔippɔ* tooth, PMP *quzan > *ʔula* ‘rain’, and PMP *zalan > *dala* ‘abdomen, path’).

Keywords: Bagobo-Klata, Philippine languages, phonology
ISO 639-3 codes: bgi

1 Introduction

This paper provides an account of the phonology of Bagobo-Klata, a Southern Philippine language largely spoken in the eastern slopes of Mt. Apo, Davao City, Davao del Sur, the Philippines, or in approximately thirty-six barangays—including Barangay Don Panaca in Magpet, Cotabato—scattered in Baguio, Calinan, and Tugbok districts. The speakers of this language call themselves Bagobo-Klata, the endonym, or just Klata, but they seem to be more commonly known as Giangan. This is quite evident in the literature of Philippine linguistics but has been rectified in the 24th edition of the *Ethnologue*, in which Bagobo-Klata replaced Giangan as the language name in its language profile (Blust 1991 & 2019; Eberhard, Simons & Fennig 2021; McFarland 1994; Walton 1979; Zorc 1986). Bagobo-Klata speakers can be generally referred to as Bagobo, the collective term for the three ethnolinguistic groups (i.e., Obo Monuvu and (Bagobo-)Tagabawa) living in Mount Apo.

1.1 Previous Research

There are two phonological accounts on Bagobo-Klata: Cagas (1991) and Evans (2017), both of which describe the basic features of its phonological system, namely, the sound inventory, syllable structure, gemination, consonant clusters, and segmental stress. Between the two accounts, Evans (2017) is more thorough, specifically in terms of the treatment of allophony, gemination, and consonant clusters but has plenty of generalizations that are unsupported by examples and must be re-examined because of their implications for the overall phonological analysis of Bagobo-Klata.

First, the voiced alveodental tap /ɾ/ is treated as phonemic, despite researchers’ not providing a minimal pair. The only supporting claim is that /ɾ/ is in unconditioned free variation with /d/ in some roots such as /hɔɾɔʔ/ ‘to stop’, but [hɔdɔʔ] is not recorded. On the contrary, Cagas (1991) states that /ɾ/ is a variant of /d/ in intervocalic positions. Second, how the allophones of Bagobo-Klata consonants and vowels are analyzed is problematic: the environments that condition these allophones are not specified. For instance, gemination is claimed to be both phonemic and phonetic without evidence. In addition to that, vowel length is also shown to be phonetic without the conditioning environment. Third, the glottal stop occurring as an onset is claimed to be epenthetic, implying that V is the basic syllable structure in Bagobo-Klata. To support such a claim, acoustic data could have been provided.

While there exist general descriptions of Bagobo-Klata phonology, a new one that re-examines such generalizations in these studies and provides a fresh analysis of Bagobo-Klata phonology is warranted.

1.2 Data

The data used in this paper were gathered during the linguistic fieldwork embarked on by the author in Barangay Sirib, Calinan District, Davao City from the 10th to the 27th of November 2019 and during a series of virtual informant works from March to June 2021. Both data-gathering activities were done with permission from National Commission of Indigenous Peoples (NCIP) in Davao Region. Moreover, the data include a 1,026-word list recorded by the datu or chieftain of Barangay Sirib on the 27th of November 2019 and fully transcribed in January 2020; and a separate list of nominal and verbal derivations was elicited from two female language consultants via Facebook Messenger from March to June 2021.

2 Consonants

Bagobo-Klata has sixteen consonants, as shown in Table 1. These consonants contrast in places and manners of articulation and voicing. It must be noted that in this paper, [r] is treated as allophonic, even though there are instances in which [r] occurs in Bagobo-Klata roots such as /ki.'rɔp/ 'to blink one's eye' and /hɔ.'rɔʔ/ 'to stop'.

Table 1: Bagobo-Klata consonants

	Bilabial	Alveodental	Palatal	Velar	Glottal
Stop	p b	t d		k g	ʔ
Nasal	m	n		ŋ	
Fricative		s			h
Tap		[r]			
Lateral		l			
Approximant			j	w	

Shown in Table 2 are the minimal pairs of the fifteen contrastive consonants in Bagobo-Klata.

Table 2: Bagobo-Klata segments exemplified in context

	Word-Initial	Word-Medial	Word-Final
/p/	[pɛ.'taʔ] 'wet' [ʔɛ.'taʔ] 'thigh'	[ʔi:.pit] 'to sleep' [ʔi:.nit] 'hot'	[ʔi:.hip] 'to whisper' [ʔi:.hiʔ] 'to whet a blade'
/b/	[ʔba:.tuk] 'cough' [pa:.tuk] 'duck'	[ʔɔ.'bɔw] 'mouse, rat' [ʔɔ.'lɔw] 'fence'	[ʔɔl.'lɔb] 'spring' [ʔɔl.'lɔt] 'between'
/t/	[tɔ.'li] 'string' [ʔɔ.'li] 'to choose'	[ʔɛ:.taʔ] 'calf (one's leg)' [ʔɛ:.maʔ] 'armpit'	[kɔk.'kɔt] 'to bury' [kɔk.'kɔp] 'to embrace'
/d/	[da:.'ʔu] 'leaf' [ba:.'ʔu] 'smell, general'	[tu:.duk] 'oar' [tu:.kuk] 'lazy'	[ʔa:.tad] 'raft' [ʔa:.taʔ] 'rice husk'
/k/	[kɛ:.wɔʔ] 'stout' [lɛ:.wɔʔ] 'narrow'	[lu:.ka] 'bowl' [lu:.wa] 'ladle'	[ʔɔ.'lɔk] 'to kiss' [ʔɔ.'lɔw] 'fence'
/g/	[gɔl.'lɔt] 'middle finger' [ʔɔl.'lɔt] 'between'	[ʔɛ:.gɔŋ] 'feces' [ʔɛ:.kɔŋ] 'tail'	[ʔa.'ʔɔg] 'to take a bath' [ʔa.'ʔɔw] 'to enter'
/ʔ/	[ʔul.'lu] 'head' [hul.'lu] 'to command'	[tɔ.'ʔi] 'to accompany' [tɔ.'li] 'string'	[ʔɔ.'tɔʔ] 'nickname' [ʔɔ.'tɔp] 'roof'

Word-Initial	Word-Medial	Word-Final
/m/ [ma.liʔ] ‘good’ [pa.liʔ] ‘wound’	[hi.mat] ‘needle’ [hi.kat] ‘fast’	[lag.gam] ‘bird’ [lag.gas] ‘to wash hands’
/n/ [nɛ.pɛs] ‘thin (object)’ [lɛ.pɛs] ‘knife’	[ʔi.nɔ] ‘mother’ [ʔi.wɔ] ‘saliva’	
/ŋ/ [ŋil.lɔ] ‘ear’ [til.lɔ] ‘flee (of a dog)’	[ʔaŋ.ŋaʔ] ‘child’ [ʔaŋ.gaʔ] ‘nickname’	[lud.duŋ] ‘fish’ [lud.dus] ‘to be lost’
/s/ [sɔ.ʔa] ‘ankle’ [hɔ.ʔa] ‘thorn’		[kɔl.lɔs] ‘strong’ [kɔl.lɔt] ‘curly (of hair)’
/h/ [hɔ.bɔw] ‘milk’ [ʔɔ.bɔw] ‘mouse’	[pu.huŋ] ‘heart’ [pu.ʔuŋ] ‘to wrap’	
/l/ [lɔs.suŋ] ‘mortar’ [pɔs.suŋ] ‘water jar’	[bu.la] ‘moon’ [bu.na] ‘to hit’	
/j/ [ju.pa] ‘centipede’ [ku.pa] ‘thick’	[ha.jup] ‘false’ [ha.ŋup] ‘watermelon’	[ʔɔ.tɔj] ‘liver’ [ʔɔ.tɔp] ‘roof’
/w/ [wɔd.dɔw] ‘afternoon’ [kɔd.dɔw] ‘noon’	[bɔ.wɔʔ] ‘to pour’ [bɔ.jɔʔ] ‘face’	[ʔɔt.tɔw] ‘person’ [ʔɔt.tɔk] ‘brain’

Shown in Table 3 is the distribution of Bagobo-Klata consonants in all word positions. Notes are in order. First, all consonants (except the allophonic [r]) occur word-initially. Second, word-initial consonants can be singletons, geminates (segmental or derived), and a member of word-medial clusters. The examples shown below are singletons only. Third, only twelve consonants (except /h, n, l/ and [r]) can occur word-finally. As will be explained later, PMP *h, *n, and *l are phonemically lost in the word-final position, although in some words, word-final /n/ is not deleted, as in (1).

- (1) *pa.ka*.wan ‘cup’
sa.bɔn ‘soap’
ma.ja.man ‘rich’
tu.rɛʔ.kan ‘I do not know’

Table 3: Bagobo-Klata consonants in word-initial, -medial, and -final Positions

#_	V_V	_#
/p/ [pud.du] ‘gall, bile’ [pit.tu] ‘seven’ [pɔ.lɔs] ‘many’	[ʔɔ.puj] ‘fire’ [ta.piʔ] ‘wall’ [lɛ.pɔs] ‘sibling’	[ʔi.hip] ‘to whisper’ [ʔɔ.kap] ‘bark (of a tree)’ [ʔa.kap] ‘monkey’
/b/ [bu.lu] ‘to cut, as one’s hair’ [bul.las] ‘to change clothes’ [bɔl.l] ‘poison’	[ʔɛ.baŋ] ‘left’ [gɔ.bɔ] ‘to deceive’ [ʔu.biʔ] ‘to request’	[la.ŋub] ‘cave’ [ŋu.ʔɔb] ‘fingernail’ [his.sib] ‘fat, grease’
/k/ [kal.laŋ] ‘scab’ [kɔ.lat] ‘thin (of a person)’ [kɔm.mi] ‘beard’	[ʔɛ.kɔŋ] ‘tail’ [ku.kaj] ‘cockroach’ [lu.ka] ‘bowl’	[lɔp.puk] ‘consumed’ [tu.kuk] ‘lazy’ [pal.lɔk] ‘sand’
[gu.hiŋ] ‘water jar’	[la.gat] ‘sea’	[la.lig] ‘happy’

#	V_V	#	
/g/	[gi.'ra] 'last' [gɔ.'tɔŋ] 'eggplant'	['mɔ:.gat] 'mote' ['ʔi.gɔʔ] 'betel leaf'	[pɔ.kɔ.'lɔŋ] 'how' ['ka:.bɔŋ] fruit 'bat'
/t/	[tu.'ʔud] 'deer' [tɔb.'biŋ] 'buttock' ['ti:.ʔi] 'pinky finger'	['ʔɛ:.taʔ] 'thigh' ['ti:.tɔŋ] 'bridge' ['pu:.tɔw] 'iron'	['ku:.lit] 'skin' [lɔ.'mɛt] 'weak' ['kɔ:.hɛt] 'cave bat'
/d/	[da.'luŋ] 'below, underneath' [da.'gɔw] 'short (of time)' [dun.'nuk] 'flood'	['lɛ:.dɔk] 'waterfall' ['ta:.dɔʔ] 'to drip' ['tɔ:.dɔ] 'to follow (as a trail)'	[tɛ.'kɔd] 'heel' [tul.'lid] 'straight (as a stick)' ['bu:.lud] 'mountain'
/ʔ/	[ʔɛ.'nɔʔ] murky [ʔɔn.'nɔp] mist ['ʔi:.hiŋ] ring (for one's finger)	['ti:.ʔɔw] 'clear (as water)' 'ʔɛ:.ʔɛ 'yes' ['tu:.ʔuw] to string together'	[lɔm.'mɔʔ] 'morning' ['ʔɛ:.maʔ] 'armpit' [ba.'juʔ] 'cheek'
/m/	['ma:.mɛ] 'leg' [mɔ.'jɔ] 'young woman' [mɔ.'ʔɔ] 'betel, areca nut'	['ŋu:.ma] 'story' ['lu:.mut] 'moss' ['lɔ:.maʔ] 'knife (of a datu)'	['ʔɛ:.lam] 'wasp' ['ʔɔ:.jam] 'yawn' [ʔik.'kam] 'mat, for sleeping'
/n/	[na.'nam] 'taste' [nɛ.'pɛs] 'thin (of an object)' [nɔ.'ʔɔs] 'sound, noise'	['la:.nu] 'sad' ['nɔ:.nɔ] 'prawn' ['pɔ:.naʔ] 'bait'	∅
/ŋ/	[ŋɔ.'nuk] 'to hunt' [ŋil.'lɔ] 'ear' [ŋit.'tɔŋ] 'dark'	['lu:.ŋa] 'shade' ['mɛ:.ŋa] 'stove' ['lu:.ŋu] 'coffin'	['tu:.baŋ] 'in front of' ['ʔɛ:.gɔŋ] 'feces' [mes.'sɛŋ] 'industrious'
/s/	[sɔs.'sɔp] 'to suck' [sa.lup.'paŋ] 'loincloth' [saj.'jɔw] 'to dance'	['ma:.si] 'salty' ['ba:.sa] 'to read' [ku.'li:.sap] 'dandruff'	['ma:.las] 'spicy' [nuw.'was] 'accustomed' ['ʔu:.pus] 'cat'
/h/	['hɔ:.lɔ] egg [his.'sip] louse (of a chicken) [hug.'gu] to push	['bɔ:.hɔw] 'arrow' ['tɔ:.hu] 'cucumber' ['ba:.hɔw] 'thirsty'	∅
/j/	[jab.baʔ.'na:.pu] 'nightmare' [jɔj.'jɔ] 'shame' [pa.'ja] 'big'	['ma:.jɔw] 'raincloud' ['lu:.jɔ] 'ginger' ['ʔi:.jup] to 'blow (using one's mouth)'	[ʔal.'lɔj] 'chin' [lus.'suj] 'gums' ['ŋu:.ʔuj] 'whistle'
/w/	[wi.'ti] 'hungry' [wɔd.'dɔw] 'afternoon' ['wɔ:.ʔɔ] 'abaca fibers, hemp'	['tu:.wɔ] 'old person' ['ʔa:.wa] 'rainbow' ['ba:.wɛʔ] 'medicine'	[ʔu.'pɔw] 'bald' [pɛ.'ʔɔw] 'lame, crippled' ['ga:.hɔw] ant
/l/	['li:.tu] to 'burn' [lab.'bus] 'poor' [lal.'lɔm] 'deep'	['wa:.lɔj] 'tired' ['da:.la] 'trail' ['bɔ:.laʔ] 'bubbles'	[pak.'sul] 'hole in the ground' [baŋ.'kil] 'canine tooth'

3 Gemimates

A geminate refers to a sequence of two identical adjacent consonants in a single morpheme (Crystal 2008:206); a single consonant is called a singleton. In terms of word position, Bagobo-Klata geminates are strictly word-medial. Moreover, they can be segmental (i.e., occurring in roots) or derived (i.e., triggered by affixes).

Table 4: Bagobo-Klata geminates

Geminate	Examples
pp	[dip.'pɔ] 'armspan' [ʔip.'pus] 'gasp, pant' [lɔp.'puk] 'all gone, consumed'
bb	[bab.'ba] 'broken in pieces' [tab.'baŋ] 'bland' [mɔb.'but] 'animal'
tt	[klat.'taŋ] 'steps in a ladder' [ʔɔt.'tɔw] 'person' [ŋit.'tɔŋ] 'dark'
dd	[tud.'duʔ] 'nape' [ʔid.'du] 'pity' [lɔd.'dɔg] 'rotten'
kk	[bik.'kɔ] 'necklace' [ʔik.'kam] 'mat, for sleeping' [lak.'kaʔ] 'jackfruit'
gg	[tug.'gɔ] 'post, house pole' [ʔɔg.'giʔ] 'cogon grass' [lag.'gu] 'while'
ss	[dis.'sɔʔ] 'sty, in one's eye' [mes.'sɛŋ] 'industrious' [lis.'sɔ] 'seed, of fruit'
mm	[dum.'mɔ] 'other' [ʔam.'muʔ] 'breast' [lam.'mi] 'new'
nn	[tɔn.'nɔb] 'honey' [ʔɔn.'nɔp] 'fog, mist' [lan.'na] 'pus'
ŋŋ	[pɔŋ.'ŋu:ʔɔ] 'chieftain' [daŋ.'ŋa] 'before' [hɔŋ.'ŋɔw] 'steam, vapor'
ll	[pul.'luʔ] 'ten' [bal.'las] 'rice (uncooked)' [lal.'lɔm] 'deep'
ww	[buw.'wa] 'hammock' [ʔuw.'wɔ] 'two' [luw.'wu] 'winnowing basket'
jj	[baj.'jɔ] 'crocodile' [gaj.'jɔ] 'south wind' [jɔj.'jɔ] 'shame'

Shown in Table 4 are segmental geminates. Only thirteen consonants (except the glottals /ʔ, h/ and the allophonic tap [ɾ]) can become geminates, but a few affixes can trigger the glottals to become geminates. As shown in (2) and (3), /h/, at least word-initially, becomes [ss] with distributive numeral affix *tɔ-* and the realis patient-voice (PV) *bɔ-*.

- (2) *tɔ-* + *hɔ. 'tu* → *tɔs. 'sɔ:tu*
 DN one one each
- (3) *bɔ-* + *hal. 'la* → *bɔs.sal. 'la*
 RLS.PV to fry rice RLS.PFV-fry rice

On the other hand, the glottal stop /ʔ/, specifically in the word-initial position, becomes a geminate and has several allophones [gg, tt, ll, jj, ww], as shown in (4), (5), (6), (7), and (8). What is certain, thus far, is that the derived geminate forms of /ʔ/ is due to fortition.

- (4) *bɔ-* + *'ʔe:.pɔk* → *bɔj. 'je:.pɔk*
 RLS.POT to cut RLS.POT-cut
- (5) *bɔ-* + *'ʔe:.pɔk* → *bɔg. 'ge:.pɔk*
 RLS.PV to cut RLS.PV-cut
- (6) *pɔ-* + *ʔaŋ. 'ŋaʔ* → *pɔl.laŋ. 'ŋaʔ*
 CAUS.RLS.PV child CAUS.RLS.PV-give birth
- (7) *tɔ-* + *ʔuw. 'wɔ* → *tɔl.lu. 'wɔ*
 DN two DN-two (two each)
- (8) *tɔ-* + *ʔɔt. 'tɔw* → *tɔt.tɔt. 'tɔw*
 NOM person NOM-person (self)

Like consonants and vowels, segmental geminates are also contrastive, as shown in Table 5.

Table 5. Contrastive geminates in Bagobo-Klata

Geminate	Singleton
[pɔt. 'tɔj] 'firefly'	[pɔ. 'tɔj] 'to die, to kill someone'
[bɔl. 'lɔj] 'to give'	[bɔ. 'lɔj] 'house'
[tap. 'pɛ] 'old (object)'	[ta. 'pɛ] 'twin'
[tɔŋ. 'ŋɔ] 'half'	['tɔ:ŋɔ] 'nipple, teat'
[kap. 'pɛŋ] 'to split into halves'	[ka. 'pɛŋ] 'wood stick'
[hud. 'duʔ] 'lad'	['hu:duʔ] 'to carry on one's head'
[lim. 'mɔ] 'hand'	[li. 'mɔ] 'five'
[ʔin. 'nɔ] 'to look'	[ʔi. 'nɔ] 'mother'
[ʔiw. 'wɔ] 'envy'	['ʔi:wɔ] 'saliva'

4 Vowels

Bagobo-Klata has five vowels that contrast in vowel quality, tongue position, and the degree of lip rounding. As shown in Table 12, only /u/ and /ɔ/ are rounded—full and half, respectively.

Table 6: Bagobo-Klata vowels

	Front	Central	Back
Close	i		u
Mid	ɛ		ɔ
Open		a	

Shown in Table 13 are the minimal pairs of Bagobo-Klata vowels in all syllable positions.

Table 7: Minimal pairs of Bagobo-Klata vowels

	Penultima	Ultima	Both Pen/Ultima
/i/	[kid.'dut] 'astonishment' [kɔd.'dut] 'to pinch'	[tɔb.'bi] 'to sew' [tɔb.'bu] 'sugarcane'	['ti:.ʔi] 'pinky finger' ['tɛ:.ʔɛ] 'hard'
/u/	[kul.'li] 'a long time ago' [kil.'li] 'eel'	['tu:.wu] 'to grow' ['tu:.wɔ] 'old'	[klut.'tuŋ] 'forehead' [klat.'taŋ] 'ladder steps'
/ɛ/	['ʔɛ:.taʔ] 'calf (of one's leg)' ['ʔɔ:.taʔ] 'to vomit'	['lɔ:.lɛ] 'valley' ['lɔ:.la] 'nest'	['hɛ:.lɛ] 'to lean' ['bɛ:.lɛ] 'to stay'
/ɔ/	[bɔl.'li] 'to buy' [bul.'li] 'evening/night'	[ʔul.'lɔd] 'larva' [ʔul.'lud] 'floor'	[bɔb.'bɔ] 'window' [bub.'bu] 'feather'
/a/	[ʔap.'pat] 'four' [ʔup.'pat] 'to jump'	[ʔaŋ.'ŋat] 'to wait' [ʔaŋ.'ŋet] 'to bite'	[hal.'la] 'to fry rice' [hul.'lu] 'to command'

Moreover, all five Bagobo-Klata vowels can occur in all syllable positions, as exemplified in Table 8.

Table 8: Bagobo-Klata Vowels in Antepenultima, Penultima, and Ultima

	Antepenultima	Penultima	Ultima
/i/	[ma.li.gɔn.'nɔj] 'beautiful' [pi.jas.'su] 'spear'	['ti:.ŋut] 'greedy, selfish' ['li.tu:] 'mole, birthmark'	[lit.'ti] 'thunder' ['tu:.liʔ] 'earwax'
/u/	[pu.'wa:.las] 'forest, woods'	['ku:.lit] 'skin' [lul.'lut] 'a Bagobo-Klata dish'	['da:.mu] 'dew' ['gi:.bud] 'fontanelle'
/ɛ/	[tɛ.nɛ.'pɔŋ] 'mens' pants'	[tɛ.'kɔd] 'heel' [lɛt.'taʔ] 'sap, resin'	[la.'wɛ] 'tall (of a person)' [ʔɔŋ.'gɛt] 'clothing'
/ɔ/	[hɔ.mɔ.bɔ.'jɔʔ] 'ugly' [bɔl.'li:.jɔj] 'drunk'	[kɔ.'lat] 'thin (of a person)' [ʔɔl.'lɔŋ] 'neck'	[kɔp.'pɔ] 'chest' [kɔm.'mɔŋ] 'fist'
/a/	[ʔa.'ri:.jus] 'earrings' [tam.'ba:.ga] 'copper'	['ʔa:.wak] 'waist' [daŋ.'ŋɔw] 'handspan'	[bɔn.'na] 'true' [ka.'jang] 'bright, as light'

5 Syllable Structure

In Bagobo-Klata, the basic syllable structure is (C₁)C₂V(C₃): the obligatory syllable elements are the onset and the nucleus, while the coda and the extra onset, which forms a cluster with the obligatory onset, are optional. Shown in Table 9 are the four permissible syllable structures in Bagobo-Klata, along with examples.

Table 9: Syllable patterns in Bagobo-Klata

Syllable Structure	Examples
CV	[bɔ.'jɔʔ] 'face' [hi.'ju] 'elbow' ['pa:.'ʔɔ] 'foot' [mɔ.'tɔ] 'eye' ['gi:.'bud] 'hair whorl'
CVC	[klut.'tuŋ] 'forehead' [ŋil.'lɔ] 'ear' [ʔɔb.'buk] 'hair, of one's head' [ʔil.'lɔʔ] 'mole, birthmark' [bok.'kɔŋ] 'shin, of one's leg'
CCV	['gka:.'waŋ] 'hip' ['mlɛ:.'dɛ] 'yellow' ['klu:.'gɔʔ] 'wart' ['gna:.'lɔ] 'throat' ['klɛ:.'wɛŋ] 'eyebrow'
CCVC	[klam.'mag] 'star' ['mlɛ:.'ʔaŋ] 'rough' [bluŋ.'ŋus] 'mouth' [mlun.'nɔw] 'green' [kjuw.'wa] 'bee'

Moreover, Bagobo-Klata has consonant clusters or a sequence of two consonants. In terms of word positions, they can be complex onsets (i.e., two onsets occurring word-initially) or word-medial clusters (i.e., a coda plus an onset occurring word-medially). On the one hand, complex onsets can be segmental or derived. Shown in (9), (10), (11), (12), and (13) are the permissible combinations of C1 and C2 in Bagobo-Klata.

- (9) **a stop** + **a lateral**
 pl 'plɛ:.'tek 'wing'
 bl 'bli:.'bud 'whirlpool'
 kl 'kli:.'tut 'anus'
 gl 'glɔ:.'puj 'bee'
- (10) **a stop** + **a stop**
 gk 'gka:.'waŋ 'hip'
- (11) **a stop** + **an approximant**
 bj 'bjɑ:.'ʔɔw 'drizzle'
 kj 'kjɔ:.'hɔŋ 'cockroach'
 gj 'gja:.'wat 'pimple'
 kw 'kwa:.'lɔ 'worm, earthworm'
 gw 'gwa:.'li 'across, opposite side'
- (12) **a nasal** + **lateral**
 ml 'mlɔ:.'ʔɔs 'smooth'
- (13) **a stop** + **a nasal**
 gn 'gna:.'lɔ 'throat'

On the other hand, Bagobo-Klata has word-medial clusters, which can be homorganic (i.e., a word-medial consonant sequence of a nasal consonant and another consonant with an identical feature) or heterorganic (i.e., a sequence of two consonants with non-identical features). However, it must be noted that these clusters are rare because as will be seen in 8.6, historical clusters of PMP forms became geminates in Bagobo-Klata.

Table 10: Word-medial clusters

Homorganic	Heterorganic
[pan.'tug] 'bladder'	[kas.'pa] 'dandruff'
[ban.'tu:gan] 'famous'	[pak.'sul] 'hole (in the ground)'
[baŋ.'kil] 'canine tooth, fang'	[lug.'waʔ] 'outside'

Like phonemes and geminates, segmental complex onsets in Bagobo-Klata are also contrastive, as shown in (14).

- (14) *klɔm.*'mɔʔ 'tomorrow'
lɔm.'mɔʔ 'morning'

Finally, the number of syllables in Bagobo-Klata words depends on whether the words in question are unaffixed or derived. In roots, monosyllabic or disyllabic words are more common than trisyllabic ones, while in derived words, disyllabic words are as common as trisyllabic ones.

6 Segmental Stress

In this study, segmental stress is defined as the prominence (i.e., loudness) of a specific syllable in a given word (Crystal 2008). First, in Bagobo-Klata, segmental stress is contrastive, as shown in Table 11.

Table 11: Contrastive stress in Bagobo-Klata

Penultima	Ultima
[ʔu:.la] 'snake'	[ʔu.'la] 'rain'
[pɔ:.tɔj] to 'quarrel'	[pɔ.'tɔj] 'to kill'
[ba:.sa] to 'read'	[ba.'sa] 'squash'
[ʔɔ:.lat] 'scar'	[ʔɔ.'lat] 'vein'
[ku:.luŋ] to 'snore'	[ku.'luŋ] 'back'
[ʔa:.tiŋ] 'sweat'	[ʔa.'tiŋ] 'if'
[la:.wu] 'cloud'	[la.'wu] to 'fall'

Second, stress is also predictable because of vowel length. Table 12 shows that stress falls on the ultima (i.e., light (CV) or heavy (CVC)) if there is no vowel length.

Table 12: Stress in ultima

CV	CVC
[tɔ.'ʔi] 'to accompany'	[hi.'jɔw] 'nine'
[ʔɔ.'ka] 'what-you-call-it'	[ʔu.'lɔb] 'knee'
[da.'jɔ] 'west'	[lɔ.'tus] 'hundred'
[mɔ.'lɔ] 'light, in weight'	[gɔ.'tɔŋ] 'eggplant'
[lɔ.'wa] 'wide'	[dɔ.'lid] 'root'
[la.'wɛ] 'tall, of a person'	[ta.'lum] 'papaya'
[pa.'ja] 'big'	[ʔu.'wɔj] 'rattan'
[ba.'sa] 'pumpkin, squash'	[pa.'ʔɛt] 'bitter'
[mɔ.'ʔɔ] 'betel, areca nut'	[ʔɔ.'lɔk] 'to kiss'
[ka.'ʔa] 'to eat'	[lɔ.'ŋuj] 'to swim'

On the other hand, Table 13 shows that if a word has vowel length, stress falls on the penultima.

Table 13: Stress in penultima

CV	CCV
['du:.luʔ] 'blood'	['gla:.ʔu] 'throat'
['pu:.huŋ] 'heart'	['kli:.tut] 'anus'
['bɔ:.gɔk] 'disease'	['gja:.wat] 'pimple'
['ka:.gɔw] 'microbe'	['mlɛ:.dɛ] 'yellow'
['ma:.liʔ] 'good'	['gnu:.wɔ] 'earth'
['hu:.kɔ] 'anger'	['bjɔ:.ʔɔ] 'year'
['ka:.pɔ] 'cold, of weather'	['blɛ:.ʔɛ] 'gecko'
['ba:.lu] 'provision, as for a journey'	['kli:.giʔ] 'hawk'
['ma:.ʔa] 'sheath, as of a bolo/knife'	['blɔ:.waʔ] 'spider'
['la:.ʔi] 'man, male'	['gka:.waŋ] 'hip'

7 Morphophonological Processes

This section discusses several morphophonological processes in Bagobo-Klata. In this study, they are called as such because they (except reduplication and cliticization) are mostly triggered by affixation. As will be seen in the following subsections, one affix, usually verbal, can trigger up to seven processes.

7.1 Syllable Deletion

In Bagobo-Klata, syllable deletion is a morphophonological process that deletes the least prominent syllable (or the antepenultima) in an affixed word. There are three identified affixes in Bagobo-Klata that trigger syllable deletion, namely, the irrealis patient-voice (PV) -ɔ, the irrealis benefactive/locative-voice (B/LV) -a, and the nominalizing -a.

There are only a few verbs to which the PV and LV verbal affixes -ɔ and -a can be attached. Once attached to verbs, these affixes trigger syllable deletion. In (15) to (17), the verb roots are all disyllabic and, after affixation, become trisyllabic, but -ɔ deletes the antepenultimas (i.e., the ones in bold), while the stress shifts to the penultima.

- (15) 'ʔɛ:.jap + -ɔ → 'ʔɛ:.ja.pɔ → 'ja:.pɔ
to count count-IRR.PV count-IRR.PV
- (16) ha.'wet + -ɔ → ha.'wɛ:.tɔ → 'wɛ:.tɔ
to hook hook-IRR.PV hook-IRR.PV
- (17) pep.'pɛʔ + -ɔ → pep.'pɛ:.ʔɔ → 'pɛ:.ʔɔ
to wash wash-IRR.PV wash-IRR.PV

Shown below are Bagobo-Klata verbs affixed with *-a*, which deletes all the antepenultimas.

- (18) 'ta:.waŋ + *-a* → 'ta:.wa.ŋa → 'kwa:.ŋa
to help help-IRR.BV help-IRR.BV
- (19) 'ka:.pɛʔ + *-a* → 'ka:.pɛ.ʔa → 'pɛ:.ʔa
to hold hold-IRR.BV hold-IRR.BV
- (20) 'ha:.kɔj + *-a* → 'ha:.kɔ.ja → 'kɔ:.ja
to ride ride-IRR.LV ride-IRR.LV

Moreover, it will be seen in the subsequent processes (e.g., assimilation, gemination, epenthesis, fortition, and lenition) that syllable deletion interacts with them and is required to take place before or after them.

7.2 Vowel Harmony

Vowel harmony is a morphophonological process in which two adjacent vowels become similar (Crowley & Bowerman 2010). The affixes that trigger this process are the verbal affixes *-ɔ* and *-a*. Derived words such as *kɔnɔ* and *ŋaja* will be used to explain this process.

In deriving the patient-voice affixed verb *kɔnɔ*, there are four morphophonological processes at play, which will be explained chronologically. First, after *-ɔ* is affixed to *ka. 'ʔa* to eat, the historically deleted word-final *n is reinstated.

- (21) Epenthesis
ka. 'ʔa + *-ɔ* → ka. 'ʔa:.no
to eat IRR.PV eat-IRR.PV

After (21), the onset /ʔ/ in the penultima becomes [k], because of adjacent assimilation to the onset /k/ of the antepenultima.

- (22) Adjacent Assimilation
ka. 'ʔa:.no → ka. 'ka:.nɔ
eat-IRR.PV eat-IRR.PV

Third, the antepenultima [ka] is deleted, because of *-ɔ*.

- (23) Syllable Deletion
ka. 'ka:.nɔ → 'ka:.nɔ
eat-IRR.PV eat-IRR.PV

Fourth, in the disyllabic form, /a/ in the penultima becomes [ɔ] because *-ɔ* triggers vowel harmony.

- (24) Vowel Harmony
'ka:.nɔ → 'kɔ:.nɔ
eat-IRR.PV eat-IRR.PV

The chronology discussed seems surprisingly incredible because it would be more convenient for Bagobo-Klata speakers to delete the penultima of [ka. 'ʔa:.nɔ], instead of the antepenultima. However, it can be seen in previous examples in 7.1 that the syllables *-ɔ* or *-a* prefer to delete in affixed verbs is the antepenultima. Thus, in (23), before the antepenultima [ka] is deleted, the onset of the penultima assimilates to that of the antepenultima. Now, regarding the vowel harmony of [a] in the penultima ['ka:] to [ɔ], it can be a case of reinstatement of a Bagobo-Klata vowel that historically shifted. As will be explained in 8.5., /ɔ/ and /a/ are reflexes of PMP *e and *a, respectively. Both proto-phonemes

sporadically assimilated to each other in some proto-forms. In the case of /ka.'ʔa/ as the reflex of PMP *kaen, *e is argued to have assimilated to *a, hence becoming /a/, but its historically unassimilated form *ɔ is reinstated through vowel harmony.

- (25) Affixation
ʔaŋ. 'ɲɔj + -a → *ʔaŋ. 'ɲɔ.ja*
 to take IRR.AV
- (26) Syllable Deletion
ʔaŋ. 'ɲɔ.ja → *'ɲɔ:.ja*
- (27) Vowel Harmony
'ɲɔ:.ja → *'ɲa:.ja*
 take-IRR.PV

In the case of *ʔaŋ. 'ɲɔj* ‘to take’, there are only two processes involved: syllable deletion and vowel harmony. Perhaps, what can be deduced here is that vowel harmony requires syllable deletion to take place first, at least, in Bagobo-Klata verbs affixed with -ɔ and -a.

7.3 Gemination

Gemination is also a morphophonological process in which a singleton becomes a geminate (Campbell, 2013). As previously discussed, geminates are not only segmental but also derived through affixes such as the irrealis PV -ɔ, the realis PV and B/LV *bɔ-C~*(...-a), the realis instrumental voice (IV) *pɔ-C~*, and the realis abilitative/potentive *kɔ-C~*.

First, -ɔ does not apply to all verbs. It appears to require a verb that has a heavy ultima with /ɔ/ as its nucleus, although it can also be /ɛ/ or /a/. Like vowel harmony, syllable deletion precedes gemination, so there are three processes involved after the -ɔ affixation in (28).

- (28) Affixation
- | | | |
|-------------------|---|-----------------------|
| <i>ʔɔp. 'pɔs</i> | → | <i>ʔɔp. 'pɔ.sɔ</i> |
| to borrow-IRR.PV | | borrow-IRR.PV |
| <i>ʔɔm. 'mɔt</i> | → | <i>ʔɔm. 'mɔ.tɔ</i> |
| to catch | | catch-IRR.PV |
| <i>kɔk. 'kɔt</i> | → | <i>kɔk. 'kɔ.tɔ</i> |
| to dig | | dig-IRR.PV |
| <i>'nɔ.tɔ</i> | → | <i>nɔt. 'tɔ</i> |
| to ask a question | | ask a question-IRR.PV |

At least, in (28), it can be noticed that after affixation, syllable deletion and degemination takes place simultaneously.

- (29) Syllable Deletion
- | | | |
|-----------------------|---|-----------------------|
| <i>ʔɔp. 'pɔ.sɔ</i> | → | <i>'pɔ.sɔ</i> |
| borrow-IRR.PV | | borrow-IRR.PV |
| <i>ʔɔm. 'mɔ:.tɔ</i> | → | <i>'mɔ.tɔ</i> |
| catch-IRR.PV | | catch-IRR.PV |
| <i>kɔk. 'kɔ:.tɔ</i> | → | <i>'kɔ.tɔ</i> |
| dig-IRR.PV | | dig-IRR.PV |
| <i>'nɔ:.tɔ</i> | → | <i>'nɔ.tɔ</i> |
| ask a question-IRR.PV | | ask a question-IRR.PV |

(30) Gemination			
	<i>'pɔ.sɔ</i>	→	<i>pɔs. 'sɔ</i>
	borrow-IRR.PV		borrow-IRR.PV
	<i>'mɔ.tɔ</i>	→	<i>mɔt. 'tɔ</i>
	catch-IRR.PV		catch-IRR.PV
	<i>'kɔ.tɔ</i>	→	<i>kɔt. 'tɔ</i>
	dig-IRR.PV		dig-IRR.PV
	<i>'nɔ.tɔ</i>	→	<i>nɔt. 'tɔ</i>
	ask a question-IRR.PV		ask a question-IRR.PV

Second, in (31-34), it is shown that Bagobo-Klata allows two geminates in one word through *bɔ-C~*, *bɔ-C~...-a*, *pɔ-C~*, and *kɔ-C~*.

(31)	<i>bɔ-</i>	+	<i>kɔk. 'kɔt</i>	→	<i>bɔb.bɔl.li</i>
	RLS.PV		to dig		RLS.PV-dig
			<i>hap. 'pɔj</i>	→	<i>bɔs.sap. 'pɔj</i>
			to hang		RLS.PV-hang
			<i>tat. 'tad</i>	→	<i>bɔt.tat. 'tad</i>
			to mince		RLS.PV-mince
(32)	<i>bɔ-...-a</i>	+	<i>bul. 'las</i>	→	<i>bɔl.las. 'sa</i>
	RLS.BV		to change		RLS-change-BV
			<i>ka 'ʔa</i>	→	<i>bɔk.kan. 'na</i>
			to eat		RLS-eat-LV
(33)	<i>pɔ-</i>	+	<i>tɔk. 'kub</i>	→	<i>pɔtɔkkub</i>
	RLS.IV		to cover		RLS.IV-cover
			<i>lag. 'gas</i>	→	<i>pɔl.lag. 'gas</i>
			to wash		RLS.IV-wash
(34)	<i>kɔ-</i>	+	<i>lɔŋ. 'ŋaʔ</i>	→	<i>kɔl.lɔŋ. 'ŋaʔ</i>
	RLS.ABI		to flee		RLS.ABIL-flee
			<i>lut. 'tu</i>	→	<i>kɔl.lut. 'tu</i>
			to cook		RLS.ABIL-cook

7.4 Epenthesis

Epenthesis is defined as the insertion of a speech sound, typically a consonant, in a given word due to affixation. This happens if a suffix is attached to a word with a light ultima, and because vowel sequences are impermissible in Bagobo-Klata, there must be an obligatory onset to intervene that takes the form of an epenthetic consonant.

Epenthetic forms [n] and [l] as in (35) have traceable origins. As previously stated, word-final *n and *l are historically deleted sounds in Bagobo-Klata, but they are reinstated through affixation, specifically epenthesis. In the examples below, word-final [n] and [l] serve as the respective onset of each ultima.

(35) Epenthesis					
	<i>ʔu. 'la</i>	+	<i>-a</i>	→	<i>ʔu. 'la.na</i>
	rain		IRR.LV		rain-IRR.LV
	<i>bu. 'na</i>	+	<i>-ɔ</i>	→	<i>bu. 'na.lɔʔ</i>
	to hit		IRR.PV		to hit-IRR.PV

As seen in (35), epenthesis precedes syllable deletion, and the same can be said for affixed verbs such as *kolana* ‘to rain’ and *nalo* ‘to hit’.

- (36) Syllable Deletion
ʔu. 'la.na → *'la:.na*
 rain-IRR.LV rain-IRR.LV
bu. 'na.lɔʔ → *'na.lɔʔ*
 hit-IRR.PV hit-IRR.PV

- (37) Affixation
kɔ- + *'la:.na* → *kɔ. 'la:.na*
 POT rain-IRR.LV POT.IRR-rain-LV

However, there are epenthetic forms with obscure origins such as in (38).

- (38) Epenthesis
'hɛ:.lɛ + *-ɔ* → *'hɛ:.lɛ.lɔ*
 to lean IRR.PV
tɔ. 'ʔi + *-a* → *tɔ. 'ʔi.na*
 to accompany IRR.BV

In *tɔ. 'ʔi*, */ʔ/* assimilates to */t/*, becoming [t] after epenthesis.

- (39) Syllable Deletion
'hɛ:.lɛ.lɔ → *lɛ:.lɔ*
 lean-IRR.PV lean-IRR.PV
tɔ. 'ti.na → *'ti:.na*
 accompany-IRR.BV accompany-IRR.BV

Then, in more examples below, it is seen that syllable deletion precedes epenthesis and that these epenthetic forms [k, g], which act as the C1 in complex onsets, have even more obscure origins.

- (40) Affixation
'ta:.waŋ + *-a* → *'ta:.wa.ŋa*
 to help IRR.BV help-IRR.BV
'bɛ:.nɛʔ + *-a* → *bɛ:.nɛ.ʔa*
 to cry IRR.BV help-IRR.BV
'dɔ:.lɔŋ + *-ɔ* → *'dɔ:.lɔ.ŋɔ*
 to honor IRR.PV honor-IRR.PV

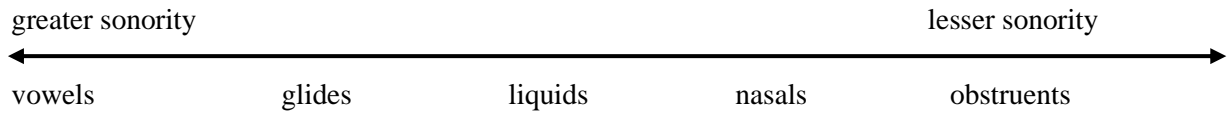
- (41) Syllable Deletion
'ta:.wa.ŋa → *wa:.ŋa*
 help-IRR.BV help-IRR.BV
bɛ:.nɛ.ʔa → *nɛ:.ʔa*
 help-IRR.BV help-IRR.BV
'dɔ:.lɔ.ŋɔ → *lɔ:.ŋɔ*
 honor-IRR.PV honor-IRR.PV

- (42) Epenthesis
wa:.ŋa → *'kwa:.ŋa*
 help-IRR.BV help-IRR.BV
nɛ:.ʔa → *'gɲɛ:.ʔa*
 cry-IRR.BV cry-IRR.BV
lɔŋ. 'ŋɔ → *glɔŋ. 'ŋɔ*
 honor-IRR.PV honor-IRR.PV

7.5 Fortition

Fortition refers to the morphophonological process in which a speech sound becomes more sonorous or acoustically louder. In general, sonority or acoustic loudness of speech sounds can be gauged through a hierarchy. As shown in Figure 1, the sounds to the left are more sonorous, while the ones to the right are less sonorous (Hayes 2009:75).

Figure 1: The Sonority Hierarchy



As mentioned in 2.2., the word-initial /ʔ/ becomes [gg, tt, ll, jj, ww] because of certain affixes that trigger gemination. To explain this, it will be argued in this study that in Bagobo-Klata, gemination requires more sonorous consonants, so /ʔ/ must undergo fortition to be eligible for gemination. To illustrate this argument clearly, derived forms such as *jɔda* ‘chair’ and *wɔga* ‘bathroom’ will be used. Note that these forms are verbal and nominal respectively.

(43) Affixation

'ʔɛ:.ʔɔd+	-a	→	'ʔɛ:.ʔɔ.da
to sit	IRR.LV		sit-IRR.LV
ʔa.'ʔɔg	-a	→	ʔa.'ʔɔ.ga
to bathe	NOM		bathe-NOM

(44) Syllable Deletion

'ʔɛ:.ʔɔ.da	→	'ʔɔ:.da
sit-IRR.LV		sit-IRR.LV
ʔa.'ʔɔ.ga	→	'ʔɔ:.ga
bathe-NOM		bathe-NOM

(45) Fortition

'ʔɔ:.da	→	'jɔ:.da
sit-IRR.LV		sit-IRR.LV
'ʔɔ:.ga	→	'wɔ:.ga
bathe-NOM		bathe-NOM

Since *bɔ-* is established to trigger gemination, it would make sense for the word-initial glottal in <oda> to undergo fortition, hence <yoda>. Thus far, what is certain is that [g] expresses patient voice, while [j], potentiative. (5) and (6) in 4. are repeated here as (46) and (47).

(46) <i>bɔ-</i>	+	'ʔɛ:.pɔk	→	<i>bɔg.</i> 'gɛ:.pɔk
RLS.PV		to cut		RLS.PV-cut

(47) <i>bɔ-</i>	+	'ʔɛ:.pɔk	→	<i>bɔj.</i> 'jɛ:.pɔk
RLS.POT		to cut		RLS.POT-cut

7.6 Lenition

Lenition is the opposite process of fortition, that is, more sonorous to less sonorous. In Bagobo-Klata, -a triggers word-final /s/ in the examples below to lenite.

(48) Affixation				
<i>lag. 'gas+</i>	<i>-a</i>	→	<i>lag. 'ga.sa</i>	
to wash	NOM		wash-NOM	
<i>kik. 'kis</i>		→	<i>kik. 'ki.sa</i>	
to scrape			scrape-NOM	
<i>lag. 'gas</i>	<i>-ɔ</i>	→	<i>lag. 'ga.sɔ</i>	
to wash	IRR.PV		wash-IRR.PV	

As seen in (48), there are two processes involved: after affixation, the antepenultimas in (49) are deleted. Then, in (50), the singleton /s/ lenites to /h/.

(49) Syllable Deletion			
<i>lag. 'ga.sa</i>	→	<i>'ga:.sa</i>	
wash-NOM		wash-NOM	
<i>kik. 'ki.sa</i>	→	<i>'ki:.sa</i>	
scrape-NOM		scrape-NOM	
<i>lag. 'ga.sɔ</i>	→	<i>'ga:.sɔ</i>	
wash-IRR.PV		wash-IRR.PV	

(50) Lenition			
<i>'ga:.sa</i>	→	<i>'ga:.ha</i>	wash-NOM (sink)
<i>'ki:.sa</i>	→	<i>'ki:.ha</i>	scrape-NOM (coconut scraper)
<i>'ga:.sɔ</i>	→	<i>'ga:.hɔ</i>	wash-IRR.PV

Perhaps, the lenition of /s/ in the examples above is a case of reinstatement of the historically deleted word-final *h. Through lenition, word-final /s/ synchronically reverts to [h].

7.7 Syncope

Syncope is a morphophonological process in which a medial segment, typically a vowel, is deleted (Crowley & Bown, 2010). In Bagobo-Klata, the unstressed nucleus in penultima of disyllabic words such as /bɔl. 'lɔj/ and /bɔl. 'li/ is deleted when these words are affixed with -ɔ.

(51) Affixation				
<i>bɔl. 'lɔj</i>	+	<i>-a</i>	→	<i>bɔl. 'lɔ.ja</i>
to give		IRR.BV		give-IRR.BV
<i>bɔl. 'li</i>	+	<i>-a</i>	→	<i>bɔl. 'li.ja</i>
to buy		IRR.BV		buy-IRR.BV
<i>bɔl. 'li</i>	+	<i>-ɔ</i>	→	<i>bɔl. 'li.jɔ</i>
to buy		IRR.PV		buy-IRR.PV
(52) Degemination				
<i>bɔl. 'lɔ.ja</i>	→	<i>bɔ. 'lɔ.ja</i>		
give-IRR.BV		give-IRR.BV		
<i>bɔl. 'li.ja</i>	→	<i>bɔ. 'li.ja</i>		
buy-IRR.BV		buy-IRR.BV		
<i>bɔl. 'li.jɔ</i>	→	<i>bɔ. 'li.jɔ</i>		
buy-IRR.PV		buy-IRR.PV		

As can be noticed in (53), syncoated nuclei in the antepenultima derive complex onsets.

(53) Vowel Syncope

bɔ. 'lɔ.ja → *blɔ. 'ja*
give-IRR.BV give-IRR.BV

bɔ. 'li.ja → *bli. 'ja*
buy-IRR.BV buy-IRR.BV

bɔl. 'li.jɔ → *bli. 'jɔ*
buy-IRR.PV buy-IRR.PV

7.8 Degemination

Degemination is a process in which a geminate becomes a singleton. In the data set, the distributive numeral affix /tɔ-/ and the irrealis benefactive/locative -a are known to reduce the geminate in a root to a singleton.

- (54) *tɔ-* + *ʔuw. 'wɔ* → *tɔl. 'lu:wɔ*
DN two two each
- tɔl. 'lu* → *tɔt. 'tɔ:lu*
three three each
- ʔap. 'pat* → *tɔl. 'la:pat*
four four each

7.9 Nasal Substitution

In Austronesian languages, nasal substitution is a very common morphophonological process (Blust, 2004). In Bagobo-Klata, this process triggers the following affixes to replace word-initial sounds (except /n, ŋ, j, w/) of roots: irrealis actor-voice (AV) *m-* and its realis counterpart (*bɔn*)*n-*, irrealis PV *məm-/mɔm-*, ir/realis B/LV *tam/m-*, and instrument nominalizer *tam-*.

- (55) *m-* + *bɔl. 'lɔj* → *mɔl. 'lɔj*
IRR.AV to give IRR.AV-give
- pan. 'nek* → *man. 'nek*
to climb IRR.AV-climb
- dɔd. 'dɔŋ* → *mɔd. 'dɔŋ*
to approach IRR.AV-approach
- 'ha:kɔj* → *ma:kɔj*
to ride IRR.AV-ride
- tɪd. 'duʔ* → *mɪd. 'duʔ*
to stand IRR.AV-stand

- (56) *(bɔn)n-* + *m-(ʔ)up. 'pat* → *nup. 'pat*
 RLS IRR.AV-jump RLS.AV-jump
- [m-(l)ɔ. 'ʔɔw* → *nɔ. 'ʔɔw*
 IRR.AV-walk RLS.AV-walk
- m-(s)ɔs. 'sɔp* → *nɔssɔp*
 IRR.AV-sip RLS.AV-sip
- m-(k)ɔm. 'mɔs* → *nɔmmɔs*
 IRR.AV-squeeze RLS.AV-squeeze
- m-(g)aj. 'ɲet* → *naj. 'ɲet*
 IRR.AV-bite RLS.AV-bite

In (55) and (56), it is shown that eleven (11) Bagobo-Klata consonants can be replaced through nasal substitution and that homorgany is not required.

7.10 Reduplication

Reduplication in Bagobo-Klata occurs only when adjectives undergo intensification. Specifically, the penultima or antepenultima reduplicates, and regardless of the syllable structure, the reduplicated syllable is always open.

- (57) *lɔ. 'wa* → *lɔ~lɔ. 'wa*
 wide very~wide
- (58) *pa. 'ja* → *pa~pa. 'ja*
 big very~big
- (59) *tɔ. 'wɔʔ* → *tɔ~tɔ. 'wɔʔ*
 fat very~fat
- (60) *ma.li.gɔn. 'nɔj* → *ma~ma.li.gɔn. 'nɔj*
 beautiful very~beautiful

7.11 Cliticization

Function words, especially those behaving like clitics, tend to attach or to cliticize to a word called a host. In Bagobo-Klata, words that often cliticize are non-personal nominal markers such as /kɛŋ/ and /nɛŋ/ and, of course, the relativizer /ŋɔ/. The plural marker /bɛ/ is also known to cliticize but only to /ʔaŋŋaʔ/.

- (61) *hɔ. 'mɔʔ kɛŋ=bɔ. 'jɔʔ=ni. 'ja* → *hɔ. 'mɔk bɔ. 'jɔʔ ni. 'ja*
 bad ABS=face=3SG.GEN
 “His/her face is ugly.”

In (61), when /kɛŋ/ cliticizes to the word it follows, it becomes [k].

- (62) *m-(k)a. 'ʔa=ʔu nɛŋ=pi. 'hit* → *ma. 'ʔa ʔuŋ pi. 'hit*
 INTR.IRR-eat=1SG.ABS GEN=shrimp
 “I will eat shrimp.”

- (63) 'ma:liʔ=ŋɔ=lɔm.'mɔʔ → 'ma:liŋ lɔm.'mɔʔ
 good=LKR=morning
 “Good morning”

In (62) and (63), when cliticized, both /nɛŋ/ and /ŋɔ/ become [ŋ].

- (64) bɛ=ʔaŋ.'ŋaʔ → bjaŋ.'ŋaʔ
 PL=child
 “Children”

In (64), when the plural marker /bɛ/ cliticizes to the word it precedes, what is retained of it is only the word-initial sound /b/. Then, the word-initial glottal stop in /ʔaŋ.'ŋaʔ/ becomes a palatal sound, while /ɛ/ is deleted. However, it must be noted that /bɛ/ cliticizes to /ʔaŋ.'ŋaʔ/ only.

8 Historical Phonology

This section discusses how the phonological system of Bagobo-Klata developed from PMP. Only the salient reflexes of PMP phonemes in Bagobo-Klata are explained. The proto-forms are based on Blust and Trussel (Ongoing).

8.1 The Reflex of PMP *n

PMP *n was retained as /n/ in Bagobo-Klata word-initially and -medially but was lost word-finally, as in (65).

- (65) PMP *bulan > *bula* ‘moon, month’
 PMP *bulaw-an > *blawa* ‘gold’
 PMP *dahun > *daʔu* ‘leaf’
 PMP *duRi-an > *duliya* ‘durian fruit’
 PMP *haRezan > *ʔadda* ‘ladder’
 PMP *ipen > *ʔippɔ* ‘tooth’
 PMP *qutin > *ʔuti* ‘penis’
 PMP *quzan > *ʔula* ‘rain’
 PMP *tian > *tiya* ‘belly’
 PSP *libun > *libu* ‘woman’

8.2 The Reflex of PMP *R and *l

PMP *R and *l merged as /l/ in word-initial and word-medial positions.

- (66) PAN *kaRi > *kɔli* ‘language, word, to say’
 PMP *beRay > *bɔllɔj* ‘to give’
 PMP *kulit > *kulit* ‘skin’
 PMP *laki > *laʔi* ‘man, male’
 PMP *lesuŋ > *lɔssuŋ* ‘rice mortar’
 PMP *Ratus > *lɔtus* ‘hundred’
 PMP *Rabun > *lawu* ‘drizzling rain, mist; fog’
 PMP *tuli > *tuliʔ* ‘earwax’
 PMP *uRat > *ʔɔlat* ‘vein’
 PMP *zuRuq > *duluʔ* ‘blood’
 PWMP *Runaw > *lunɔw* ‘landslide’

However, in the word-final position, both *R and *l were phonemically lost:

- (67) PAN *qebel > *ʔəbbə* ‘smoke’
 PMP *bitil > *witi* ‘hungry’
 PMP *hulaR > *ʔula* ‘snake’
 PMP *gatel > *gattə* ‘itchy’
 PMP *suqaR > *həʔa* ‘thorn’

However, when *R is metathesized, it is preserved as, in (68).

- (68) PMP *keseR > *kəlləs* ‘strength, force, vigor’

8.3 The Reflexes of PMP *b

PMP *b has two reflexes in Bagobo-Klata: /b/ and /w/. Shown in (69) are examples of the retention of PMP *b.

- (69) *b > b
 PMP *batuk > *batuk* ‘cough’
 PMP *baRah > *ballə* ‘embers, glowing coals’
 PMP *tekub > *təkkub* ‘to cover’

Like in the Tëduray and Danaw languages, in Bagobo-Klata, PMP *-b- became /w/ intervocalically.

- (70) *b > w
 PMP *besuR > *wəssu* ‘satiated’
 PMP *belay > *wəlj* ‘tired, weary’
 PMP *qabaRa > *wəllə* ‘shoulder’
 PMP *qubi > *ʔuwwi* ‘yam’
 PMP *tubuq > *tuwuʔ* ‘to grow’
 PMP *tabaŋ > *tawaŋ* ‘help’
 PMP *tabeq > *təwəʔ* ‘fat; grease’
 PMP *ibeR > *ʔiwwə* ‘envy’
 PPh *abaka > *wəʔə* ‘Manila hemp’

8.4 The Reflexes of PMP *s

PMP *s split into /h/ and /s/ only in word-initial and word-medial positions.

- (71) PAN *sikux > *hiju* ‘elbow’
 PMP *pusuŋ > *puhuŋ* ‘heart’
 PMP *Rusuk > *luhuk* ‘ribcage’
 PMP *saqebit > *hawet* ‘to hook on to something’
 PMP *sakay > *hakəy* ‘to ride on something’
 PMP *suat > *huwat* ‘to comb one's hair’

In word-final position, it was retained as /s/, which sporadically lenited to [h] through nominal or verbal affixation, as in seen 7.6.

- (72) PMP *hapejes > *pəddəs* ‘painful’
 PMP *kemes > *kəmməs* ‘to squeeze’
 PMP *tebus > *təbbus* ‘to redeem’

8.5 The Reflexes of PMP *e and *a

PMP *a and *e merged and became /a/ and /ɔ/ in Bagobo-Klata. This can be seen in both penultima and ultima.

- (73) PMP *batu > *bɔtu* ‘stone, testicle’
 PMP *ina > *ʔinɔ* ‘mother’
 PMP *manuk > *mɔnuk* ‘chicken’
 PMP *mata > *mɔtɔ* ‘eye’
 PMP *qatay > *ʔɔtɔy* ‘liver’
 PSP *igaʔ > *ʔigɔʔ* ‘betel leaf’
 PSP *samaʔ > *hɔmɔʔ* ‘bad’
- (74) PMP *beli > *bolli* ‘to buy something’
 PMP *enem > *ʔonnom* ‘six’
 PMP *lebeŋ > *lobboŋ* ‘to bury’
 PMP *sejem > *sɔlɔm* ‘sugar ants’
 PMP *telu > *tollu* ‘three’
 PMP *tebuh > *tobbu* ‘sugarcane’

There are instances in which PMP *e sporadically assimilated to PMP *a, as in (75).

- (75) *e > *ɔ > a
 PAN *kaen > *kaʔa* ‘to eat’
 PMP *beRas > *ballas* ‘uncooked rice’
 PMP *epat > *ʔappat* ‘four’
 PSP *eŋaʔ > *ʔaŋŋaʔ* ‘child’

Then, PMP *a sporadically assimilated PMP *e, as in (76).

- (76) *a > ɔ
 PMP *hajek > *ʔɔlɔk* ‘to kiss’
 PMP *qatep > *ʔɔtɔp* ‘roof’
 PMP *tazem > *tɔlɔm* ‘sharp and pointed’

Finally, shown in (77) are retentions of PMP *a.

- (77) PMP *qatiŋ > *ʔatiŋ* ‘sweat, perspiration’
 PMP *palaj > *palad* ‘palm (of a hand, foot)’
 PWMP *bala(n)tik > *blatik* ‘a spring-set spear trap’
 PWMP *balun > *balu* ‘provisions for a journey’
 PPh *bakkak > *pabbak* ‘frog’
 PSP *aŋat > *ʔaŋŋat* ‘to wait’

8.6 Historical Geminaton

In Austronesian languages, consonant gemination has three well-established sources (Blust 1995:125-127): elimination of an articulatory gesture, syncope of an unstressed vowel, and consonant lengthening after schwa. All these three sources can account for the historical development of gemination in Bagobo-Klata.

First, shown in (78) are Bagobo-Klata geminates that developed through the elimination of an articulatory gesture (i.e., manner, place, or both) or the assimilation of homorganic or heterorganic consonants. In (78), the homorganic consonants *mp, *nd, and *ŋd respectively became /pp/ and /dd/ through regressive assimilation.

- (78) PMP *ampu > *ʔappuʔ* ‘grandparent’
 PMP *empu > *ʔɔppu* ‘grandchild’
 PMP *daŋdaŋ > *dɔddɔŋ* ‘near’
 PMP *sampay > *happɔj* ‘to hang clothes’
 PMP *tindes > *tiddɔs* ‘to crush lice with the fingernails’

Then, in (79), the heterorganic consonants *pk, *sk, and *dk all became /kk/, while *ps and **kt, /ss/ and /tt/ respectively.

- (80) PAN *sepsep > *sɔssɔp* ‘to sip’
 PAN *tektek > *tɔttɔk* ‘house lizard’
 PMP *kaskas > *kakkas* ‘to scratch’
 PMP *kepkep > *kɔkkɔp* ‘to embrace’
 PMP *panahik > *pannek* ‘to climb’
 PPh *kakkad > *kɔkkɔt* ‘to dig’

Second, Bagobo-Klata geminates also developed from clusters of non/identical consonants through vowel syncope followed by regressive or progressive assimilation, as in (81).

- (81) PMP *qalejaw > *ʔaddɔw* ‘day, sun (*lj > dd)’
 PMP *qahelu > *ʔɔllu* ‘pestle (*hl > ll)’
 PMP *bageqaj > *bagganj* ‘molar tooth (*gq > gg)’
 PMP *saŋelaR > ‘*halla* to fry rice (*ŋl > ll)’
 PMP *tuqelan > *tulla* ‘bone (*ql > ll)’
 PMP *tuq(e)lid > *tullid* ‘straight (*ql > ll)’
 PWMP *benehiq > *binniʔ* seedling (*nh > nn)
 PWMP *seŋ(e)qaw > *hɔŋŋɔw* ‘steam, vapor (*ŋq > ŋŋ)’
 PWMP *taq(e)baŋ > *tabbaŋ* tasteless, bland (*qb > bb)’
 PPh *bulbul > *bubbu* ‘hair (of one’s body) (*lb > bb)’
 PPh *luh(e)naw > *lunnɔw* ‘green, as vegetation (*hn > nn)’

Third, as seen in 8.5, PMP singletons became geminates usually after PMP *e and sometimes before and between PMP *e

- (82) *e_
 PAN *lemek > *lɔmmɔk* ‘soft’
 PMP *bener > *bɔnna* ‘true’
 PMP *betem > *bɔttɔm* millet’
 PMP *qelet > *ʔɔllɔt* ‘between’
 PMP *tebuh > *tɔbbu* sugarcane’
 PMP *teŋaq > *tɔŋŋɔ* ‘half’
 PPh *Rebaq > *lɔbbɔ* ‘to collapse’
 PPh *tebaR > *tɔppɔ* ‘to answer (*R > ø)’

- (83) *_e
 PMP *dalem > *dallɔm* ‘inside’
 PMP *liteq > *lettaʔ* ‘sap (of a tree)’
 PMP *qulej > *ʔullɔd* ‘larva’
 PMP *piqek > *piyyak* ‘baby chick’
 PWMP *qali-matek > *mattɔk* ‘leech’
 PSP *parek > *pallɔk* ‘sand’

It must be noted that not all PMP singletons before, after, and between PMP *e became geminates in Bagobo-Klata, as shown in (84).

- (84) PMP *beŋel > *bɔŋɔ* ‘deaf’
 PMP *etaq > *ʔataʔ* ‘rice bran’
 PMP *heyup > *ʔijup* ‘to blow, using one’s mouth’
 PMP *lisehaq > *lihɔʔ* ‘nit, egg of a hair louse’
 PMP *peŋa > *pɔŋɔʔ* ‘branch (of a tree)’

Finally, there are quite a few Bagobo-Klata words with geminates, which cannot be accounted for by these sources, as shown in (85), although the singletons in *aNay, *duma, *kita, *talaw, *qalima, *baba, *hikam, *iak, and *laña probably geminated because of the merger of PMP *e and *a. Another curious case here is that of PMP *b and PPh *b, both of which became /pp/, as in *bappa* < PMP *baba and *ʔapput* < PPh *qabut.

- (85) PAN *aNay > *ʔannɔj* ‘termite’
 PAN *duma > *dummɔ* ‘other’
 PAN *kita > *kittɔ* ‘to see, to show’
 PAN *pitu > *pittu* ‘seven’
 PAN *qalima > *limmɔ* ‘hand (c.f., PAN *lima > *limɔ* five)’
 PAN *talaw > *tallɔw* ‘fear’
 PMP *baba > *bappa* ‘to piggyback somebody’
 PMP *buni > *bunni* ‘to hide something’
 PMP *iak > *ʔiyyak* ‘to cry out’
 PMP *hikam > *ʔikkam* ‘mat (for sleeping)’
 PMP *kutu > *kuttu* ‘louse (of one’s head)’
 PMP *laña > *lannɔ* ‘cooking oil’
 PMP *qulu > *ʔullu* ‘head’
 PMP *qupa > *ʔuppɔ* ‘hen’
 PMP *puluq > *pulluʔ* ‘ten’
 PMP *putiq > *puttiʔ* ‘white’
 PMP *suRuq > *hullu* ‘to command, to send on an errand’
 PPh *piqqiq > *peppɛʔ* ‘to wash clothes’
 PPh *qabut > *ʔapput* ‘to reach something’
 PSP *bunal > *buna* ‘to beat somebody up’
 PSP *daŋaw > *daŋŋɔw* ‘handspan’

9 Conclusion

Synchronically, Bagobo-Klata has 20 phonemic speech sounds—15 consonants and five vowels. Besides speech sounds, geminates, consonant clusters, and segmental stress are shown to be contrastive. Moreover, geminates and consonant clusters both occur in roots and are derived through morphophonological processes. These are crucial in understanding the internal structure of Bagobo-Klata words. Further data will help unravel and elucidate the complicated nature of the fortified forms of the glottal stop as well as the epenthetic forms triggered by the affixes -a and -ɔ. Finally, by tracing the historical development of Bagobo-Klata phonemes from Proto-Malayo-Polynesian, the synchronic aspect of this phonological description has been calibrated to shed light on some of synchronic facts that can be sufficiently explained by historical facts.

References

- Blust, Robert. 1991. The Greater Central Philippine Hypothesis. *Oceanic Linguistics* 30.2:73-129.
 Blust, Robert. 1995. Notes on Berawan Consonant Gemination. *Oceanic Linguistics* 34.1:123-138.
 Blust, Robert. 2019. The Resurrection of Proto-Philippines. *Oceanic Linguistics* 58.2:153-256.

- Blust, Robert and Stephen Trussel. Ongoing. The Austronesian Comparative Dictionary (web edition). <https://www.trussel2.com/ACD/>
- Cagas, Luzviminda M. 1990. *Davao Minor Languages: A Description, Comparison, and Classification*. Unpublished PhD dissertation. University of the Philippines, Diliman, Quezon City.
- Crowley, Terry and Claire Bowern. 2010. *An Introduction to Historical Linguistics. Fourth edition*. Oxford University Press.
- Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics*. 6th edn. Massachusetts: Blackwell Publishing.
- Eberhard, David M., Simons, Gary F. and Charles D. Fennig (eds.). 2021. *Ethnologue: Languages of the World. Twenty-fourth edition*. SIL International. <https://www.ethnologue.com/>
- Evans, Angharad. 2017. *Phonology Essentials of Bagobo-Klata*. Unpublished MA thesis. University of Gloucestershire.
- McFarland, Curtis. 1994. Subgrouping and Number of the Philippine Languages, or How Many Philippine Languages Are There? *Philippine Languages* 25.
- Walton, Charles. 1979. A Philippine Language Tree. *Anthropological Linguistics* 21.2:70-98.
- Zorc, R. David. 1986. The Genetic Relationships of Philippine Languages. *FOCAL II: papers from the Fourth International Conference on Austronesian Linguistics*, ed. by Paul Geraghty, Louise Carrington, and Stephen Wurm, 147-173. Pacific Linguistics.
- Zorc, R. David. 2019. Klata/Giangan: A New Southern Philippine Subgroup. *The Archive: Special Publications No. 16 Current Studies in Philippine Linguistics*, 33-52.

Appendix A: Bagobo-Klata 1,026-Word List

	Bagobo-Klata	English
/p/	'pa:.tuk	duck
	pa.ka.'wan	cup
	pa.'ʔet	bitter
	'pa:.'ʔo	foot
	'pa:.'naʔ	bow
	pa.ha.gi.ra.paw.'wə	youngest child in a family
	pa.'lad lim.'mə	palm (of one's hand)
	pa.'lad 'pa:.'ʔo	sole (of one's foot)
	pa.'liʔ	wound, injury
	pa.'ja	big
	'pa:.'juŋ	umbrella
	pe.'taʔ	wet
	'pe:.'kəʔ 'pa:.'ʔo	back of knee
	pe.'ʔed	to carry, bring, take with oneself
	pe.'ʔəw	lame, crippled
	pi.'lə	how many, how much
	pi.jas.'su	spear
	pə	still, yet
	'pə:.'təj	quarrel; to fight someone
	pə.'təj	to die; to kill someone
	pə.kə.'ləg	how
	'pə:.'ʔes	to open something
	pə:.'naʔ	bait
	pə:.'ŋəʔ	branch
	pə.'ləs	many
	'pu:.'təw	iron (of metal)

Bagobo-Klata	English
pi. 'hit	shrimp
'pi:hiʔ	to pinch
pu. 'ʔuŋ	to wrap something up
'pu:huʔ	coconut heart
'pu:hu neŋ 'bu:lig	banana heart
'pu:huŋ	heart
'pu:lut	to pick something up
pu. 'wa:las	forest, woods
pak. 'sul	a hole in the ground
pan. 'tug	bladder
pab. 'bak	frog
pat. 'tig	to fracture, as a twig or one's arm
pad. 'deʔ	stinging pain
pad. 'dɔ	price, debt
pan. 'nek	to climb, as a mountain
paŋ. 'ŋa	to miss someone
pal. 'lək	sand
pas. 'sik	also, too
pəp. 'pɛʔ	to wash clothes
pip. 'pis	small (object)
pit. 'tu	seven
pis. 'sɔʔ	boil, abscess; to press or crush something
pij. 'jak	chick, baby fowl
pət. 'tɔj	firefly
pəd. 'dɔs	painful, sick
pən. 'nɔʔ	full, to the brim
pən. 'nuʔ	turtle
pəŋ. 'ŋɔ	to finish; after
pəŋ. 'ŋu:ʔɔ	chieftain
pəl.laŋ. 'ŋaʔ	born
pəl. 'ləd	acid, tart, sharp flavor
pud. 'du	gall bladder
pul. 'luʔ	ten
pul. 'luʔ hɔ. 'tu	eleven
pul. 'luʔ ʔuw. 'wɔ	twelve
pul. 'luʔ təl. 'lu	thirteen
pul. 'luʔ ʔap. 'pat	fourteen
pul. 'luʔ lim. 'mɔ	fifteen
pus. 'sɔd	navel
put. 'tiʔ	white
put. 'tiŋ ʔɔb. 'buk	gray hair
'pla:ta	silver
'plɛ:tɛk	wings
'pli:ma	phlegm
'pwi:di	can, could (ability)
/b/	
ba. 'pu	boat, larger than a canoe
'ba:bɔŋ	to shut, as a door
'ba:tuk	cough
'ba:diʔ	bolo
'ba:gaʔ	lungs
'ba:guŋ	to throw, toss, as a stone
'ba:ʔu	smell (general)

Bagobo-Klata	English
'ba:ʔu ʔɛ:maʔ	smell (as of underarm, body odor)
ba.'ʔow	north
'ba:həw	thirsty
ba.'li:tiʔ	banyan tree
ba.'li:juŋ	axe
'ba:sa	to read
ba.'sa	pumpkin, squash
ba.'ra:tu	cheap, inexpensive
ba.'ra:suʔ	upper arm
'ba:lu	provisions (as for a journey)
'ba:jad	to pay
'ba:jaw	brother-in-law
ba.'juʔ	cheeks
'ba:wɛʔ	medicine
bɛ.'la:gat	ocean
bɛ pat.'tad	plain
'bɛ:kəŋ	coconut shell
bɛ.'kəs	to shoot, as an arrow from a bow
'bɛ:nɛʔ	to cry
'bɛ:lɛ	to stay, remain
bɛ.'ləd	river
'bi:taʔ	tadpole
'bɔ:tu	testicles
'bɔ:ŋiʔ	rice, burnt lower crust
bɔ.'ŋɔ	deaf
bɔ.'ŋət	pubic hair
bɔ.'ŋu:lɔ	insane, mentally ill
'bɔ:həw	arrow
'bɔ:laʔ	bubbles, foam
'bɔ:lɔg	blind
bɔ.'ləj	house
bɔ.'jɔʔ	face
bɔ.sɔ.'mɔʔ	broken, not working or out of order
'bɔ:wɔʔ	to pour, as liquid
bu.'tuŋ	coconut palm
bu.'na	to hit someone with something
bu.'nut	coconut husk
'bu:ŋɔ	to cook something covered in ash
'bu:la	moon
'bu:lak	flower
'bu:lig	banana
bu.'li:jɔj	drunk, intoxicated
'bu:lu	to cut something (e.g., hair) with scissors
'bu:lud	mountain
bu.'sɔw	stout mythical creatures that are believed to take children
bu.'waʔ	lanzones fruit
ban.'tu:gan	famous
ban.'taʔ	threat
baŋ.'kaʔ	canoe
baŋ.'kil	canine tooth
bab.'ba	broken in pieces
bap.'pa	to carry someone on one's back

	Bagobo-Klata	English
	bag.'gaŋ	jaw, molar tooth
	bal.'las	rice (uncooked)
	bal.'leŋ	to return home
	bal.'lə	embers, glowing coals
	bal.'lu	widow
	baj.'jə	crocodile
	baw.'wəw	high (object)
	bik.'kə	necklace
	bin.'niʔ	rice seed
	bil.'ləd	semen
	bəb.'bə	hole
	bət.ti.'ja	pregnant
	bət.'təm	(foxtail) millet
	bək.ka.'ləs	wound (scratch)
	bək.'kəʔ	not (constituent negation)
	bək.'kəŋ	shin (of one's leg)
	bək.'kəs	fruit bunch (in a basket)
	bəg.gət.'təw	adopted (child)
	bəl.'li	to buy something
	bəl.'lə	poison
	bəl.'ləj	to give something to someone
	bən.'na	true; to believe in someone
	bən.nə.'ʔəs	loud, noisy
	bəs.'sak	earth, soil
	bəl.'la:l.u	earthquake
	bub.'bu	body hair; feather
	bun.'ni	to hide
	buŋ.'ŋə	fruit
	bul.'las	to change something
	bul.'li	night, evening
	buw.'wa	hammock
	'bla:wa	gold
	'ble:ʔe	gecko
	'bli:bud	whirlpool
	'bli:la	rafter
	'blə:waʔ	spider
	'blu:buk	dust
	bja.'ʔəw	drizzle, light rain
	'bjə:ʔə	year
	blab.'baʔ	rice beer
	bluŋ.'ŋus	mouth
	ta	in, on, at (oblique nominal marker)
/t/	ta.'pe	twin
	'ta:piʔ	wall
	'ta:dəʔ	to drip, as from a faucet
	ta.gis.'kwi:la	school
	ta.la.tu.'lu	teacher
	ta.'lum	papaya
	ta.ma.'wət	hook, from which to hang something
	ta.'me:ŋa	stove
	ta.'mi:təj	small bridge
	ta.'mi:hiŋ	ring finger

Bagobo-Klata	English
ta.'mi:.luŋ	shelter
'ta:.lu	to defecate
ta.'rəj	to hang up, as on a hook
'ta:.waŋ	to help
'ta:.wis	trousers (female)
tak.pul.'luʔ	tenth
tak.pul.'luʔ ʔuw.'wə	twelfth
tak.təl.'lu	third
tak.hi.'jəw	ninth
tak.lap.'pat	fourth
tak.li.'mə	fifth
tak.lən.'nəm	sixth
tak.luw.'wə	second
tak.wə.'lu	eighth
tɛ.'kəd	heel
'tɛ:.ʔɛ	hard
'tɛ:.ʔə	Job's tears
'ti:.təj	bridge
'ti:.tuʔ	puppy
'ti:.ŋut	greedy
'ti:.ʔi	little finger
'ti:.ʔəw	clear (as water)
ti.'ja	abdomen, belly
tə.'pə	to broil
tə.pəd.'dəŋ	quiet, silent
tə.bu.'lug	round
'tə:.də	to follow, as a trail
tə.'ʔi	to accompany
tə.'muk	mosquito
'tə:.ŋə	nipple
'tə:.hu	cucumber
tə.'li	string
tə.'ləm	sharp
tə.'wəʔ	fat
'tə:.wuʔ	smart, intelligent
'tu:.baŋ	in front
'tu:.big	water
'tu:.diʔ	vagina
'tu:.duk	oar
'tu:.kuk	lazy
tu.'ʔud	deer
'tu:.ʔuw	to string something together
tu.'la:.gəŋ	spine, backbone
'tu:.liʔ	earwax
tu.'lu	to teach
tu.'rɛʔ.kan	X do not know
'tu:.wə	old (person)
'tu:.wu	to grow
tab.'la	board, wood plank
tam.'ba:.ga	copper
tim.'pu mu.'la	rainy season
tug.'sip	bedbug
tap.'pɛ	old (object)

Bagobo-Klata	English
tab. 'baŋ	bland
tak. 'ka	to arrive
tak. 'kɛʔ	to laugh
tag. 'gɛʔ	urine
tan. 'nɔʔ	snot, nasal mucus
taŋ. 'ŋa	home
tal. 'lɔw	to fear
tɛp. 'pɔʔ	to spill, throw
tɛd. 'dɔʔ	index finger; to point at someone
tib.ba.pul. 'luʔ	only ten
tib.ba.təl. 'lu	only three
tib.ba.ʔap. 'pat	only four
tib.ba.ʔuw. 'wɔ	only two
tib.ba.ho. 'tu	only one
tib.ba.li. 'mɔ	only five
tid. 'dɔ	to stand
tid. 'dɔs	to crush lice with one's fingernail
tik. 'ka	dry; to dry in the sun
tig. 'gaŋ	to cook rice
tig. 'guʔ	to hit the mark
tim. 'mɔs	to clean something
tɔp. 'pɔʔ	to answer, reply
tɔp. 'pɔ:.kaŋ neŋ ʔul. 'lu	headhunting
tɔb. 'bi	to sew
tɔb. 'biŋ	buttocks
tɔb. 'bɔk	to stab
tɔb. 'bɔʔ	coconut fruit (young)
tɔb. 'bu	sugarcane
tɔb. 'bus	to redeem
tɔt.tɔt. 'tɔw	self
tɔt. 'tɔk	house lizard
tɔt. 'tɔ:.lu	three each
tɔk. 'kɔ:.li	language
tɔk. 'kub	to cover, put a cover on something
tɔg.gaŋ. 'gaʔ	parent
tɔd. 'dɔj	fruit bunch (still on a tree)
tɔd. 'du:.wa	ancestors
tɔm.ma. 'na	want (modal)
tɔn. 'nɔb	honey
tɔŋ. 'ŋɔ	half
tɔŋ. 'ŋɔŋ bul. 'li	midnight
tɔs. 'sɔ:.tu	one each
təl. 'lu	three
təl. 'lu pul. 'luʔ	thirty
təl. 'lu:.wɔ	two each
tuk. 'kɔd	cane, walking stick
tud. 'duʔ	nape
tug. 'gɔ	post, house pole
tug. 'guŋ	to increase
tul. 'la	bone
tul. 'lid	straight
tuw. 'waŋ	to drop, fall down

	Bagobo-Klata	English
	da. 'gəw	short (of time)
/d/	'da: 'ʔit	name
	'da: 'ʔu	leaf
	'da: 'mu	dew
	'da: 'la	path, trail
	da. 'luŋ	low
	dajt	what
	da. 'jə	west
	'dɛ: 'laʔ	tongue
	'dɛ: 'ja	to turn, as around a corner
	'di: 'pa 'la: 'gat	island
	'di: 'ta	name
	'dɔ: 'ʔɔs	lime for betel nut chewing
	'dɔ: 'mɔ	to remember
	'dɔ: 'nɔw	to wake someone up
	dɔ. 'lid	root
	du. 'li: 'ja	durian (<i>Durio zibethinus</i>)
	'dɔ: 'lɔŋ	to honor, praise
	'du: 'pu	to chase away
	'du: 'luʔ	blood
	dat. 'tɔj	to carry on one's shoulder
	daŋ. 'ŋɔw	handspan
	dip. 'pɔ	armspan
	dit. 'tu	to come near the speaker
	dik. 'kɔt	to stick, adhere to
	dim. 'mɔt	rice bran
	dis. 'sɔʔ	sty (in one's eye)
	dɔd. 'dɔŋ	near; to approach someone
	dɔn. 'nak	broom
	dal. 'lɔm	inside; content
	dɔl. 'lɔm	heavy
	dug. 'gɔj	long time; slow
	dum. 'mɔ	other
dun. 'nuk	flood	
	'ka: 'peŋ	split bamboo
/k/	'ka: 'peʔ	to hold, grasp
	'ka: 'pɔ	cold season, weather
	'ka: 'bɔg	fruit bat
	'ka: 'tig	outrigger
	'ka: 'da	each
	ka. 'ʔa	to eat
	ka. 'ʔi. 'la: 'ŋa	need, must
	ka. 'ʔɔb	to lie face down
	ka. 'si: 'du	knife (for all genders)
	ka. 'hi. 'ra: 'pa	difficult
	ka. 'laʔ	frying pan
	ka. 'li. 'baŋ. 'baŋ	butterfly
	ka. 'jaʔ	to lie on one's back
	ka. 'jang	bright
	'ka: 'wad	barb (as a fish hook)
	'kɛ: 'led	to lie on side
	'kɛ: 'wɔʔ	short (of a person)

Bagobo-Klata	English
'ki:.bət	civet
ki.'tik	small (as a young person)
'ki:.lat	lightning
ki.'lid bɛ.'lɛd	riverbank
ki.'lid 'la:.gat	seashore
ki.'rəp	to blink one's eyes
kə.tə.'lɛ:.bɛs	diarrhea
kə.'mət	blanket
'kə:.nə	rice (cooked)
'kə:.hɛt	cave bat
kə.'lat	thin (of a person)
'kə:.lɛs	to write
kə.luk.'ku?	dirty
'kə:.li	to say
'kə:.wəj	snail
'ku:.pa	thick
'ku:.da?	horse
'ku:.kaŋ	cockroach
ku.'li:.səp	food particles stuck between teeth
'ku:.lat	mushroom
'ku:.lit	skin
'ku:lit 'hu:.ʔuŋ	lip
'ku:.luŋ	to snore
ku.'luŋ	back
kam.'biŋ	goat
kas.'pa	dandruff
kum.'biŋ	Jew's harp
kap.'pɛŋ	to split into halves
kap.'pi	Philippine eagle
kad.'di:.ru	cauldron
kak.'kas	to scratch up the soil, as a chicken
kal.'laŋ	scab
kaj.'ju	tree
kək.'kɛs	to scrub the floor
kɛj.'jɛ	absolute medial demonstrative (free)
kit.'tə	to see; to show
kid.'dʊt	astonishment
kik.'kis	to scrape, as a coconut
kin.'na	have (existential verb)
kin.'na 'bə:.gək	sick
kil.'li	eel
kəp.'pə	chest
kət.tol.'lu	three times
kət.'tu?	absolute distal demonstrative (free)
kəd.'da:.la	knowledge
kəd.'dʊt	to pinch someone
kək.'kəp	to embrace
kək.'kət	to dig
kəl.'la:.pat	four times
kəl.'li:.mə	five times
kəl.'lət	curly (hair)
kəl.'ləs	strong
kəl.'lu:.wə	twice

	Bagobo-Klata	English
	kəm.'mi	beard
	kəm.'mən	fist
	kəm.'məs	to wring
	kən.'ni	absolute proximal demonstrative (free)
	kən.'nə	dream
	kən.nəd.'dəʔ ŋə 'la:ʔi	son-in-law
	kən.nəd.'də ŋə 'li:bu	daughter-in-law
	kən.'ŋə	later, in a while
	kə.ŋəd.'dəw	yesterday
	kəs.'sə	once
	kuk.'kus	to scrub one's body
	kut.'tu	head louse
	kul.'lə	pot for cooking
	kla.'mat mə.'tə	eyelashes
	'kle:weŋ	eyebrows
	'kli:tut	anus
	'kli:giʔ	hawk
	klə.'ni	now
	'klu:gəʔ	wart
	'kjə:hən	crab
	kwa.'lə	earthworm
	'kwa:ja	bamboo
	klak.'kam	ringworm, scabies
	klat.'taŋ	a step in a ladder
	klut.'tuŋ	forehead
	klam.'mag	star
	kləb.'bəw	water buffalo
	kləm.'məʔ	tomorrow
	kjuw.'wa	bee
	ga.'bas	saw (tool)
/g/	ga.'taʔ	coconut milk
	'ga:həw	ant
	ga.'huʔ	easy
	'ga:liʔ	to do, make
	'ga:nas	door
	'gɛ:nət	slow
	'gɛ:hət	domestic pig
	gɛ.'ləʔ	correct
	'gɛ:lək	tickle
	'gi:bəʔ	south
	'gi:bud	fontanelle, hair whorl
	gi.bul.'li	last night
	gi.'ra	last
	gi.ra.paw.'wə	younger sibling
	gi.'ra nəŋ baŋ.'kaʔ	stern, back of boat
	'gə:bə	to lie
	gə.'təŋ	eggplant
	'gə:laʔ	to sell
	'gu:hiŋ	water jar
	gat.'tə	itch(y)
	gak.'ka	to deceive someone
	gaj.'jə ŋə ʔən.'nus	wind from the south

	Bagobo-Klata	English
	gam.'mi	to start
	gen.'nɔ	earlier, recently
	gɔg.'gɔt	a sharp weapon
	gɔl.'lɔt	middle finger
	'gka:.waŋ	hip
	'gna:.lɔ	throat
	'gla:.ʔu	Adam's apple
	gla.'ʔud	east
	'glɔ:.puj	bee
	'gja:.wat	pimple
	gwal.'li	across, opposite
	'ʔa:.paŋ	locust
/ʔ/	ʔa.pi.'li:.du	surname
	'ʔa:.tad	raft
	'ʔa:.taʔ	rice husk
	'ʔa:.tiŋ	sweat
	ʔa.'tiŋ	if (conditional)
	'ʔa:.tu	to revenge
	ʔa.'kap	monkey
	ʔa.'gad kɛ.'la ma	never
	'ʔa:.hɔs	garlic
	ʔa.'ʔɔg	to take a bath
	ʔa.'ʔɔw	to enter
	ʔa.'ri:.jus	earring
	ʔaw.'wɔ:	to weave
	'ʔa:.wa	rainbow
	'ʔa:.wak	waist
	'ʔɛ:.pa	to cross a river, road, etc.
	'ʔɛ:.pɔk	to cut off, sever, as a piece of rope
	'ʔɛ:.baŋ	left
	'ʔɛ:.taʔ	calf, of one's leg
	ʔɛ.'tɔd	to escort, bring someone somewhere
	'ʔɛ:.kɔŋ	tail
	'ʔɛ:.gɔŋ	feces
	'ʔɛ:.ʔɛ	yes
	'ʔɛ:.ʔɔd	to sit
	'ʔɛ:.maʔ	armpit
	ʔɛ.'nɔʔ	turbid, murky; mud
	'ʔɛ:.lam	wasp
	'ʔɛ:.jap	to count
	ʔi.'pat	to take care of someone/something
	'ʔi:.pag	sister-in-law
	'ʔi:.pit	to sleep
	'ʔi:.gɔʔ	betel leaf
	'ʔi:.nit	hot (as weather)
	ʔi.'nɔ	mother
	ʔi.'nɔ gwal.'li	aunt
	ʔi.'num	to drink
	'ʔi:.ŋa	envious
	'ʔi:.hi	to whet a blade
	'ʔi:.hip	to whisper
	'ʔi:.hiŋ	ring

Bagobo-Klata	English
ʔi.'laʔ	to lie down
ʔi.'las	to slice (e.g., meat)
ʔi:.'lis	to move to another location
ʔi:.'lus	to massage
ʔi:.'jaʔ	do not (negative imperative)
'ʔi:.'jup	to blow, using one's mouth
ʔi:.'jug	coconut fruit (mature)
'ʔi:.'wə	saliva
ʔə	or
'ʔə:.'puj	flame
ʔə.'bəw	mouse, rat
'ʔə:.'taʔ	to vomit
ʔə.'təp	roof, thatch
ʔə.'təj	liver
ʔə.'ka	what-you-may-call-it
ʔə.'kap	tree bark
'ʔə:.'meŋ	to cook
ʔə.'mə	father
ʔə.'mə gwal.'li	uncle
'ʔə:.'hap	to chew (sugarcane), bite off
ʔə.'hi	salt
'ʔə:.'hət	rice straw
'ʔə:.'lat	scar
ʔə.'lat	vein, root
ʔə.'ləʔ	and
ʔə.'li	to choose, select
ʔə.'lək	to kiss
ʔə.'ləw	fence
ʔə.'luŋ	to speak to someone
'ʔə:.'jam	yawn; to yawn
ʔə.'wak	crow
ʔə.'wu	ash
ʔu.'pəw	bald
'ʔu:.'pus	cat
'ʔu:.'biʔ	to ask for, request
'ʔu:.'ti	penis
'ʔu:.'gaŋ	co-parent
'ʔu:.'gaŋ 'la:.'ʔi	father-in-law
'ʔu:.'gaŋ 'li:.'bu	mother-in-law
'ʔu:.'hə	wild boar
'ʔu:.'la	snake
ʔu.'la	rain
ʔu.'ləb	knee; to kneel
'ʔu:.'lit	to repeat; to do something again
'ʔu:.'liŋ	charcoal
'ʔu:.'ŋa	to call
ʔu.'wəj	rattan
ʔan.'tap	to think
ʔan.'daʔ	not have (negative existential)
ʔan.'daʔ nən.'nəŋ	stupid
ʔim.'pə	thing
ʔin.'diʔ	not (standard negation)
ʔin.'diʔ tə.'ləm	dull, blunt

Bagobo-Klata	English
ʔin. 'diʔ ku. 'waŋ	hoarse (of one's voice)
ʔom. 'bɔʔ	do not like (negative desiderative)
ʔap. 'pat	four
ʔap. 'put	to reach something
ʔap. 'puʔ ʔu. 'lɔb	great-grandparent
ʔap. 'puʔ 'la. ʔi	grandfather
ʔap. 'puʔ 'li. bu	grandmother
ʔas. 'su	dog
ʔat. 'tap	winnow, sift rice
ʔad. 'da	ladder
ʔak. 'kat	to climb, as a tree
ʔak. 'kaŋ	owl
ʔal. 'lɔj	chin
ʔam. 'mi	right
ʔam. 'muʔ	breast; to suck, breastfeed
ʔan. 'nɔj	termite
ʔaŋ. 'ŋat	to wait for somebody
ʔaŋ. 'ŋaʔ	baby
ʔaŋ. 'ŋaʔ gwal. 'li	nephew, niece
ʔaŋ. 'ŋet	to bite
ʔaŋ. 'ŋɔj	to get something
ʔal. 'lɔs	to flow
ʔas. 'sɔ	fish poison
ʔel. 'lɔs	to masturbate
ʔej. 'je	absolute medial demonstrative (bound)
ʔip. pɔ	tooth
ʔip. 'pɔd	spouse
ʔip. 'pɔd ŋɔ 'la. ʔi	husband
ʔip. 'pus	gasp, paint
ʔit. 'tɔm	black
ʔik. 'kam	mat, for sleeping
ʔik. 'kɔt	rope; to tether an animal
ʔid. 'di	to boil, as water
ʔid. 'duŋ	nose
ʔiŋ. 'ŋɔ	deaf
ʔil. 'lɔʔ	mole, birthmark
ʔij. 'jak	to shout loudly
ʔɔp. 'pɔʔ	grasshopper
ʔɔp. 'pɔs	to borrow
ʔɔp. 'pu	thumb; grandchild
ʔɔp. 'puŋ	to smell something
ʔɔp. 'pu ʔu. 'lɔb	great-grandchild
ʔɔb. 'bid	roe, fish eggs
ʔɔb. 'bɔ	smoke, as from fire
ʔɔb. 'buk	hair
ʔɔt. 'tad	to separate
ʔɔt. 'tɔk	brain
ʔɔt. 'tɔw	person
ʔɔt. 'tu	to harvest
ʔɔt. 'tut	flatulence
ʔɔt. 'tuʔ	absolute distal demonstrative (bound)
ʔɔd. 'dɛ	unripe (of a fruit)
ʔɔd. 'dɔk	to pound, as rice

Bagobo-Klata	English
ʔəd. 'dow	day, sun
ʔəg. 'gət	clothes
ʔəg. 'giʔ	cogon grass
ʔəm. 'mət	to catch
ʔəm. 'məj	rice plant, unhusked rice
ʔən. 'nəm	six
ʔən. 'nəm pul. 'luʔ	sixty
ʔəl. 'lik	to thresh
ʔəl. 'ləb	spring (of water)
ʔəl. 'lət	between
ʔəl. 'ləŋ	neck
ʔəl. 'lu	pestle
ʔən. 'ni	absolute proximal demonstrative (bound)
ʔən. 'nəp	fog, mist
ʔən. 'nus	wind
ʔən. 'nus gid. 'du ta 'la: .gat	monsoon wind
ʔup. 'pat	to jump over, leap
ʔup. 'pə	hen
ʔub. 'bi	spiderweb, cobweb
ʔuk. 'kat	to strike with intention to hurt
ʔun. 'nə	first
ʔun.nə.paw. 'wə	older sibling
ʔəs. 'səm	sour; pomelo fruit
ʔul. 'ləd	larva
ʔul. 'lu	head
ʔul. 'lud	floor
ʔus. 'saʔ	to put, place
ʔus. 'siŋ	soot
ʔuw. 'wi	(purple) yam
ʔuw. 'wə	two
ʔuw. 'wə pul. 'luʔ	twenty
ʔuw. 'wə ʔul. 'li klə. 'ni	day after tomorrow
ʔuw. 'wə lə. 'tus	two hundred
/m/	
'ma: .ʔa	sheath
ma. 'ha	expensive
'ma: .mə	leg
'ma: .si	salty
'ma: .las	spicy
'ma: .liʔ	good
ma.li. 'ba: .ʔu	fragrant
ma.li.gən. 'nəj	beautiful, pretty, attractive
ma.li.na. 'nam	beautiful
ma. 'li: .nis	clean
ma. 'niʔ	peanut
'ma: .raŋ	marang fruit
'ma: .jəw	raincloud
'mə: .ŋa	fireplace
'mə: .gat	mote, dirt in one's eye
mə. 'pəw	shallow
mə. 'tə	eye
mə. 'haŋ	betel chew, quid
mə. 'ʔə	betel, areca nut

	Bagobo-Klata	English
	mə.'nuk	chicken
	mə.'lə	light (in weight)
	mə.'jə	young woman, lass
	mu.'ʔud	spirit, soul
	mu.'ʔud nəŋ nə.'təj	spirit, of dead person
	'mu:.wu kəŋ 'la:.gat	high tide
	maŋ.'ga	mango
	maŋ.'pəŋ ʔəd.'dəw	sunrise
	mab.'bag	swollen, swelling
	mat.'tək	leech
	məs.'səŋ	industrious, hardworking
	məb.'but	animal
	məm.'mis	sweet
	mən.'ni	smell (as of fish)
	məs.'səd	smell (as of urine)
	məs.'sək ʔəd.'dəw	sunset
	məl.'ləŋ	dry season
	mud.'du nəŋ 'pa:.ʔə	toe finger
	mum.'mə	crumbs (i.e., rice or food)
	mul.'lə	to plant
	'mlɛ:.ʔaŋ	rough, coarse
	mlə.'tus	one hundred
	'mlə:.ʔəs	smooth
	'na:.ta	vegetable
/n/	'na:.də	naked, bare
	na.'gəʔ	mine
	'na:.ʔɛ	intestine
	'na:.ʔəw	to steal
	'na:.ʔu	to descend a hill
	'na:.na	to leave behind, abandon
	na.'nap	to crawl
	na.'nam	taste
	nɛ.'pɛs	thin (of an object)
	'ni:.paʔ	nipa palm
	ni:.dəm	to think about something carefully
	nə.'təj	corpse
	nə.'ʔəs	noise
	'nə:.nə	lobster, prawn
	'nu:.du	why
	nu.'ma	why
	nu.'lit ŋə ʔəd.'dəw	day before yesterday
	nu.'wə	to hear, listen
	naw.'wə	to love someone; to breathe
	nɛj.'jɛ	genitive distal demonstrative
	nəb.'bə	field
	nəd.'dəʔ	to live at, reside at
	nəm.'mə	to search for
	nuw.'was	accustomed
	'ŋu:.ʔəb	finger nail
/ŋ/	'ŋu:.ʔuj	whistle
	'ŋu:.ma	story

	Bagobo-Klata	English
	ŋit. 'təŋ	dark
	ŋil. 'lə	ear
/s/	sa.lu. 'wa	trousers (male)
	si. 'laʔ	sweet potato
	si. 'laʔ kaj. 'ju	cassava
	'sə: ʔa	ankle
	sa.lup. 'paŋ	loincloth, g-string
	saj. 'jəw	to dance
	səs. 'səp	to suck, sip
/h/	'ha: puʔ	meat
	ha. 'but	sieve, strainer
	'ha: kəj	to ride
	ha. 'gəʔ	I
	'ha: lud	to catch water with a pail
	ha. 'jup	wrong, incorrect
	ha. 'wet	to hang something up, not on a hook
	'hə: lə	to lean
	'hi: paʔ	to kick
	hi. 'kat	fast, quick
	hi. 'tə	we (inclusive)
	hi. 'la	they; absolutive plural nominal marker
	'hi: mat	needle
	hi. 'ja	s/he
	hi. 'jəw	nine
	hi. 'ju	elbow
	hə	absolutive singular nominal marker
	hə. 'pə: təj	enemy
	hə. 'bəw	(breast)milk
	hə.tə. 'ʔi	companion
	hə. 'tu	one
	hə. 'tuŋ 'mla: la	one thousand
	'hə: ʔa	thorn
	hə.məʔ. 'ba: ʔu	bad smell, stink
	hə.mə.bə. 'jəʔ	ugly
	hə. 'məʔ	bad, evil
	hə. 'mu	you (plural)
	'hə: lə	egg
	hə. 'rəʔ	to stop
	'hu: kiʔ	bamboo water container
	'hu: kə	anger, angry
	'hu: duʔ	to carry on one's head
	'hu: ŋəj	horn (of an animal)
	'hu: lat	to write something
	'hu: lud	to comb for lice with a fine tooth comb
	hu. 'wat	to comb one's hair
	hu. 'wiŋ	beak
	hap. 'pəj	to hang clothes on a clothesline
	had. 'də	to know
	ham. 'mə	we (exclusive)
	hid. 'duʔ	to spit
	hik. 'kə	you (singular)

	Bagobo-Klata	English
	hiŋ. 'ŋɔ	to blow one's nose
	his. 'sip	louse (of a chicken)
	his. 'sib	fat, grease
	hɔb. 'bɛʔ	dry stream
	hɔb. 'bɔt	grass
	hɔŋ. 'ŋiʔ	knife (for all purposes)
	hɔŋ. 'ŋɔw	steam, vapor
	hud. 'duʔ	young man, lad
	hud.du. 'ʔa	hiccough
	hup. 'pu neŋ 'bu:lud	summit
	hal. 'la	to fry rice
	hug. 'gu	to push
	hul. 'lu	to command, send on an errand
	'la:.baʔ	brave, fearless
/l/	'la:.bid	ringworm
	'la:.ta	taro
	la. 'tiʔ	swamp, wetland
	'la:.gat	sea
	'la:.ʔi	man, male
	la. 'ʔɔt	to carry with both hands, as a heavy load
	'la:.ʔu	to bail water, as from a canoe
	'la:.nu	sad, unhappy
	'la:.ŋit	sky
	'la:.ŋub	cave
	la. 'hat	all
	'la:.lak	friend
	la. 'lɛ	to run
	'la:.lig	happy, joyful
	'la:.jaŋ	to fly
	'la:.was	body
	'la:.was neŋ kaj. 'ju	tree trunk
	la. 'wɛ	high (of a person)
	'la:.wu	cloud
	la. 'wu	to fall (from a higher place)
	lɛ. 'pɛs	knife (for women)
	'lɛ:.pɔs	sibling
	'lɛ:.pɔs gwal. 'li	cousin
	'lɛ:.dɛ	yellow
	'lɛ:.dɔk	waterfall
	'lɛ:.haʔ	torch
	'lɛ:.wɔʔ	narrow
	'li:.bu	woman, female
	'li:.tu	fire, to burn
	'li:.hɔʔ	louse egg
	'li:.liŋ	to look back
	li. 'mɔ	five
	li. 'mɔ pul. 'luʔ	fifty
	'li:.mut	to forget
	'li:.nɔw	lake
	'li:.juŋ	to turn around
	lɔ. 'tus	hundred
	lɔ:.kat	to open up one's eyes

Bagobo-Klata	English
'lɔ:.kuʔ	shirt
lɔ.'ʔɔw	to walk
'lɔ:.maʔ	knife (of a datu)
lɔ.'met	to burp, belch
'lɔ:.nɔd	to come back
lɔ.'ɲuj	to swim
'lɔ:.la	nest
'lɔ:.jɔb	weak
'lɔ:.jɔd	to pull, drag
lɔ.'wa	wide
lu.'tɔw	to float
'lu:.tuʔ	ripe
'lu:.ka	bowl
lu.'ga	place
'lu:.huk	ribs
'lu:.mut	moss, algae
'lu:.na	placenta (afterbirth), pillow
'lu:.ɲa	shade
lu.'lu	to broil something over coals
'lu:.jɔ	ginger
'lu:.wa	ladle, as of coconut shell
'lu:.ɲu	coffin
lug.'waʔ	to go out, exit
lab.'bus	poor
lak.'kaʔ	jackfruit
lag.'gam	bird
lag.'gas	to wash one's face or hands or dishes
lam.'ma	fever
lam.'mas	to sow seeds
lam.'mi	new
lan.'na	pus
lan.'nɔ	cooking oil
laŋ.'ɲɔw	housefly
lal.'lad	to move (be in motion, not be still)
lal.'lɔm	deep
laj.'ju	far
law.'wɔw	tear(drop)s
let.'taʔ	sap, resin
les.'set	to tear, rip
les.'set kɛŋ 'hu:.ʔuŋ	harelip
les.'seŋ	to annoy someone
lip.'pɔ	mold, mildew
lid.'dɔ	slippery
lik.'kut	tight
lit.'ti	thunder
lig.'gaʔ	red
lim.'mɔ	hand
liŋ.'ɲu	shadow
lis.'sɔ	sesame
lɔp.'puk	all gone, consumed
lɔb.'bɔ	to collapse, cave-in
lɔb.'bɔŋ	to bury, inter
lɔt.'tu kɔd.'dɔw	noon, midday

	Bagobo-Klata	English
	lɔd. 'dɔŋ	rotten
	lɔk. 'kɔ	granary
	lɔk. 'kɔd	to blame; to regret, be sorry
	.lɔm. 'mɔd	to swallow down
	lɔm. 'mɔʔ	morning
	lɔn. 'nɔt	mortar, for pounding grain
	lɔŋ. 'ŋaʔ	to flee, run away
	los. 'suŋ	to ask a question
	lud. 'dɔ	cotton
	lud. 'duŋ	to lose one's way
	lud. 'dus	fish
	lun. 'nɔd	to drown
	lun. 'nɔw	green
	lul. 'lut	to cook something in bamboo
	lus. 'suk	bead (as for jewelry)
	lus. 'suŋ	green
	luw. 'wu	winnowing basket
	lug. 'waʔ	outside
	'ju:.pa	centipede
/j/	jab.baʔ. 'na:.pu	nightmare
	jɔj. 'jɔ	shame
	'wa:.lɔj	tired, weary
/w/	wi. 'ti	satiated
	'wɔ:.ʔɔ	abaca fibers, hemp
	wɔ. 'lu	eight
	wal. 'lɔ	shoulder
	wɔd. 'dɔw	afternoon
	wɔs. 'su	hungry

PSEUDOREGISTER IN PACOH: PRELIMINARY ACOUSTIC ANALYSIS AND IMPLICATIONS FOR A GENERAL MODEL OF PSEUDOREGISTER FORMATION IN AUSTROASIATIC

Ryan GEHRMANN
University of Edinburgh
r.m.gehrmann@sms.ed.ac.uk

Abstract

Register is a suprasegmental contrast cued by a combination of voice quality, vowel quality and pitch found in many languages of Southeast Asia. Conventionally, registers develop from the transphonologization of historical onset voicing contrast; however, there are several Austroasiatic languages which have developed apparent register contrasts that correspond to historical vowel quality contrasts. The term *pseudoregister* is introduced here to refer to such contrasts. A general model for the formation and evolution of pseudoregister is proposed and, as quantitative phonetic analysis of pseudoregister languages is sorely lacking, the results of a preliminary study on the acoustic correlates of pseudoregister in Pacoh (< Katuic) are presented. F1 is shown to be the only acoustic correlate which reliably covaries with pseudoregister in Pacoh, while subtle differences in F0 and spectral tilt (H1*-H2*) are also detectable.

Keywords: historical phonology, phonetics, register, Austroasiatic, Katuic
ISO 639-3 codes: pac, tto, tth, irr, oog, sed, cog

1 Introduction

The Pacoh language of Vietnam and Laos has been described as a *register language* (Watson 1964, 1966, 1996; Watson et al. 1979, 2013; Diffloth 1982; Sidwell 2005; Alves 2006; Gehrman 2015). Register is a binary suprasegmental contrast, upheld by a combination of voice quality, vowel quality and pitch cues. Register, in its conventional form, develops under conditioning from the historical voicing specification of onset consonants and emerges in conjunction with the loss of said voicing contrasts (i.e., *registrogenesis*) (Huffman 1976). However, the distribution of register in Pacoh is unrelated to historical onset voicing, corresponding instead with historical vowel quality contrasts (Diffloth 1982, Sidwell 2005, Gehrman 2015). As such, Pacoh, a language from the Katuic branch of Austroasiatic, is part of a small class of languages which I propose be called *pseudoregister* languages. Other documented pseudoregister languages include languages of the Bahnaric and Pearic branches of Austroasiatic and one sister language of Pacoh in the Katuic branch: Ta'oi (Sidwell 2015, 2019; Gehrman 2015, 2019).

Unfortunately, apart from the well-documented Chong language (< Pearic) (L-Thongkum 1991, Edmondson 1996, DiCanio 2009), acoustic phonetic descriptions of pseudoregister are sorely lacking. Pacoh phonology has been described qualitatively by Watson (1996), but there has been no acoustic phonetic investigation of the phenomenon in this language. In this paper, I will demonstrate using a small set of archived Pacoh word list recordings that differences of vowel height (F1) are by far the most reliable acoustic correlates of the Pacoh pseudoregister contrast. Nevertheless, in addition to F1, there are secondary correlates, including a relatively lower fundamental frequency (F0) and spectral tilt (H1*-H2*) in the high pseudoregister.

This investigation into the acoustic correlates of the pseudoregister contrast in Pacoh begins the process of establishing a quantitative baseline description of the Pacoh pseudoregister phenomenon. Furthermore, the results of this preliminary study have important implications for a general model of pseudoregister formation. Two primary typologies of pseudoregister are apparent in the pseudoregister languages described thus far, but Pacoh pseudoregister does not match either of these prototypes. A

general model of pseudoregister formation is proposed here, based on the patterns of pseudoregister emergence and evolution among all of the pseudoregister languages documented thus far.

2 Pacoh

Pacoh is spoken primarily in the Thừa Thiên Huế province of Vietnam and adjacent areas of Salavan province, Laos. Speakers of Pacoh are included under the officially recognized *Tà Ôi* ethnic group in Vietnam and, indeed, some speakers of Pacoh refer to their language as *Ta’oi*.¹ The language is not particularly well documented apart from the dialect which is the focus of this study, *High Pacoh* (hereafter, simply Pacoh). Phonological descriptions (Watson 1964, 1966, 1996), grammatical descriptions (Alves 2006, 2007, 2015; Watson 1977, 2011; S. Watson 1964, 1966, 1976; Watson et al. 2013) and a sizable dictionary (Watson et al. 2013) are available for the High Pacoh dialect. The phonology of the Cado dialect has been described (Gehrman 2015, Vitrano-Wilson et al. 2018) and brief descriptions of two varieties referred to as Pakôh and Taôih are also available (Nguyen et al. 1986). Another variety, Bahi, is mentioned in passing by Watson (1996).

Watson’s characterization of pseudoregister in Pacoh evolves over three decades from the 1960s to the 1990s until, in a final paper on the topic, Watson (1996) presents his conclusions on the matter. Having been influenced by Gregerson’s (1976, 1984) hypothesis that differential tongue root position underlies phonological register in Southeast Asia and by his own experience working with African ATR harmony languages, Watson settled on a bifurcation of Pacoh vocalism into two registers: a *tongue root retracted* (+RTR) register and a *tongue root advanced* (-RTR) register. The +RTR register is reportedly the more marked of the two and is described as *tense* or *pharyngealized*. Watson draws parallels between Pacoh register and register in Sedang, both of which are described as having the “...‘creaky, raspy’ vowels of a retracted tongue root articulation.” (Watson 1996, 200). Hereafter, we will refer to the tenser, +RTR register as high pseudoregister and the laxer, -RTR register as low pseudoregister.

Watson’s ultimate interpretation of Pacoh vocalism divided the language’s 30 vowel phonemes into fifteen vowel qualities doubled for register contrast. This inventory is presented in Table 1, with slight alterations in the spelling of phonemes compared to Watson’s (1996) inventory.

Table 1: *Pacoh vocalism*

/	ia ^H	iə ^L	ia ^H	iə ^L	ua ^H	uə ^L									
	i: ^H	i: ^L	i: ^H	i: ^L	u: ^H	u: ^L	i ^H	i ^L	i ^H	i ^L	u ^H	u ^L			
	e: ^H	e: ^L	a: ^H	a: ^L	o: ^H	o: ^L	e ^H	e ^L	a ^H	a ^L	o ^H	o ^L	/		
[ia	iə	ia	iə	ua	uə									
	e:	i:	ə:	i:	ɔ:	u:	e	i	ə	i	ɔ	u			
	æ:	ɛ:	a:	ɜ:	ɒ:	o:	æ	ɛ	a	ɜ	ɒ	o]		

3 Experiment

The phonetic correlates of the proposed Pacoh pseudoregister contrast are currently unknown, apart from Watson’s qualitative description. Here, I present an introductory acoustic study on Pacoh pseudoregister, the goal of which is to confirm on empirical grounds, if possible, the pseudoregistrical interpretation of Pacoh vocalism. The data was segmented and annotated in *Praat* (Boersma & Weenink 2020) and acoustic measures were extracted using the *PraatSauce* script (Kirby 2020). The resulting data was analyzed in python and graphical representations of the data were produced using the *plotnine* package (Kibirige et al. 2021).² Unfortunately, only a small amount of data was available for this study, which amounts to 303 words spoken one time each in isolation by a single male speaker. The recordings were made on tape by Watson, working with a native speaker in 1972.³ Not every vowel in the Pacoh

¹ Not to be confused with the Ta’oi language [tto, tth, irr, oog] spoken in Laos, which is a separate Katuic language (Watson 1996).

² *Plotnine* is based on the *ggplot2* package for R (Wickham 2016).

³ I wish to express my gratitude to Dick Watson for sharing these audio files with me for this project.

vocalic inventory is represented in this data set and the general paucity of data precludes the possibility of addressing this issue definitively here and now.

3.1 Vowel Quality

Watson (1996:200) explains that that high pseudoregister vowels are phonetically more open than corresponding low pseudoregister vowels, and that the degree of vowel height difference “increases from front to back”. The close back high pseudoregister vowel /u(:)^H/ is so much more open than its low pseudoregister counterpart, in fact, that it is found at a more open vowel height position [ɔ(:)] than the low register non-close vowel /o(:)^L/ [o(:)]. This is asymmetrical with respect to the corresponding non-back vowels (cf. **Error! Reference source not found.**), but the explanation for this lies in the relatively recent retraction of a historical mid vowel, *ə(:), which has disrupted the vowel height – pseudoregister relationship in the Pacoh back vowel inventory (Gehrman 2015, 2019).

Figure 1 presents a graphical representation of mean F1 values for each token across the middle 50% of vowel duration (i.e., $0.25 < t < 0.75$ with respect to regularized time). Low pseudoregister vowels are indeed consistently closer in vowel quality than their high pseudoregister counterparts and Watson’s observation that /u(:)^H/ is unexpectedly more open in vowel quality than /o(:)^L/ is confirmed. This suggests that all of the pseudoregister vowel pairs would be differentiable one from the other based solely on vowel quality measures, even in the absence of additional, co-varying register cues (see F1-F2 charts in Figure 2).

Figure 1: F1 measurements for vowels

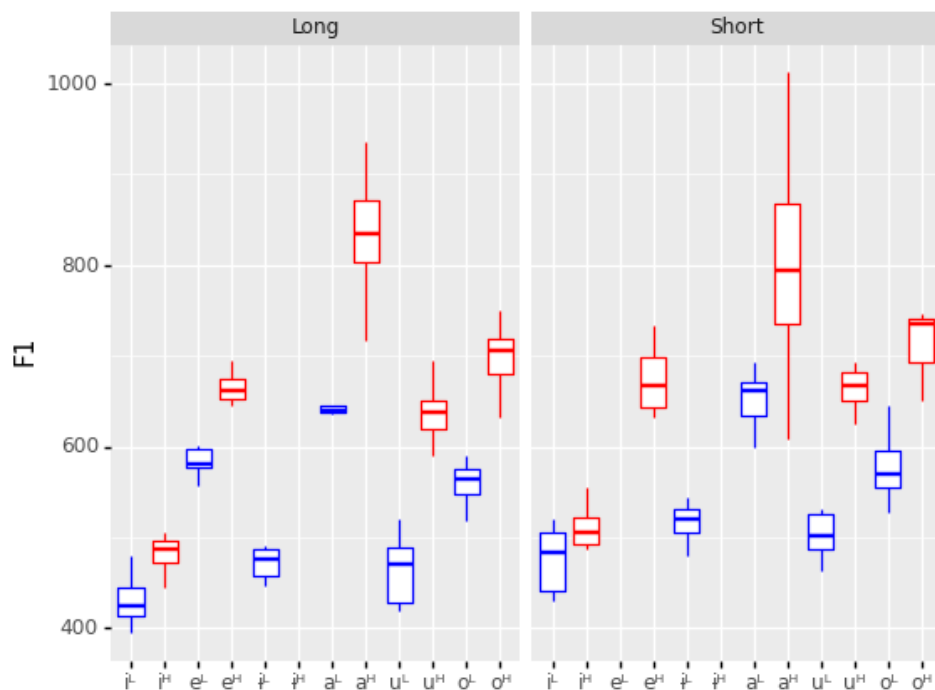
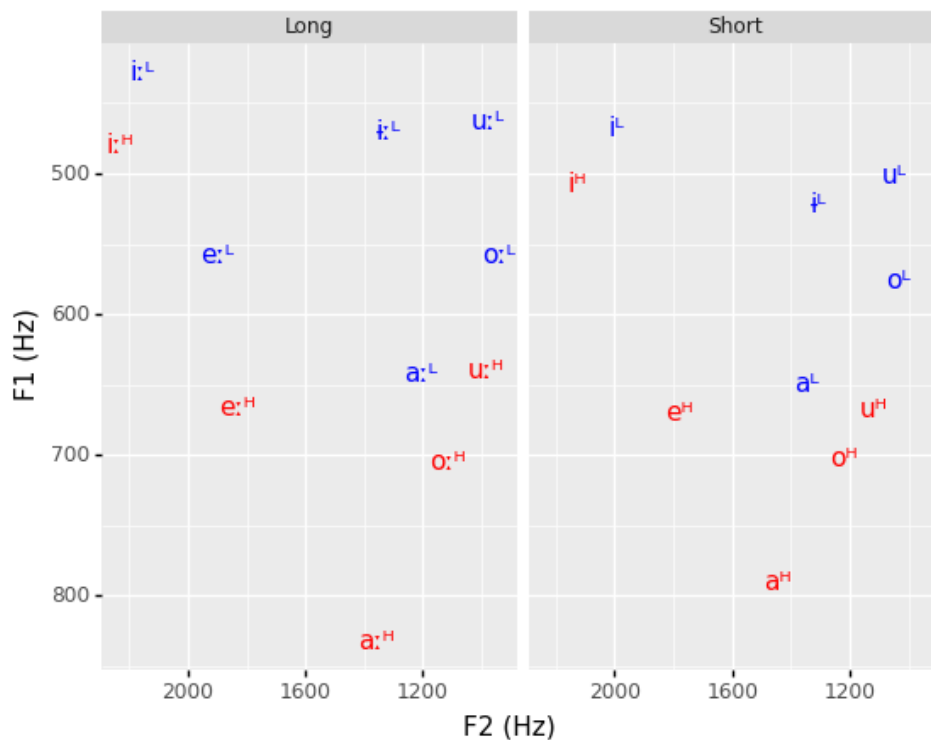
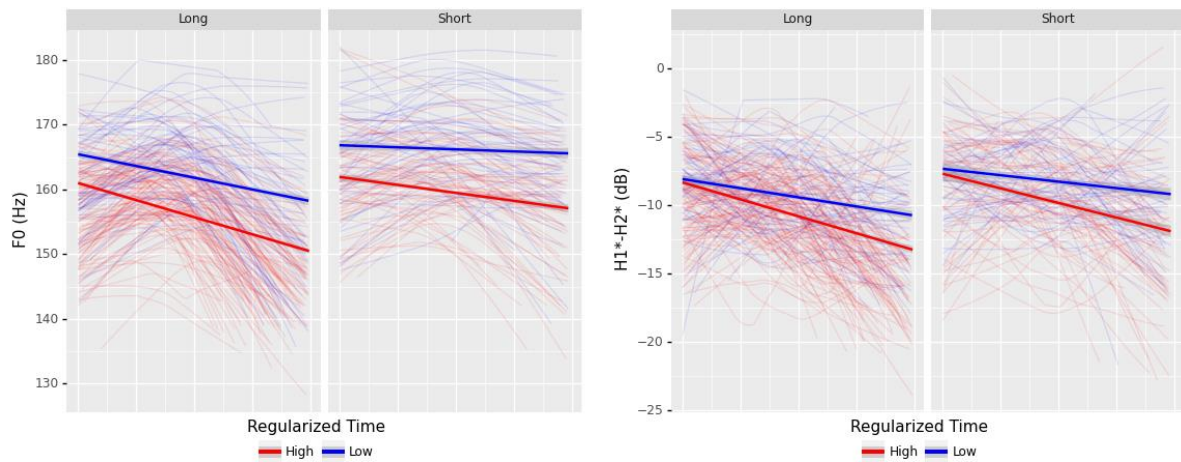


Figure 2: F1-F2 charts

3.2 Pitch and Voice Quality

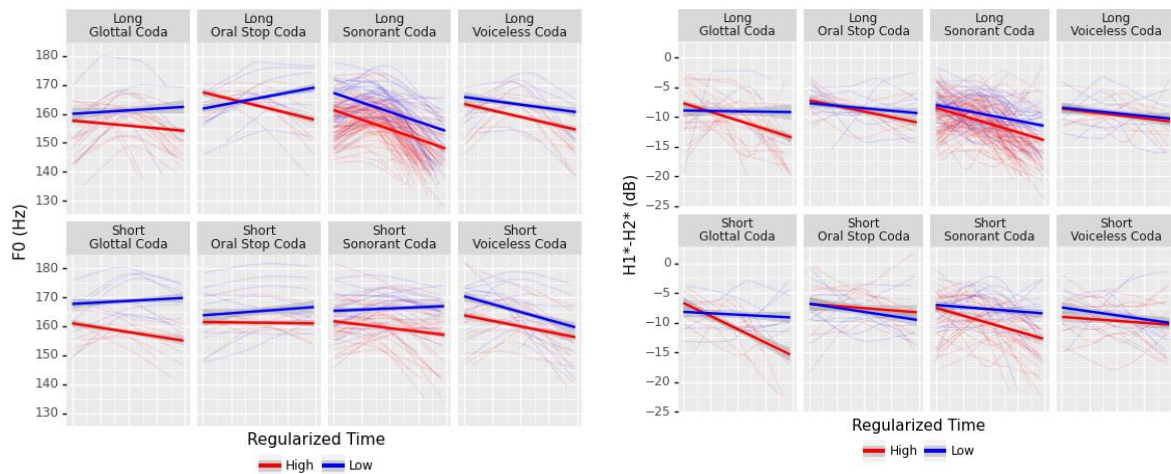
Fundamental frequency (F0) and spectral tilt (H1*-H2*), one of the frequent acoustic correlates of voice quality differences, do co-vary with pseudoregister here, though the effect is rather weak. Low pseudoregister is marked by higher F0 and H1*-H2* measures on average and high pseudoregister is accompanied by lower F0 and H1*-H2* measures on average. Figure 3 demonstrates this. The falling H1*-H2* contour indicates an increase in laryngeal tension toward the end of the vowel duration. This coincides with a decrease in F0, which indicates that glottal pulses are becoming elongated and more irregular. We would expect to see this if laryngealization were increasing towards the end of the rime, as is often the case in high pseudoregister words (more on this below), however the degree of laryngealization indicated here does not approach the strong creak achieved toward the right edge of high pseudoregister words in Ta'oi and Sedang, for example (Gehrman 2019, Smith & Sidwell 2015). Note also that the H1*-H2* measures in Figure 3 are all negative in value. Consequently, breathy voice quality is not indicated in this data and we are only dealing with degrees of modal to tensor-than-modal voice quality here.

Figure 3: F_0 and $H1^*-H2^*$ over regularized time for all vowels by vowel length



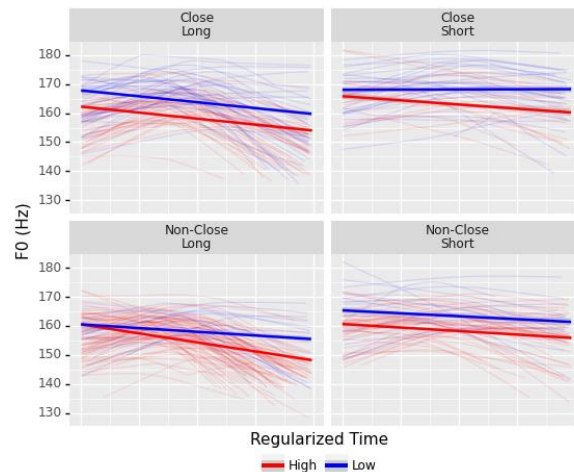
The trend evident in Figure 3 is not equally evident in all phonetic environments. While some factors, such as onset voicing, vowel height or vowel backness were *not* found to have a significant effect on this trend, coda manner of articulation does have a small effect. Figure 4 shows how the spectral tilt difference between the two pseudoregisters weakens or disappears in words with oral stop or voiceless fricative codas. Otherwise, in the high pseudoregister, spectral tilt measures are relatively low towards the right edge of words.

Figure 4: F_0 and $H1^*-H2^*$ over regularized time by vowel length and terminance type



Because the difference in F_0 between the two pseudoregisters is small, one might suspect that intrinsic F_0 is to blame for the higher F_0 measurements in the low pseudoregister. All things being equal, it is expected that close vowels will have higher F_0 than non-close vowels and, suspiciously, all phonetically close vowels in this language are in the low pseudoregister. Nevertheless, the same difference in F_0 between the two pseudoregisters is apparent for all monophthongs in Pacoh, as demonstrated in Figure 5. This indicates that the intrinsic F_0 effect is being overridden here, in service of the phonological pseudoregister distinction, to which F_0 is a phonologized, extrinsic cue.

Figure 5: F0 over regularized time by vowel length and vowel height



To summarize, it is clear that vowel quality differences are the primary and most reliable phonetic cue to the pseudoregister contrast, but there is also support for differences of pitch and voice quality playing a secondary role as redundant cues, especially in words ending with a sonorant or glottal(ized) coda. The low pseudoregister may be described as involving relatively close vowel quality and, optionally, relatively high, level pitch and relatively lax voice quality. The high pseudoregister, by contrast, is characterized by relatively open vowel quality and may optionally demonstrate relatively low-falling pitch coinciding with an increase in laryngeal tension towards the end of the vowel.

4 Register and Pseudoregister

The term register is used to describe various natural phenomena in linguistics, speech pathology and vocal pedagogy, but in Southeast Asia, the term is used to refer to a particular type of suprasegmental contrast. Even within the specific sub-domain of Southeast Asian phonology, however, register is not a well-defined concept. There is no established definition of the term, and different researchers have used the term differently. For some, phonological register may be used to refer to any suprasegmental contrast for which differential voice quality is considered the primary phonetic cue. Others are less inclusive, constraining register to only those voice quality-based contrasts which are cognate with the historical voicing of onset consonants. Some use the term register to refer to any tonal system in which differences of pitch and of voice quality work in tandem to mark tonal contrasts (i.e., a register-tone system) while others use register to refer to sets of tonemes which result from a phonemic tone split conditioned by historical onset voicing differences.

In order to properly differentiate register from pseudoregister then, a working definition of register is needed first. I propose the following definition:

Register: a binary, suprasegmental contrast upheld by a certain suite of naturally co-varying phonetic cues (or a subset thereof), which frequently arises to enhance and ultimately replace historical onset voicing contrast.

This “suite of naturally co-varying cues” may be referred to collectively as the *register bundle* of cues. These include, most notably, voice quality, vowel quality and pitch, as summarized in Table 2.

Table 2: The primary phonetic cues associated with register

	High Register	Low Register
Voice Quality	Modal	Breathy
Vowel Quality	More Open	Closer
Pitch	Higher	Lower

I define register with overt reference to the historical conditioning environment in which it develops. Typologically similar contrasts which resemble register in terms of their phonetics and/or phonology

are thus excluded if they do not arise under conditioning from differences of historical onset phonation. In addition to having a different point of origin, these other, register-like contrasts exhibit certain evolutionary potentialities which, based on the language evidence available, do not appear to be available for conventional register contrasts. These crucial differences in point of origin and ultimate evolution are why I propose the term *pseudoregister*, as a means of differentiating this phenomenon from register proper.

5 Pseudoregister formation and evolution

What, then, is the origin of pseudoregister if historical onset voicing is not involved? In all of the pseudoregister languages described thus far, pseudoregister evolved out of historical vowel quality differences, with phonetically close vowels conditioning low pseudoregister and phonetically open vowels conditioning high pseudoregister.⁴ This pattern is concordant with the well known patterns of vowel height-register interaction found in conventional register languages, whereby close vowels are stable in the low register and open vowels are stable in the high register, but high register close vowels and low register open vowels are unstable and typically become restructured over time in terms of vowel quality (Huffman 1985, Gehrman 2015). This leaves all phonetically close vowels as low register vowels and all phonetically open vowels as high register vowels. The fundamental association of close vowel height with low (pseudo)register and open vowel height with high (pseudo)register is clear, even if it less clear why this should be so.

A promising potential explanation casts tongue roots position and larynx height as the articulatory gestures which underlie both register and pseudoregister. This would explain some of the intriguing parallels between binary register contrasts in Southeast Asia and binary tongue root harmony contrasts found elsewhere in the world. We may well hypothesize that tongue-root harmony languages and register languages are both drawing on the same suite of naturally co-varying phonetic cues associated with the expansion or reduction of supraglottal cavity volume, as Gregerson (1976, 1984) suggests.

Table 3: Examples of Pseudoregister formation in Rengao

Bahnar		Rengao		Bahnar		Rengao	
PB *ɛ:				PB *i:			
babɛ:	goat	babi: ^H	goat	bri:	woods	bri: ^L	wild (forest)
ɾɛ:	rattan	ri: ^H	rattan	ɾri:	banyan tree	ɾri: ^L	banyan tree
kanɛ:	rat	kani: ^H	rat	si:	louse	ci: ^L	louse
ʔakɛ:	horn	ki: ^H	antlers	ti:	hand	ti: ^L	hand
PB *ɔ:				PB *u:			
ʔɔ:ŋ	bee	ʔo:ŋ ^H	wasp	tu:ŋ	carry	to:ŋ ^L	carry
ʔlo:ŋ	tree	lo:ŋ ^H	wood	ku:ŋ	ladder	go:ŋ ^L	stairs
bɔ:ŋ	casket	bo:ŋ ^H	coffin	su:ŋ	axe	co:ŋ ^L	axe
gɔ:ŋ	beat gong	go:ŋ ^H	gong	ʔju:ŋ	stand up	jo:ŋ ^L	sit up
PB *a				PB *ə			
nam	go	nam ^H	go	ka [?] nəm	under	ka [?] nam ^L	under
padam	five	padam ^H	five	hatəp	dig hole	tanap ^L	bury
ʔakan	woman	kan ^H	female	bət	make a dam	bat ^L	dam
mat	eye	mat ^H	eye	ʔət	hold breath	ʔat ^L	stop breathing
ɾraŋ	house post	ɾraŋ ^H	post	glək	drown	glak ^L	drown
maŋ	night	maŋ ^H	night	katəŋ	hear	taŋ ^L	hear

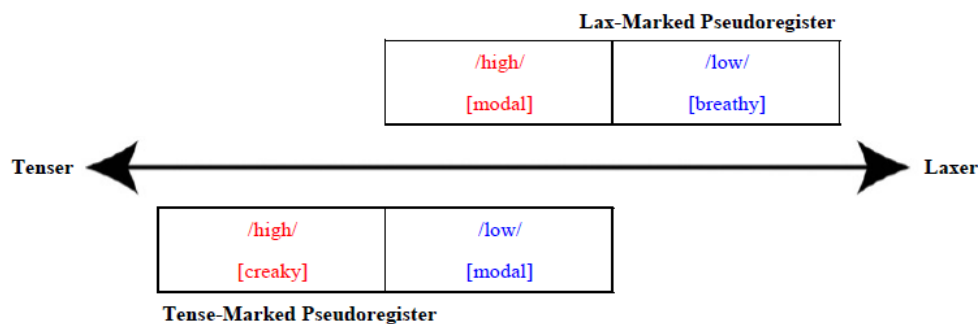
By way of example, Table 3 demonstrates the formation of pseudoregister contrast out of historical vowel height contrasts in Rengao, a North Bahnaric language, with reference to the more conservative

⁴ See examples from the Bahnaric (Smith 1972, Sidwell 2015), Katuic (Diffloth 1982; Sidwell 2005; Gehrman 2015, 2019) and Pearic (Sidwell 2019) branches of Austroasiatic.

Bahnar language (< Central Bahnaric), which retains the Proto-Bahnaric vowel contrasts in these examples.⁵

A survey of the pseudoregister languages described thus far reveals two basic typologies: *lax-marked* pseudoregister and *tense-marked* pseudoregister (Gehrman 2015). In lax-marked pseudoregister languages, the low pseudoregister is the more marked of the two in terms of voice quality, being realized with breathy voice. In tense-marked pseudoregister languages, it is the high pseudoregister that is more marked, being characterized by creaky voice / laryngealization. The North Bahnaric languages testify to the fact that historically cognate pseudoregister contrasts may be realized synchronically as either lax-marked or tense-marked. The Sedang language has a tense-marked pseudoregister contrast but cognate contrasts in other North Bahnaric languages such as the previously mentioned Rengao are lax-marked. This suggests that unlike conventional register, the phonetic voice quality cues associated with pseudoregister exist along a continuum of laryngeal tension, as illustrated in Figure 8.

Figure 6: Lax-marked vs. tense-marked pseudoregister on a continuum of laryngeal tension



That is not to say that pseudoregister languages move back and forth along this laryngeal tension continuum at random, however. On the contrary, a historical progression from more conservative lax-marked pseudoregister typology to a more innovative tense-marked typology is evident. Firstly, lax-marked pseudoregister languages retain conservative differences in vowel height between pseudoregister vowel pairs as a phonetic cue to the pseudoregister contrast, even as new cues of pitch and/or voice quality come in to enhance the original vowel quality contrast. This is not true of documented tense-marked pseudoregister languages, in which historical vowel quality differences have been erased by full phonetic merger of vowel quality while the job of cueing the pseudoregister contrast has been assumed by differential pitch and/or voice quality. Secondly, all documented tense-marked pseudoregister languages have in common the loss of historical onset voicing contrasts, while onset voicing contrasts have been retained in all documented lax-marked pseudoregister languages. This is highly suggestive that tense-marked pseudoregister languages are formerly lax-marked pseudoregister languages that have undergone a *general tensing* of the pseudoregister contrast in conjunction with the devoicing of onsets.

The catalyst for the general tensing of pseudoregister then is actually *registrogenesis*. Consequently, register and pseudoregister may in fact occur together in a language, intertwining into a complex, tone-like, register-pseudoregister hybrid as has been well documented in Chong (Ferlus 1979, 2011; Headley 1985; Sidwell 2019). The possibility of such interactions between register and pseudoregister is another argument for disambiguating the two phenomena terminologically, as I propose here.

When a pseudoregister language undergoes registrogenesis, the devoicing of voiced stop onsets begins to impart low register cues. As a result, four combinations of register and pseudoregister are produced as illustrated in Table 4. Focusing now on voice quality, a problem arises when the two “low” categories combine. Low pseudoregister has no room to become laxer, as it is already associated with breathy voice quality on the lax end of the laryngeal tension continuum (see Figure 8). However, high

⁵ Bahnar data from Banker et al. (1979), Rengao data from Gregerson & Gregerson (1977), pBahnaric reconstructions from Sidwell (2011).

pseudoregister, being associated with modal phonation, does have room to become tenser and move into the creaky/laryngealized end of the spectrum. This is exactly what we find in Chong and other tense-marked register languages, where the historical high pseudoregister shifts tenser to creaky/laryngealized and the historical low pseudoregister shifts tenser to modal voice. As a result, breathy voice is removed as a cue to the low pseudoregister category and becomes associated instead with the nascent low register category.

Table 4: *General tensing and the register-pseudoregister interaction*

		Register	
		High	Low
Pseudoregister	High	[creaky]	[breathy-creaky]
	Low	[modal]	[breathy]

The pattern in Table 4 describes exactly the pattern of development seen in the Chong register-pseudoregister complex. However, in the other documented tense-marked pseudoregister languages, we find a simplification of this scheme. In Ta'oiq and Sedang, the pseudoregister contrast remains intact and has undergone the same general tensing seen in Chong, but there is no modern reflex of conventional register. Historically voiced stop onsets have been devoiced and fully merged into the historically voiceless stops, but, there is no register contrast crosscutting and splitting the pseudoregister contrast in these languages. Register must have at least begun to develop in these languages in the past in order to catalyze the general tensing of pseudoregister, but register either failed to trigger a phonemic split in pseudoregister in these languages as it did in Chong or, if it did do so in the past, the expected 4-way contrast has since simplified to a 2-way contrast. Table 5 illustrates this state of affairs in Ta'oiq and Sedang.

Table 5: *General tensing without phonemic register (as in Ta'oiq and Sedang)*

		Register	
		Neutralized	
Pseudoregister	High	[creaky]	
	Low	[modal]	

6 What Kind of Pseudoregister Language is Pacoh?

Pacoh pseudoregister is unique. It does not fit comfortably in either the lax-marked or the tense-marked pseudoregister typology. Like tense-marked pseudoregister, Pacoh has undergone onset stop devoicing and a general tensing of the pseudoregister contrast; however, unlike tense-marked pseudoregister, vowel height remains a reliable phonetic cue for the pseudoregister vowel pairs in Pacoh. The phonetic merger of vowel quality among pseudoregister pairs, which is characteristic of other documented tense-marked pseudoregister languages, has not occurred in Pacoh. In fact, vowel quality differences are unexpectedly *the most reliable cue* to the pseudoregister contrast here, while the obvious difference of voice quality between high and low pseudoregisters which is documented in Ta'oiq, Sedang and Chong (i.e., laryngealization or lack thereof, respectively) is not found in Pacoh. The voice quality difference between the two Pacoh pseudoregisters is detectable, as demonstrated in Section 3, but much more subtle. Pitch differences are likewise subtle.

Based on the above discussion, the expected pattern of pseudoregister formation and evolution may be summarized as shown in Table 6.

Table 6: A model of pseudoregister formation and evolution

Stage	*D-	Voice Quality	Vowel Height	Examples
1: Lax-Marked	Voiced	Modal : Breathy	More Open : Closer	<i>N. Bahnaric</i> ⁶
2: Transitional	Devoiced	(Tense) : Modal	More Open : Closer	<i>Pacoh</i>
3: Tense-Marked	Devoiced	Creaky : Modal	Phonetically Merged	<i>Ta'oiq, Sedang, Chong</i>

Pacoh would appear to offer us a rare glimpse at a language intermediate between the lax-marked and tense-marked stages of pseudoregister development. However, based on the results of this admittedly preliminary acoustic study of Pacoh, it seems unlikely that this language is actually evolving in the direction of a tense-marked pseudoregister language. Historical *D onsets have clearly devoiced, and the voice quality cues to the pseudoregister contrast have clearly shifted for the tenses when compared with the lax-marked pseudoregister languages of North Bahnaric. However, while this shift has moved the historical low pseudoregister away from breathy voicing, the historical high pseudoregister has not yet become associated with creaky voice as we would expect based on the known tense-marked pseudoregister languages. Rather than a stronger emphasis on voice quality and pitch cues and a diminished role for vowel quality, we find instead the marginalization of voice quality and pitch cues and a doubling down on vowel height as the primary and only reliable cue to Pacoh pseudoregister. This hints at the possibility that pseudoregister is, in fact, fading away in Pacoh as a contrastive property of the language's phonology, with a reversion to contrasts of vowel quality. In other words, at an earlier phase, Pacoh entered into the pseudoregister formation process and saw historical vowel height contrasts re-interpreted as pseudoregister contrasts marked by a complex of cues from the register bundle which came in alongside differences of vowel height (i.e., Stage 1). However, the pseudoregister vowel pairs subsequently failed to coalesce in terms of vowel quality and remained quite strongly differentiated in terms of vowel quality, even as the language began the transitional phase with onset devoicing, registrogenesis and general tensing (i.e., Stage 2). As a result, the language is exiting the pseudoregister life cycle out of a side door, so to speak, abandoning pseudoregister and returning instead to purely segmental contrasts of vowel quality. Strengthening the argument that pseudoregister loss is in fact in progress in the High Pacoh dialect under discussion here is the fact that two other dialects of Pacoh, Cado (Gehrman 2015, 2019) and Bahi (Watson 1996), have already experienced pseudoregister loss.⁷

7 Summary and Outlook

In this paper, it has been demonstrated that the pseudoregister contrast of Pacoh splits Pacoh vocalism into a high pseudoregister cued by more open vowel quality, slightly lower pitch and slightly tenser voice quality and a low pseudoregister cued by closer vowel quality, slightly higher pitch and slightly laxer voice quality. After establishing a typological profile for pseudoregister languages based on documented instances of the phenomenon, it was shown that Pacoh does not fit neatly into either the more conservative lax-marked pseudoregister typology or the more innovative tense-marked typology. Instead, Pacoh appears to be an abortive pseudoregister language, which is abandoning the contrast and reverting to a language without register or pseudoregister, as the connection between pseudoregister vowel pairs breaks down and contrasts of vowel quality take their place.

⁶ Including Rengao, Hre, Kayong, Jeh & Halang (cf. Smith 1972, Sidwell 2015)

⁷ It remains possible that Cado and Bahi never underwent pseudoregister formation in the first place. However, the devoicing of *voiced stops without register reflexes and certain developments in their vowel inventories suggest that pseudoregister formation was common to an earlier phase of Pacoh.

This has been a mere introduction to the study of Pacoh pseudoregister, based on the speech of one speaker recorded half a century ago. Further documentation of the Pacoh language is a research priority, as there may be more conservative dialects which preserve more faithfully the old Pacoh pseudoregister contrast or, perhaps, a different iteration on it with a separate evolutionary history. This would inform the model of pseudoregister evolution put forward here.

More generally, further documentation of the pseudoregister phenomenon is sorely needed, especially when it comes to early-stage, lax-marked pseudoregister. Researchers who have the opportunity to gather data on the few remaining pre-registral / pre-tonal Austroasiatic languages should prioritize quantitative investigations into differences of voice quality and pitch among vowels of different vowel height series in these languages, in addition to looking for differences in cues from the *register bundle* after onsets of different laryngeal settings. Acoustic and perceptual studies on such languages will be necessary to make progress in understanding how vowel height contrasts may be transformed into pseudoregister contrasts.

References

- Alves, Mark. 2006. *A Grammar of Pacoh: A Mon-Khmer language of the Central Highlands of Vietnam*. (Pacific Linguistics 580). Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.
- Alves, Mark. 2007. Pacoh pronouns and grammaticalization clines. In Shoichi Iwasaki, Andrew Simpson, Karen Adams & Paul Sidwell (eds.), *SEALSXIII: Papers from the Thirteenth Meeting of the Southeast Asian Linguistics Society 2003*, 1-12. Canberra: Pacific Linguistics.
- Alves, Mark. 2015. Pacoh. In Mathias Jenny & Paul Sidwell (eds.), *The Handbook of Austroasiatic Languages*. Leiden: Brill.
- Banker, John, Elizabeth Banker & Siu Mò. 1979. *Bahnar Dictionary*. Huntington Beach, California: Summer Institute of Linguistics.
- Boersma, Paul & David Weenink. 2021. *Praat: Doing phonetics by computer (version 6.1.41)*. Online: <http://www.praat.org/>. (Accessed 05 April 2021).
- DiCanio, Christian T. 2009. The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association* 39.2:162-188.
- Diffloth, Gérard. 1982. Registres, deivoisement, timbres vocaliques: leur histoire en Katouique. [Registers, devoicing and voice quality: their history in Katuic]. *Mon-Khmer Studies* 11:47-82.
- Edmondson, Jerold A. 1996. Voice qualities and inverse filtering in Chong. *Mon-Khmer Studies* 26, 107-116.
- Ferlus, Michel. 1979. Formation des registres et mutations consonantiques dans les langues Mon-Khmer. [The formation of registers and consonant mutations in the Mon-Khmer languages]. *Mon-Khmer Studies* 8:1-76.
- Ferlus, Michel. 2011. Toward Proto-Pearic: Problems and historical implications. In Sophana Srichampa & Paul Sidwell (eds.), *Austroasiatic Studies: Papers from ICAAL4: Mon-Khmer Studies Journal Special Issue No. 2*. Dallas: SIL International; Canberra: Pacific Linguistics; Salaya: Mahidol University.
- Gehrman, Ryan. 2015. Vowel height and register assignment in Katuic. *Journal of the Southeast Asian Linguistic Society* 8:56-70.
- Gehrman, Ryan. 2019. On the origin of rime laryngealization in Ta'oiq: a case study in the phonatory restructuring of vowel quality contrasts. Paper presented at the 8th *International Conference on Austroasiatic Linguistics*, Chiang Mai University, Chiang Mai, Thailand, Aug. 29-31, 2019.
- Gregerson, Kenneth. 1976. Tongue-root and register in Mon-Khmer. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic Studies* (Oceanic Linguistics Special Publication No. 13). Honolulu: University of Hawai'i Press.
- Gregerson, Kenneth. 1984. Pharynx symbolism and Rengao phonology. *Lingua* 62:209-238.

- Gregerson, Kenneth, and Marilyn Gregerson. 1977. *Ngữ Vựng Rongao* [Rengao vocabulary]. Manila, Philippines: SIL.
- Headley, Robert. 1985. Proto-Pearic and the classification of Pearic. In Suriya Ratanakul, David D. Thomas & Suwilai Premsrirat (eds.), *Southeast Asian Linguistic Studies Presented to André-G Haudricourt*, 428-478. Bangkok: Mahidol University.
- Huffman, Franklin. 1976. The register problem in fifteen Mon-Khmer languages. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic Studies*, 575-590. Honolulu: The University Press of Hawaii.
- Huffman, Franklin. 1985. Vowel permutations in Austroasiatic languages. In Graham Thurgood, James A. Matisoff and David Bradley (eds.), *Linguistics of the Sino-Tibetan Area: The State of the Art. Papers Presented to Paul K. Benedict for his 71st Birthday*, 141-45. Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.
- Kibirige, Hassant, Greg Lamp, Jan Katins, gdowning, austin, matthias-k, Tyler Funnell, Florian Finkernagel, Jonas Arnfred, Dan Blanchard, Sergey Astanin, Eric Chiang, Paul Natsuo Kishimoto, Evan Sheehan, stonebig, Bernard Willers, Robert Gibboni, smutch, Yaroslav Halchenko, Pavel, Brian King, Min RK, John Collins, zachcp, Anthony, Bevan Koopman, Carlos H. Grohmann, Dan Becker, Dan Brown & Daniel Saiz. 2021. *Plotnine (v0.8.0)*. Online: <https://github.com/has2k1/plotnine>. (Accessed March 27, 2021).
- Kirby, James. 2020. *PraatSauce*. Online: <https://github.com/kirbyj/praat sauce> (Accessed 05 April 2021).
- L-Thongkum, Theraphan. 1991. An instrumental study of Chong registers. In Jeremy Davidson (ed.), *Essays in Mon-Khmer Linguistics in Honour of H. L. Shorto*, 141–160. London: School of Oriental and African Studies.
- Nguyễn Văn Lợi, Đoàn Văn Phúc & Phan Xuân Thành. 1986. *Sách học Tiếng Pakôh-Taôih*. [Study on the Pacoh-Taoh Language]. Hanoi, Tỉnh Bình tri Thiên.
- Sidwell, Paul. 2005. *The Katuic Languages: Classification, Reconstruction and Comparative Lexicon*. Munich: Lincom Europa.
- Sidwell, Paul. 2011. *Proto-Bahnaric*. Unpublished manuscript. Online: <http://sealang.net/monkhmer/database/>
- Sidwell, Paul. 2015. Local drift and areal convergence in the restructuring of Mainland Southeast Asian languages. In N. J. Enfield and Bernard Comrie (eds.), *The Languages of Mainland Southeast Asia: The State of the Art*, 51-81. Berlin: De Gruyter Mouton.
- Sidwell, Paul. 2019. Proto-Pearic and the role of vowel height in register formation. Paper presented at the *8th International Conference on Austroasiatic Linguistics*, Chiang Mai University, Chiang Mai, Thailand, Aug. 29-31, 2019.
- Smith, Kenneth & Paul Sidwell. 2015. Sedang. In Matthias Jenny & Paul Sidwell (eds.), *The Handbook of Austroasiatic Languages*, 789-836. Leiden: Brill.
- Smith, Kenneth. 1972. *A Phonological Reconstruction of Proto-North-Bahnaric*. (Language Data: Asian-Pacific Series, No. 2). Santa Ana, California: Summer Institute of Linguistics.
- Vitrano-Wilson, Seth, Ryan Gehrman, Carolyn Miller & Cheung Xaiyavong. 2018. Tone marks as vowel diacritics in two scripts: Repurposing tone marks for non-tonal phenomena in Cado and other Southeast Asian languages. *Writing Systems Research* 10.1:43-67.
- Watson, Richard, Sandra Watson & Cubuat Canxoiq. 1979. *Pacoh-Vietnamese-English Dictionary*. Huntington Beach, CA: SIL.
- Watson, Richard, Sandra Watson & Cubuat Canxoiq. 2013. *Pacoh-Vietnamese-English Dictionary*. SIL International. Online: <http://pacoh.webonary.org/> (Accessed 16 July 2021)
- Watson, Richard. 1964. Pacoh phonemes. *Mon-Khmer Studies* 1:135-148.
- Watson, Richard. 1964. Pacoh phonemes. *Mon-Khmer Studies* 1:135-148.

- Watson, Richard. 1966. *Reduplication in Pacoh*. Hartford, CT: Hartford Seminary Foundation MA thesis.
- Watson, Richard. 1977. Discourse elements in a Pacoh narrative. *Mon-Khmer Studies* 6:279-322.
- Watson, Richard. 1996. Why three phonologies for Pacoh? *Mon-Khmer Studies* 26:197-205.
- Watson, Richard. 2011. A case for clitics in Pacoh. In Sophana Srichampa, Paul Sidwell & Ken Gregerson (eds.), *Austroasiatic Studies: Papers from ICAALA*, 222-232. (Mon-Khmer Studies Special Issue No. 3). Dallas, SIL International; Salaya, Mahidol University; Canberra, Pacific Linguistics.
- Watson, Sandra. 1964. Personal pronouns in Pacoh. *Mon-Khmer Studies* 1:81-97.
- Watson, Sandra. 1966. Verbal affixation in Pacoh. *Mon-Khmer Studies* 2:15-30.
- Watson, Sandra. 1976. The Pacoh noun phrase. *Mon-Khmer Studies* 5:219-232.
- Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. Online: <https://ggplot2.tidyverse.org>. (Accessed 05 April 2021).

TOWARDS A PROSODIC TYPOLOGY OF KHERWARIAN MUNDA LANGUAGES: SANTALI OF ASSAM

Luke HORO

*Living Tongues Institute for Endangered
Languages*
luke.horo@livingtongues.org

Gregory D. S. ANDERSON

*Living Tongues Institute for Endangered
Languages*
gdsa@livingtongues.org

Abstract

In this article, we detail the prosodic typology of disyllables in Santali spoken in Sonitpur district, Assam, India. Previous investigations present a chaotic picture of contradictory analyses. For Santali, Ghosh (2008:30) states “(s)tress is always on the second syllable of the word regardless of whether it is an open or a closed syllable”. Neukom (2001:8), on the other hand, claims that in disyllabic stems, “(s)tress falls on the first syllable; however, if the first syllable is light and the second heavy (iambic structure), stress falls on the second syllable”, i.e., a Quantity Sensitive system. Bodding (1922) detailed a partly morpholexical system of prominence assignment. To test this, we recorded data from male and female speakers of various ages in a noise-free environment in the field. Forms were recorded in isolation, in a quasi-focal frame “I ____ said”, and in an out of focus frame “I ____ LOUDLY said, not SOFTLY” to control for focal intonation effects and those of phrasal or utterance intonation. Based on subsequent instrumental analysis, we suggest that Assam Santali always shows prominence on the second syllable of disyllables, cued by intensity, f₀ and duration—a pattern like that attested in Assam and Odisha lects of Sora (Horo and Sarmah 2015, Horo 2017, Horo et al. 2020), contra Donegan and Stampe (2004)—an iambic pattern and not quantity sensitive, at least in disyllables.

Keywords: Santali, Sora, Munda, Prosody, Prominence
ISO 639-3 codes: sat

1 Introduction

In this article, we detail a preliminary study on the prosodic typology of the Kherwarian Munda languages, a group of Austroasiatic languages spoken in eastern and northeastern India, focusing here on one lect of Santali spoken in Sonitpur district, Assam, India. This variety has not been described previously, nor have any Kherwarian lects been studied experimentally using instrumental phonetic methods before.

The major group of Santali speakers resides to the southwest of this region in West Bengal, Jharkhand, and Northern Odisha. Some of these varieties of the language have been subjected to previous analysis in some domains, but not to date in terms of instrumental phonetic analysis. Thus, only impressionistic statements have been made to date in the literature about the prosodic or intonational system of Santali. Indeed, the previous investigations present a chaotic picture of contradictory analyses.

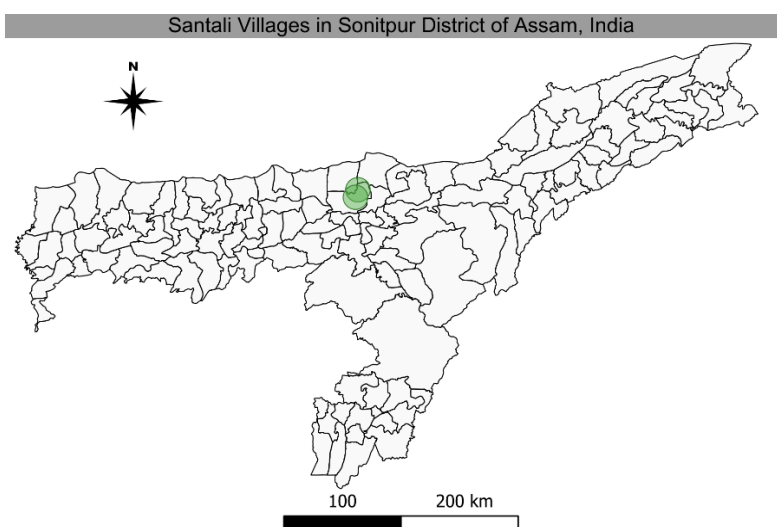
For Santali, Ghosh (2008:30) states Santali has fixed second position stress while Neukom (2001:8) claims a Quantity Sensitive system where a heavy syllable following a light one takes stress. Otherwise, it is the initial syllable. Bodding (1922) for his part detailed a partly morpholexical system of prominence assignment. Thus, subjecting this analysis to the rigors of instrumental verification is an ongoing process. In this preliminary study, we limit ourselves to uninflected lexical items that are in their underlying structure disyllabic, with a few that surface as trisyllabic with a weak medial syllable resolving a word medial onset cluster. Moreover, forms were recorded in isolation, in a quasi-focal frame “I ____ said”, and in an out of focus frame “I ____ LOUDLY said, not SOFTLY” to control for

focal intonation effects and those of phrasal or utterance intonation. Based on the data acquired in the above formats and subsequent phonetic analysis, prominence in Santali disyllables is explored and the findings, supported by acoustic evidence, are reported in this paper.

2 Santali of Assam

According to the Census of India (Registrar General of India 2011) there are 7,368,192 Santali speakers in India, of which 213,139 speakers are reported in Assam. The Santali variety of Assam has not been analyzed previously. The current study presents a preliminary finding that is based on speech data recorded from Santali speakers living in two villages of Assam's Sonitpur district, namely, Erasuti and Borbil (see Figure 1). There are approximately eight hundred Santali individuals in Erasuti and one hundred and fifty Santali individuals in Borbil, and both villages have a major concentration of Santali speakers in Sonitpur District of Assam. Also, the inhabitants claim to have lived in those villages for at least four generations until now. Moreover, unlike most ethnic Munda inhabitants of Assam, who are reported to have migrated to Assam as indentured tea laborers from parts of Eastern India in the nineteenth century (Tea Districts Labour Association-India 1924), the Santali community recorded in this work claim that they are native inhabitants of the land, not migrant laborers.

Figure 1: Map of Assam highlighting locations of Santali speech data collection



For the purposes of this study, a total of six native Santali speakers (three male and three female) living in the two villages were interviewed and recorded to acquire the data. The average age of the participants is thirty-one years with a standard deviation of eleven years. Each participant is a multilingual speaker, and besides Santali they also speak Sadri, Assamese, and Hindi as their second, third and fourth languages. Among the six participants, only one male and one female have completed their high school education, whereas the others dropped out of their formal education either during high school or even earlier.

2.1 Data collection

To collect Santali speech samples, a dataset was created from a list of basic Santali vocabulary wherefrom, the text data, including fifty-one Santali disyllabic words having (C)V(C).(C)CV(C)¹ syllable structures, were generated. All words in the dataset are non-derived nouns, including words for

¹ Four targeted disyllabic words that have onset clusters in the second syllable, [tʃepre], [sikri], [kʰapri] and [pandʒra], were interchangeably produced as trisyllabic words with the insertion of an epenthetic vowel [ə] breaking the obstruent and rhotic consonant clusters present in the words. Such variants are not included in the analysis.

body parts, animal names and words for natural objects. Table 1 presents a subset of the text data used in this study to generate the speech data.

Table 1: Subset of Santali disyllabic data used in the study

	Santali	English	Syllable Structure
1.	<i>aɔo</i>	'urine'	V.CV
2.	<i>ipil</i>	'star'	V.CVC
3.	<i>supu</i>	'arm'	CV.CV
4.	<i>lutur</i>	'ear'	CV.CVC
5.	<i>ninda</i>	'night'	CVC.CV
6.	<i>kaʃkom</i>	'crab'	CVC.CVC

Subsequently, the speech data is derived from eliciting and recording the text data once in isolation and once in each of the two sentence frames shown in (1) and (2).

- (1) *ij* _____ *men-kediŋ*
 1SG _____ say-PST.TR/ACT.1SUBJ
 'I said _____'
- (2) *ij* _____ *gula=te* *men-kediŋ* *lahe=te=do* *baŋ*
 1SG _____ loud=ADV say-PST.TR/ACT.1SUBJ soft=ADV=CONJ NEG
 'I said _____ loudly not softly'

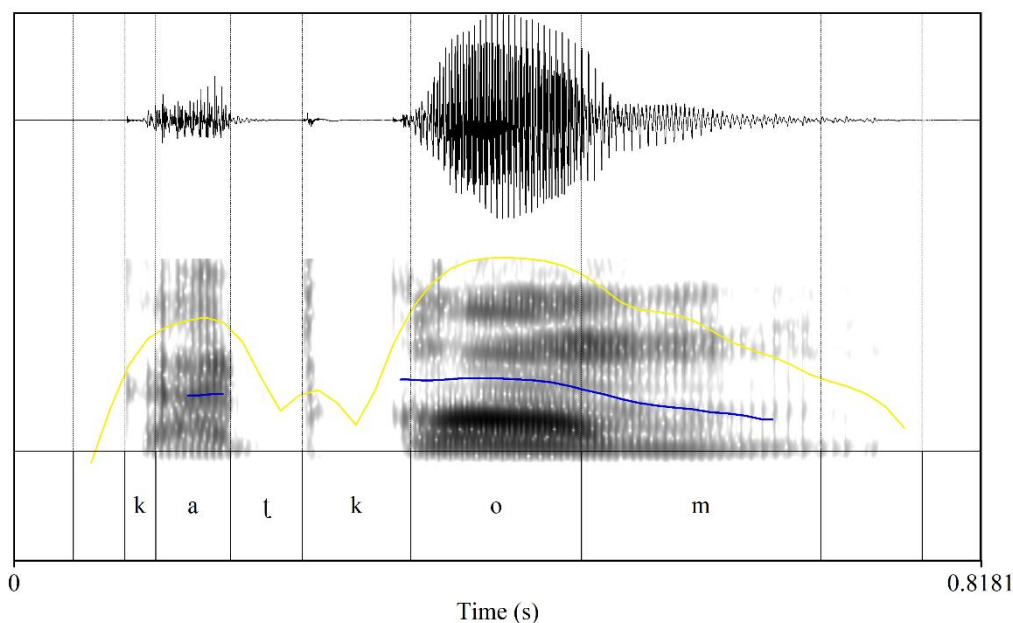
In both (1) and (2), the blank space is replaced by a target word from the text data. The phrasal position in (1) is intended to capture phrasal prominence in the target words and control for any speech perturbation that may be caused during the production of words in isolation, and the target position in (2) is intended to record the words in an out of focus or an unaccented intonational context to control for possible information structure effects. The fifty-one unique lexical items recorded thrice from six individuals produced a total sample size of 918 Santali disyllables, of which 46.13% have the CV.CV syllable structure with the vowels [a, i, e, o, u]² as the syllable nucleus of both the first and second syllables. The data were recorded in a noise-free environment in the field using a head-worn unidirectional Shure mic connected by XLR cable to a Tascam linear PCM recorder, and the digital data are stored at a sampling frequency of 44.1 kHz and a bit depth of 32 bits in .WAV format.

2.2 Data analysis

The Santali speech data collected from the field were subjected to phonetic analysis by means of acoustic phonetic methods, for which purpose the data were manually annotated for word boundary and phoneme boundary in Praat (Boersma and Weenink 2020) using the spectral and temporal cues of speech sounds. Thereby, vowel sounds are annotated between the beginning and end of glottalic pulses, and sonorant sounds are annotated in low amplitude regions. In the case of the obstruents, they are annotated between the release of the oral closure and the beginning of the glottalic pulses in onset position and between the end of glottalic pulses and the point of oral closure in coda position (see Figure 2).

² The data set also includes three nasal vowels [ã, ï, ê] with [ã] appearing in the first syllable, [ï] appearing in the second syllable and [ê] appearing in both first and second syllables; a diphthong [ai] appearing in the second syllable and a lax front vowel [ɛ] appearing in the second syllable; they are for now not treated as unique syllable nuclei in the general analysis. However, for examining the interaction between syllable prominence and vowel types, the nasal vowels, the diphthong(s) and the lax front vowel are not included. Also, Ghosh (2008) includes schwa in his Santali vowel phoneme inventory, but Bodding (1922) considers it to be an allophone of [a] in words with high vowels. Until we have done instrumental analysis on the vowel system of various Santali lects, we reserve judgment on this issue too.

Figure 2: Spectrographic illustration³ of phonetic annotation of a Santali disyllable as produced by a male Santali speaker living in Sonitpur district of Assam



2.2.1 Analyzing prominence in Santali disyllables

The analysis of prominence in Santali disyllables is based on three acoustic cues, namely, vowel duration, vowel intensity and fundamental frequency (Fry 1955; 1958). While vowel duration is calculated from the absolute length of the vowel nuclei in first and second syllables, vowel intensity is measured from the mean amplitude of the entire length of the vowel nuclei. Likewise, fundamental frequency is estimated from the mean of the entire vowel segment. Additionally, all values for the three acoustic parameters are normalized for speaker variability using the z -score normalization method ($z = (x - \mu) / \sigma$)⁴. The normalized data is then analyzed and visually represented in R Version 3.5.3 (R Core Team, 2019) using the *ggplot2* package (Wickham et al. 2016) through its built-in functions of *geom_boxplot* and *geom_density* where the normalized values of vowel duration, vowel intensity and average fundamental frequency are treated as dynamic variables. Syllable positions and the contexts of the utterance (isolation, phrasal frame, unaccented frame) are treated as factor variables. Further, the same normalized data is tested with a one-way analysis of variance (ANOVA) using the *aov* function in R Version 3.5.3 (R Core Team 2019). For this purpose, normalized values of vowel duration, vowel intensity and average fundamental frequency are treated as dependent variables and syllable position (first versus second) and context of utterance (isolation, phrasal and unaccented) are treated as independent variables.

2.2.2 Analyzing segmental effect on prominence in Santali disyllables

This study also includes an investigation of segmental effects on syllable prominence in Santali disyllables. First, to examine the interaction between syllable prominence and vowel types, a subset of data, having only the five frequently occurring vowels [a, i, e, o, u] in first and second syllables of Santali disyllables, is used. In this analysis, vowel duration, vowel intensity and average fundamental frequency of the five vowels are measured separately in the two syllable positions. The same data is then visually represented using the built-in functions of *geom_boxplot* and *facet_grid* in the *ggplot2* package (Wickham et al. 2016) of R Version 3.5.3 (R Core Team, 2019), in which the normalized values

³ This sound file can be heard and accessed in the Santali Living Dictionary at

<https://livingdictionaries.app/santali/entries/f0nG6dDN9GlqY0WPs6jC>.

⁴ z = Normalized value; x = Individually extracted values; μ = Mean of x ; σ = Standard Deviation of x .

of vowel duration, vowel intensity and average fundamental frequency are treated as dynamic variables and syllable positions, vowel types and contexts of utterances are treated as factor variables.

Second, the effect of different consonantal environments on syllable prominence was examined. To do the analysis, a subset of data, including only the frequently occurring CV.CV syllable structure, that have onset consonants [h, k, m, r, tʃ] in both syllables, was used. Thereby, vowel duration, vowel intensity and average fundamental frequency in first and second syllables that have the five onset consonants were examined and compared with each other. This analysis is also represented using the same visualization method described above for examining the interaction between syllable prominence and vowel types except that the vowel type factor is changed to onset consonant type factor for this analysis.

3 Findings: Prominence in Santali disyllables

3.1 Vowel duration

Vowel duration is a common acoustic cue to determine prominence in languages whereby longer vowel duration is considered a robust indicator of prominence in languages of the world (Gordon and Roettger 2017). In the case of Santali (Ghosh 2008:23), vowel length is not reported to be phonemically distinct, and the data presented here exhibits only the phonetic length of vowel segments as produced in syllable nuclei of disyllabic words. By examining the speech data as produced by the Santali speakers recorded in this study, it was observed that vowel duration is generally longer in the second syllable as compared to vowel duration in the first syllable in Santali disyllables. Figure 3 demonstrates the vowel duration distinction in first and second syllables in the targeted disyllables as recorded in the three utterance contexts.

From Figure 3, it is evident that in Santali disyllables, average vowel duration in the second syllable is always longer than in the first syllable. Additionally, in Figure 4, the density plots reveal that the distribution of vowel duration in first and second syllables is distinct in all three utterance contexts. Also, a distinct skewing was observed in the data, indicating that longer vowel durations occur in second syllables, whereas shorter vowel durations occur in first syllables.

Figure 3: Average Vowel Duration in First (1) and Second (2) syllables in Santali Disyllables.

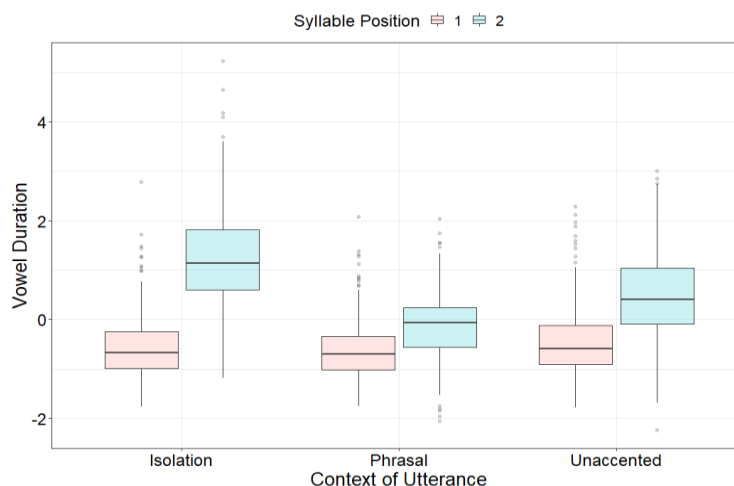
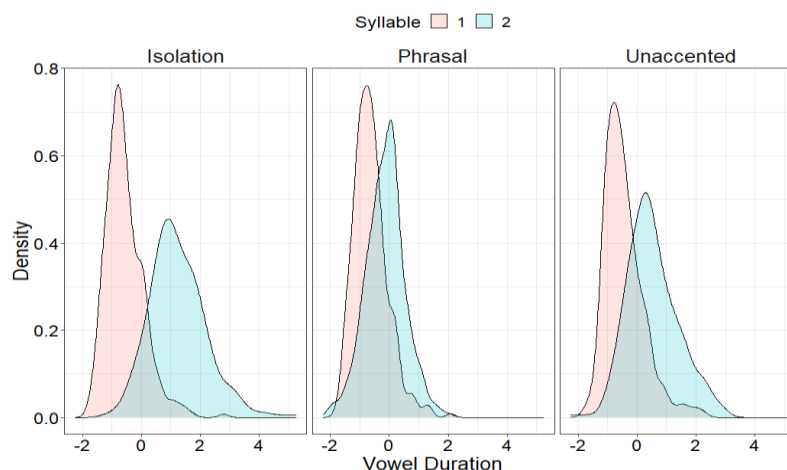


Figure 4: Density distribution of vowel duration differences in first (1) and second (2) syllables in Santali disyllables.



In this regard, it is notable that the distinction in vowel duration is greater when the disyllabic words are produced in isolation than when they are produced in the phrasal and in the unaccented intonational contexts.⁵ Moreover, the difference in vowel duration in first and second syllables of Santali disyllables is found to be statistically significant in isolation [$F(1,642) = 796.4, p < 0.001$], in phrasal contexts [$F(1,629) = 101.3, p < 0.001$], and in unaccented intonational contexts [$F(1,622) = 248.9, p < 0.001$].

3.2 Vowel intensity

Vowel intensity refers to the acoustic energy in a vowel segment which is normally greater in prominent syllables than in non-prominent syllables. Unlike vowel duration, vowel intensity is a less robust cue for diagnosing prominence in languages, yet there is evidence that prominent syllables are distinct from non-prominent syllables with respect to their mean intensities (Gordon and Roettger 2017). Accordingly, in the present data, it is observed that both the averages as well as the overall distribution of average vowel intensity in first and second syllables of Santali disyllables are distinct from each other (see Figure 5 and 6).

From Figure 5, it is revealed that in Santali disyllables the average vowel intensity in second syllables is greater than the average vowel intensity in first syllables. Also, this is found to be true for Santali disyllables produced in all three utterance contexts. Likewise, the density plots in Figure 6 reveal that the distribution of average vowel intensity in first and second syllables is distinct wherein a skewing towards higher vowel intensity is observed in second syllables but a skewing towards lower vowel intensity is observed in first syllables. Also, the average vowel intensity difference in each of the three utterance contexts, namely, isolation [$F(1,642) = 30.62, p < 0.001$]; phrasal [$F(1,629) = 74.8, p < 0.001$] and unaccented intonational contexts [$F(1,622) = 33.72, p < 0.001$] are found to be statistically significant.

⁵ This may indicate that vowel duration plays a role in demarcating the end of utterances, but this remains to be examined more systematically.

Figure 5: Average vowel intensity in first (1) and second (2) syllables in Santali disyllables

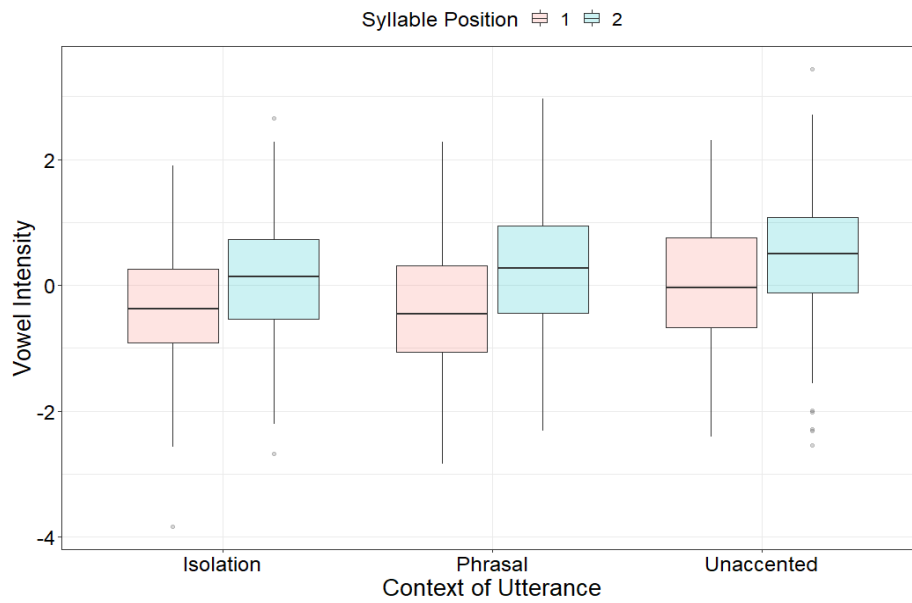
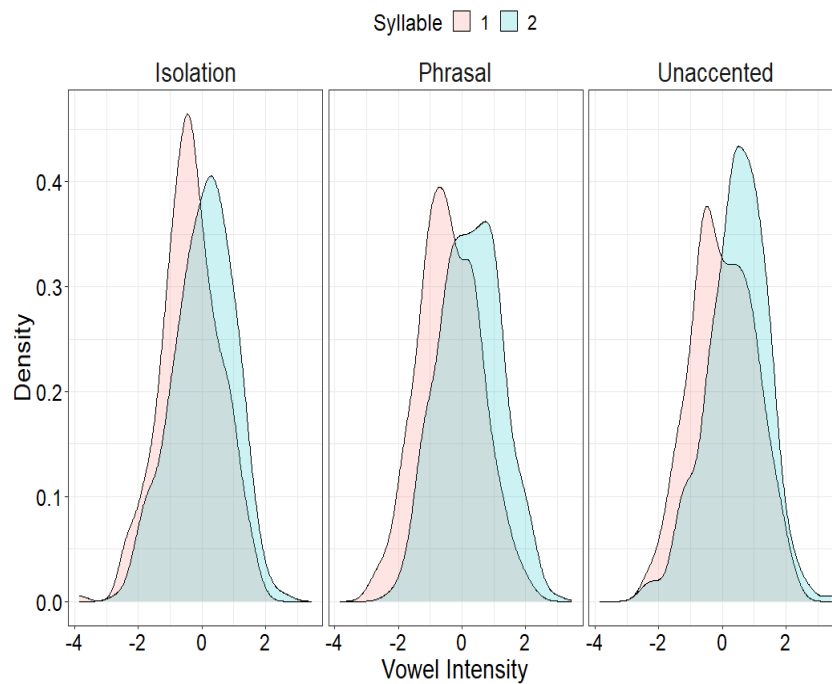


Figure 6: Density distribution of vowel intensity differences in first (1) and second (2) syllables in Santali disyllables



3.3 Fundamental frequency

Fundamental frequency represents pitch variation in speech sounds, and a systematic variation in pitch across syllables is known to be an indicator of syllable prominence in various languages (Gordon and Roettger 2017). Generally, prominence is associated with higher pitch which is expressed by greater fundamental frequency in the prominent syllable as opposed to lower fundamental frequency realized in the non-prominent syllable. The Santali data examined in this work reveals a similar pattern, but an exception is also observed in the analysis. Figure 7 shows the average fundamental frequency differences in first and second syllables of Santali disyllables.

Figure 7: Average fundamental frequency in first (1) and second (2) syllables in Santali disyllables

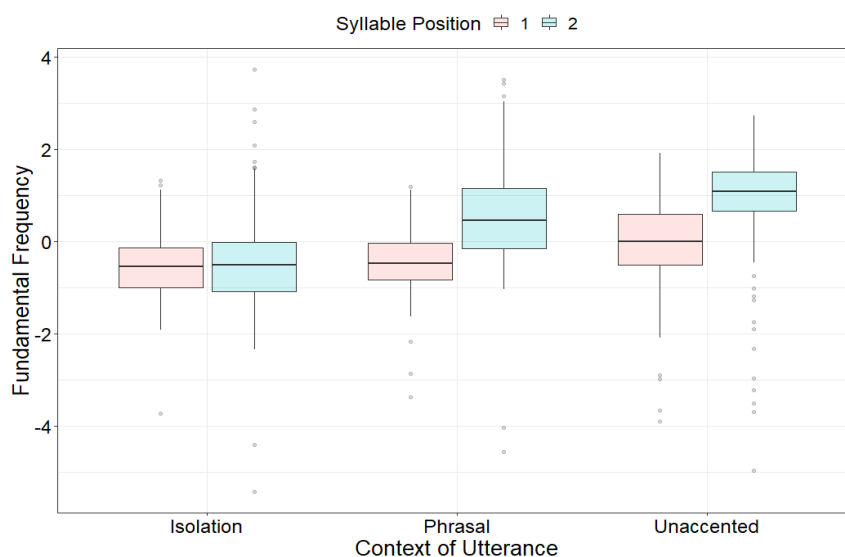


Figure 8: Density distribution of Fundamental Frequency in First (1) and Second (2) syllables in Santali Disyllables

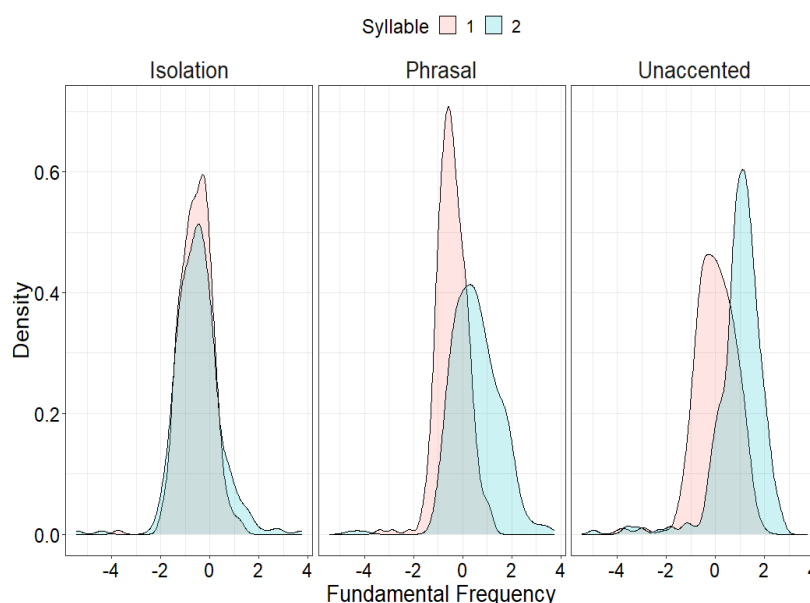


Figure 7 reveals that, in Santali disyllables, while the average fundamental frequency is higher in the second syllable than in the first syllable in the phrasal and unaccented intonational contexts, variation is absent when the disyllabic words are uttered in isolation. The same pattern is observed in the density plots presented in Figure 8, which shows that the distribution of average fundamental frequency in first and second syllables is not distinct in words that are spoken in isolation, whereas the same distributions appear to be distinct in words that are spoken in both the phrasal and unaccented intonational contexts. In this regard, the findings are also confirmed through statistical analysis whereby the average fundamental frequency difference in first and second syllables of Santali disyllables is found to be distinct with statistical significance when spoken in phrasal [$F(1,629) = 229.8$, $p < 0.001$] and unaccented [$F(1,622) = 183.1$, $p < 0.001$] intonational contexts but not when spoken in the isolation [$F(1,642) = 1.354$, $p = 0.245$] context. Thus, the analysis here suggests that although prominence, realized by higher pitch, in second syllable is present in Santali disyllables, the distinction is likely to be neutralized in words that are produced in isolation. However, this observation requires further

investigation⁶ with substantial data evidence which could not be achieved in this preliminary study of prominence in Santali disyllables.

4 Findings: Segmental effects on syllable prominence

To explore the interaction between syllable prominence and segmental types and to see if the prominence of the second syllable over the first syllable in Santali disyllables is consistently maintained in different vowel types and in different consonantal environments, the three acoustic cues of prominence, namely, vowel duration, vowel intensity and fundamental frequency are further examined in this study using two separate subsets of data. Firstly, the five vowels in Santali [a, e, i, o, u], separated by their syllable positions, were examined to see if all vowel types render similar patterns of syllable prominence or not. Secondly, a subset of data containing only the CV.CV syllable structure and having the onset consonants [h, k, m, r, tʃ] in both syllables was examined to see if changes in the consonantal environment also effects syllable prominence or not.

4.1 Segmental effects on duration

With respect to vowel duration in first and second syllables having different vowel nuclei, it is observed that all vowel types have longer vowel duration only in the second syllable in all three contexts (see Figure 9). This implies that in Santali disyllables, the five vowels [a, e, i, o, u] are phonetically longer only when they occur in the second syllable but are relatively shorter when they occur in the first syllable, and that prominence of the second syllable manifested by longer vowel duration is not affected by changes in vowel types in the syllable nuclei. Likewise, by examining vowel duration in the environment of the five Santali consonants [h, k, m, r, tʃ], it is observed that in Santali disyllables that have any of the five consonants in the onset position, vowel duration is longer in the second syllable only, and not in the first syllable (see Figure 10). Also, the pattern is observed to be consistent in all three utterance contexts included in this study. This implies that in Santali disyllables, different onset consonants do not have an impact on syllable prominence cued by vowel duration.

⁶ A further investigation of pitch variation in first and second syllables can be achieved by extracting pitch values (fundamental frequencies) at different pitch timings including initial, medial, and final. This can show potential pitch changes that may be undisclosed in the average pitch distinction in the two syllable positions in the isolation context. Also, maximum, and minimum pitch values can be extracted, which can be utilized for examining potential separation of pitch ranges in the two syllable positions when the disyllables are produced in isolation.

Figure 9: Average vowel duration in first (1) and second (2) syllables in Santali disyllables having different vowel nucleus

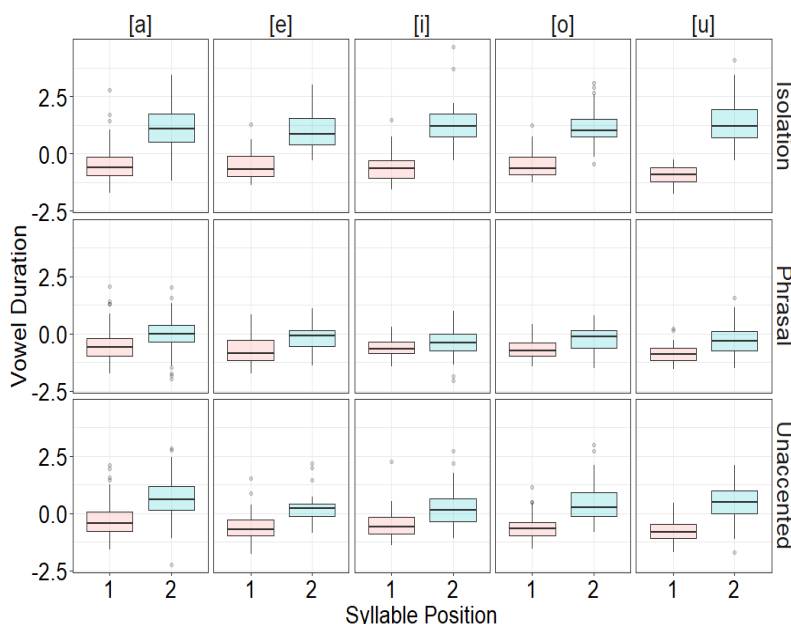
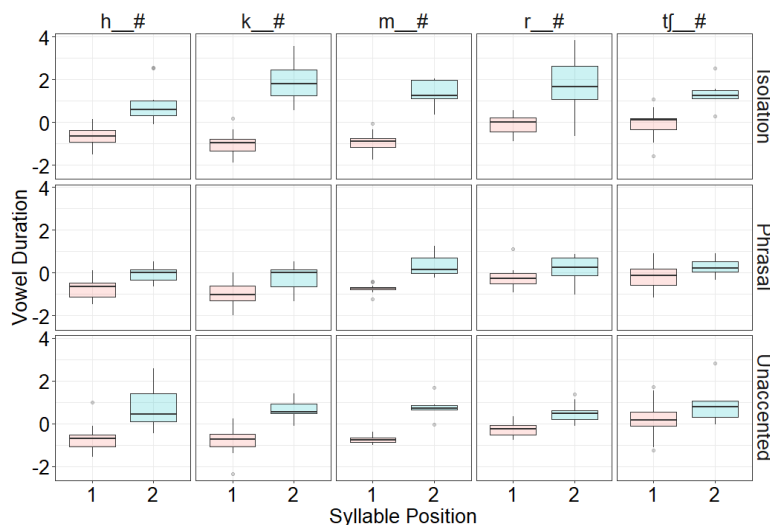


Figure 10: Average vowel duration in First (1) and Second (2) syllables in Santali disyllables with CV.CV syllable structure having different consonantal environments



Thus, based on the data presented in Figures 9 and 10, it is confirmed that vowel duration is a robust cue for identifying prominence in Santali disyllables whereby neither the vowel types nor the onset consonants in CV.CV syllable structures affect the manifestation of prominence in the second syllable.

4.2 Segmental effects on intensity

The general examination of vowel intensity in first and second syllables of Santali disyllables revealed that vowel intensity is normally higher in the second syllable. In the micro analysis, while the tendency to produce higher intensity in the second syllable is found consistently in the mid vowels [e, o] and the low vowel [a], in the case of the two high vowels [i] and [u], the vowel intensity difference between the two syllables is observed to be neutralized and even reversed for the high front vowel [i] in the unaccented intonation context (see Figure 11). This indicates that the two high vowels [i] and [u] in

Santali disyllables do not have distinct vowel intensities between the two syllables. Also, while the distinction is observed to be equally neutralized in all three utterance contexts for the back high vowel [u], in case of the front high vowel [i], the difference is neutralized in isolation and in the phrasal context but is reversed in the unaccented intonational context. Thus, from analysing the interaction between vowel types and vowel intensity in Santali disyllables, it is revealed that prominence of second syllable cued by higher vowel intensity is maintained in disyllables that have either the mid vowel or a low vowel in the syllable nuclei, whereas the pattern is likely to be neutralized or reversed if the disyllables have only the high vowels in their syllable nuclei.

Figure 11: Average vowel intensity in First (1) and Second (2) syllables in Santali disyllables having different vowel nucleus

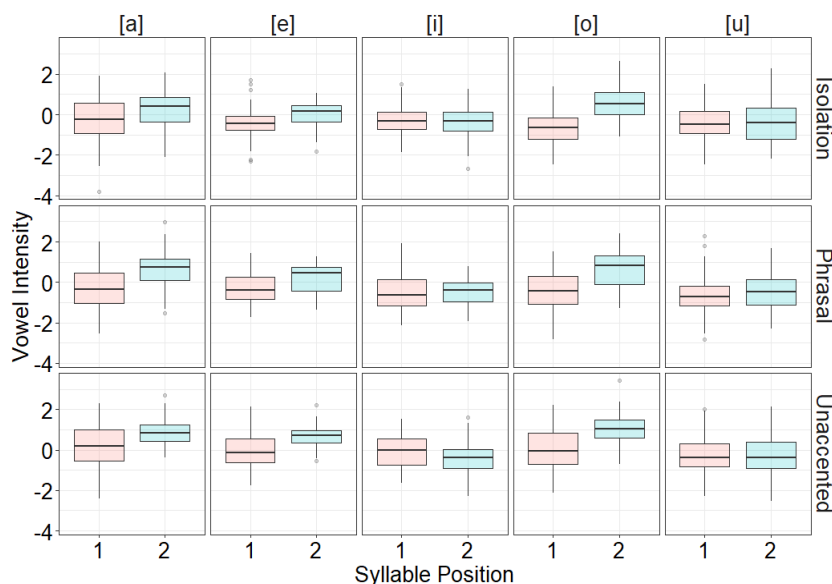
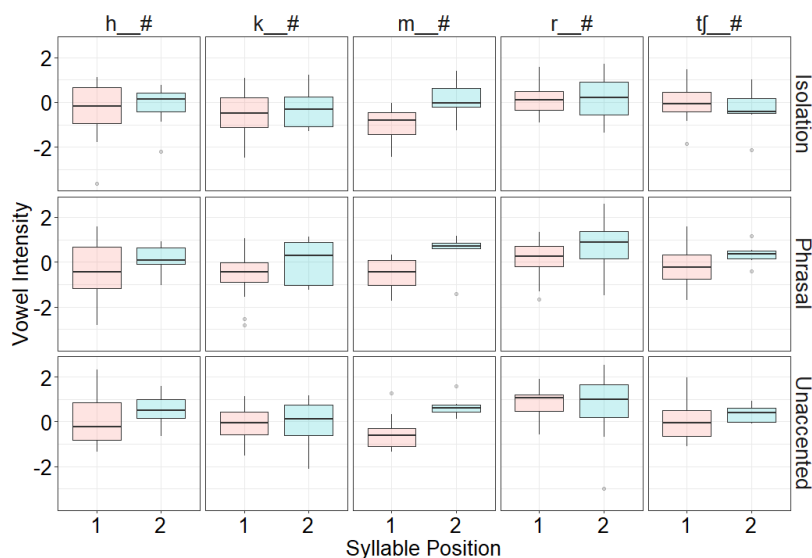


Figure 12: Average vowel intensity in First (1) and Second (2) syllables in Santali disyllables with CV.CV syllable structure having different consonantal environments



In addition to the variations in vowel intensity arising from vowel types, an examination of the interaction between vowel intensity and different onset consonants reveals that vowel intensity differences in first and second syllables may be neutralized or reversed even in the environment of

certain consonant types in the onset position. Of the five onset consonants examined in this study, the analysis here reveals that the general tendency to produce higher vowel intensity in the second syllable has only been consistently maintained in all three contexts of utterances in the environment of the bilabial nasal consonant [m] (See Figure 12). On the other hand, vowel intensity distinctions between first and second syllables are observed to be neutralized in the environment of the voiceless velar consonant [k] and alveolar trill consonant [r] in the isolation and unaccented intonational contexts. The difference is seen to be reversed in the environment of the voiceless palatal affricate consonant [tʃ] in the isolation context only. These observations imply that like vowel types, onset consonant types in the CV.CV syllable structure also have an impact on the rendering of vowel intensity in Santali disyllables, whereby consonants such as [m] and [h] appear to have no or minimal impact. In contrast, consonants such as [k] and [r] may neutralize the intensity distinction, and consonants such as [tʃ] may even reverse the vowel intensity distinction in Santali disyllables.

Thus, from analysing the interaction between vowel intensity and vowel types and onset consonant types it is revealed that, unlike vowel duration, vowel intensity is a less robust cue for identifying syllable prominence in Santali disyllables. Specifically, at the micro level, it is found that there are certain segmental exceptions that probably suppress the general tendency to produce higher intensity vowels in the second syllable of Santali disyllables.

4.3 Segmental effects on fundamental frequency

Fundamental frequency difference in Santali disyllables has been shown to be sensitive to various utterance contexts whereby higher fundamental frequency in the second syllable is exhibited only in the phrasal and unaccented intonational contexts but not in the isolation context of utterance. Significantly, the same pattern of fundamental frequency distinction is observed when the five Santali vowels [a, e, i, o, u] are analysed separately with respect to their syllable positions in the disyllables. Figure 13 presents the vowel-wise fundamental frequency distinction in first and second syllables of Santali disyllables in the three contexts of utterances included in this study. From Figure 13 it is evident that all five Santali vowels have higher fundamental frequency in the second syllable when they are produced in the phrasal and unaccented intonational contexts, but as an exception not in the isolation context. This implies that differences in vowel types do not impact the second syllable prominence depicted by higher fundamental frequency in Santali disyllables.

Similarly, by examining the effect of different onset consonants on the fundamental frequency of the vowel nuclei of Santali disyllables, it is observed that besides the lack of fundamental frequency distinction in first and second syllables in the isolation context of utterance there is only a minimal impact of onset consonant types even in the phrasal and unaccented intonational contexts of utterances. Figure 14 presents the fundamental frequency difference in first and second syllables of Santali disyllables that are grouped according to their onset consonants [h, k, m, r, tʃ] and the contexts of utterances that are included in this study.

Figure 13: Average fundamental frequency in first (1) and second (2) syllables in Santali disyllables having different vowel nucleus

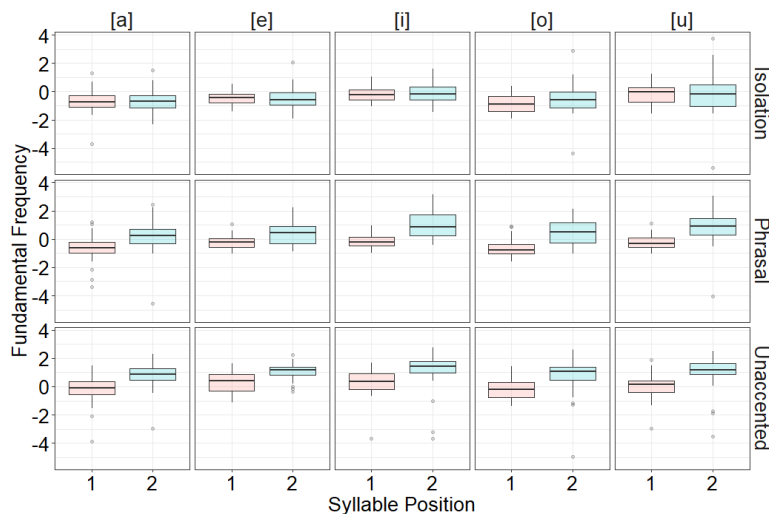
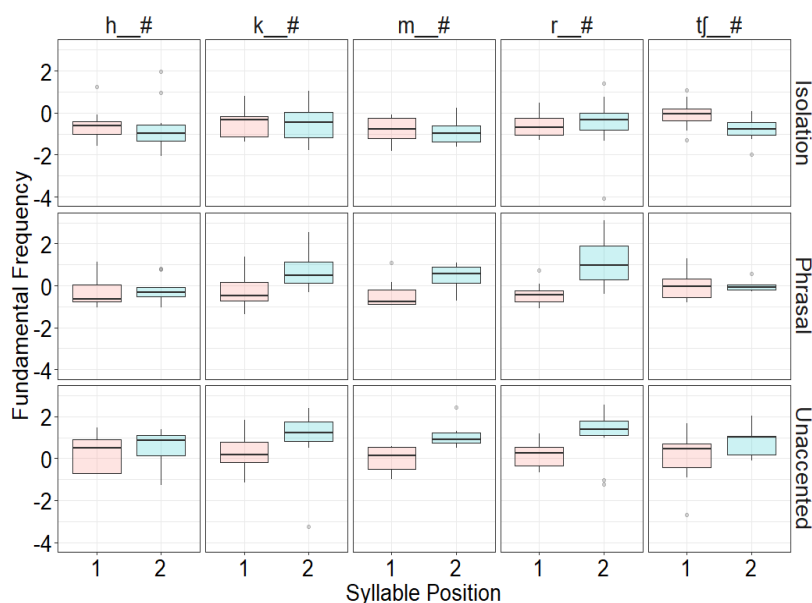


Figure 14: Average fundamental frequency in first (1) and second (2) syllables in Santali disyllables with CV.CV syllable structure having different consonantal environments



From Figure 14, it is observed that while a fundamental frequency difference between first and second syllables of Santali disyllables may remain neutralized in different consonantal environments when the target word is said in isolation for words beginning in [k, m, r], in the very same context, the difference may even be reversed in the environment of the glottal fricative consonant [h] and voiceless affricate consonant [tʃ]. Also, the same consonantal environments appear to have resulted in neutralizing or minimizing the fundamental frequency difference between first and second syllables of Santali disyllables even when they are produced in the phrasal and unaccented intonational contexts of utterances.

Thus, the analysis of vowel-wise fundamental frequency differences in first and second syllables of Santali disyllables revealed that vowel types do not impact syllable prominence in the language. Additionally, the exceptional case of neutralizing fundamental frequency differences in first and second syllables in the isolation context of utterance is also confirmed from the examination of the interaction of vowel types and fundamental frequency in Santali disyllables. However, an analysis of

different onset consonants indicates that two consonant types, namely, fricatives and affricates may either minimize the fundamental frequency difference in first and second syllables or even reverse the difference when Santali disyllables with CV.CV syllable structures bearing the two consonants in the onset position are recorded in isolation only.

5 Comparison with Sora

Despite claims in the literature to the contrary (Donegan 1993, Donegan and Stampe 1983, 2004), instrumental analyses show that Sora clearly has second syllable prominence in disyllabic forms (Horo and Sarmah 2015, Horo 2017, Horo, Sarmah and Anderson 2020). Phonetic data representing Sora speech varieties of four geographical locations in Assam, namely, Singrijhan, Sessa, Lamabari and Koilamari and one geographical location in Odisha, namely, Raiguda, provide evidence that vowels in the second syllable are longer, louder and pitched higher than the vowels in the first syllable of disyllables. Accordingly, the acoustic cues of prominence reveal that vowel duration, vowel intensity and fundamental frequency (f_0) are generally higher in second syllables only; see Figures 15 to 17.

Figure 15: Average vowel duration in first and second syllable of Sora disyllables (Horo 2017)

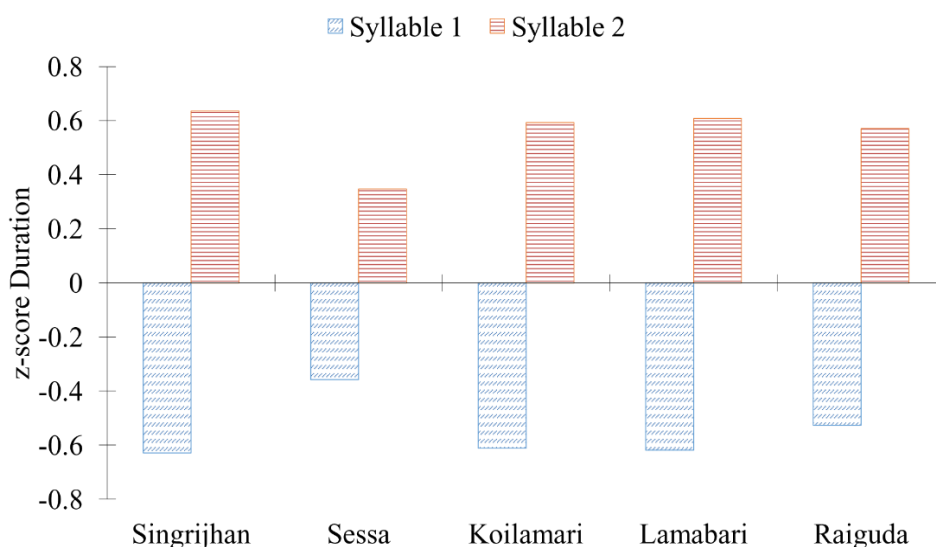


Figure 16: Average f_0 in first and second syllable of Sora disyllables. (Horo 2017)

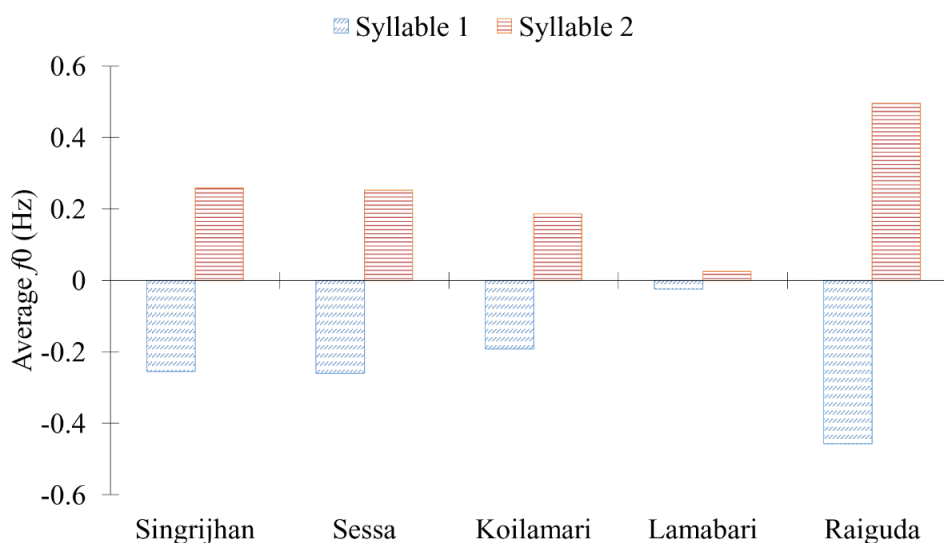
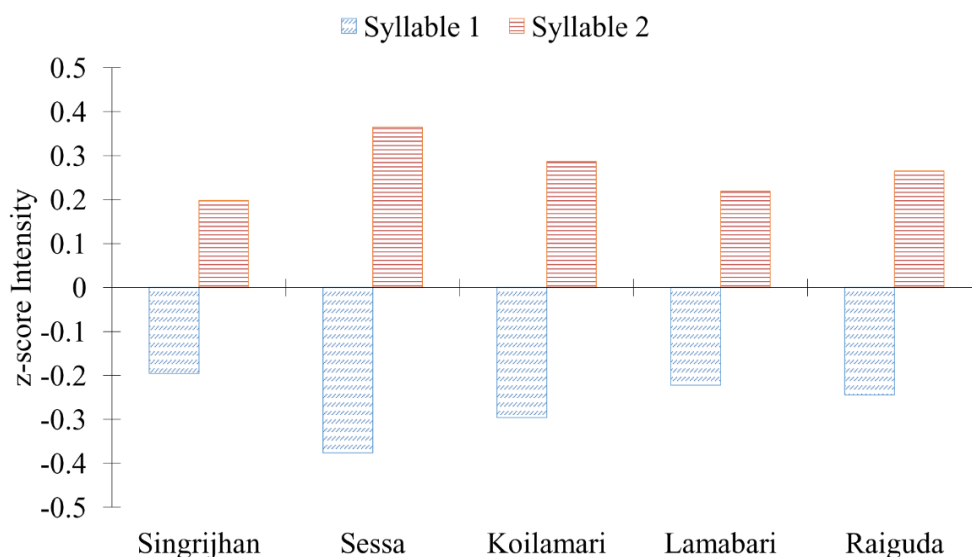


Figure 17: Average vowel intensity in first and second syllable of Sora disyllables. (Horo 2017)

Likewise, other Munda languages have been claimed to show similar patterning as well (e.g., Remo (Bhattacharya 1968) or West Bengal Santali (Ghosh 2008)), but no instrumental data has been offered in support of this, even while we believe those published observations to be accurate. However, Assam Santali, based on preliminary phonetics analysis and supported by acoustic evidence, shows that second syllable prominence is the pattern attested in disyllabic words.

On the other hand, Indo-Aryan and Dravidian languages, including those of Odisha, at least those studied instrumentally to date which may in fact be limited to Oriya (Mahanta 2010) and Telugu (which is not an official language of Odisha but is spoken in Parlakhemundi in the extreme south of Odisha, on the border) typically are trochaic or first-syllable prominent in similar contexts having more peripheral vowels in the first syllable and more phonological contrasts attested, all pointing to first-syllable prominence (Khan 2016).

6 Discussion

The phonetic analysis supported by acoustic evidence above suggests that duration and intensity are largely consistent cues of prominence in the Assam lect of Santali when investigated across all three utterance contexts, viz., in isolation, in a phrasal frame and in an unaccented frame. In each instance, the second syllable reflects greater prominence with respect to these cues than does the corresponding first syllable in these disyllabic words. Fundamental frequency is also distinct and points to prominence on the second syllable over the first syllable in Assam Santali as well. However, unlike duration and intensity, fundamental frequency is only statistically relevant as an acoustic cue of prominence in phrasal and unaccented utterance contexts. In the isolation context on the other hand, fundamental frequency does not cue word-level prominence *per se* in the Assam Santali lect examined. Simultaneously, as the isolation context may potentially reflect word-level intonation as well as intonation of the full utterance level—two areas where pitch variation can well be distinct—there may be intonational parameters that operate on the full utterance level that interact with and potentially override word-level intonational or prominence cueing parameters. However, what these may be requires further research. Also, duration does not appear to show sensitivity to specific vowel types or consonant contexts in their functions as cues of prominence. Intensity differences between first and second syllables on the other hand appear to be sensitive to both the presence of high vowels and the presence of [k], [r] or [tʃ] in onset position, while fundamental frequency may show some sensitivity to consonantal environment, specifically the presence of an initial fricative [h] and affricate [tʃ].

Furthermore, that the phrasal frame data in our study—which is a potentially inherently a quasi-focus position—and the unaccented position frame—which is explicitly an out of focus—largely pattern

together with respect to the three acoustic cues of prominence suggest that focal intonation is not active *per se* in determining the distribution of the three examined cues of word-level prominence.

7 Broader Munda perspectives

That our data show that the Assam Santali lect under investigation showing second syllable prominence in disyllables may be surprising to some scholars, but others will note this might be expected due to previous analyses of West Bengal and Jharkhand Santali lects. Those who may find this surprising might be scholars who would either predict that Santali should be similar to data from other South Asian language groups such as the Indo-Aryan lects that it is in contact with, or they may assume this based on claims that have been made in previous investigations, not of Santali specifically, but hypotheses that are said to apply to all Munda languages, and thus Santali by extension.

To be sure, one camp of scholarship has long asserted that Munda languages exclusively and only show trochaic prosodic patterns (i.e., first syllable prominence in disyllables). This is best and most succinctly encapsulated in the theory of ‘rhythmic holism’ put forth by Donegan (1993) and Donegan and Stampe (1983, 2004). They suggest that there was a one-time shift from iambic ‘rhythm’ to trochaic rhythm in Munda at the proto-Munda level, and that it was this prosodic shift that triggered a series of cascading changes that caused the wholesale typological restructuring of proto-Munda from isolating to agglutinative-synthetic and to verb-final syntactic structure, and so on, which resulted in modern Munda languages ultimately representing the mirror image of their Austroasiatic sister languages that remained in Southeast Asia. However, such claims have been refuted by various scholars in recent work on Munda, such as Horo and Sarmah (2015), Horo (2017a), Horo et al. (2020), Anderson (2015b), Anderson (2020) and Ring and Anderson (2018).

As mentioned above, other scholars have made claims about the prosodic structure of Santali in print that do not align with our experimental findings, although it should be mentioned here that no acoustic or statistical data have been offered by any of the other scholars mentioned below with respect to the Santali data, so their claims must therefore be considered impressionistic and preliminary as a result. Thus, Neukom (2001:8) claims that in disyllabic stems, in Santali, “(s)tress falls on the first syllable; however, if the first syllable is light and the second heavy (iambic structure), stress falls on the second syllable”. Therefore, he suggests that Santali reflects a Quantity Sensitive system. Our data as presented above on disyllables do not support this view. Regardless of syllable shapes in the words, the pattern is always the same: the acoustic cues that indicate prominence in Santali disyllables, including intensity, fundamental frequency, and duration, conspire to make second syllables prominent in Santali disyllables over the first syllable in such words. This is also true regardless of whether the word appears in a quasi-focal frame or in an explicitly out of focus frame for all three examined cues of prominence, and in isolation as well for duration and intensity.

Note as also mentioned above that not all previous researchers agree with Neukom’s take on the Santali data. Thus, Ghosh (2008:30) states “(s)tress is always on the second syllable of the word regardless of whether it is an open or a closed syllable”. Our instrumental acoustic and statistical data do in fact support Ghosh’s view, at least in uninflected disyllables.

As the present investigation is just a preliminary study, we have for the time being limited ourselves to only examining disyllabic lexemes in this one lect, Assam Santali. Boddington’s (1922) musings on the topic of the system of prominence attested in Santali are rather involved. In short, casting things into a modern typology of prosodic systems or prominence assignment, he suggests that when taking into consideration all Santali words, including inflected forms of verbs which can create rather lengthy morphological strings (or g[rammatical] words), the system of prominence should be considered to be morphologically specified and constrained. He suggests that there may be certain morphemes capable of bearing stress while others do not. Whether this will be verified instrumentally remains a task for future research.

It may also be the case that there are mismatches between phono-prosodically defined word domains (such as prominence or vowel harmony, see Anderson, Horo and Harrison 2022) and larger morphological complexes (g-words) that may constitute more than one phono-prosodic ‘word’ or

represent phrases prosodically, even if functioning syntactically as units.⁷ Moreover, it appears that in Munda languages more generally there are morphological word elements or grammatical morphemes that appear to be more integrated into the phono-prosodic domain of roots and form single phono-prosodic words with such lexical hosts, but others remain outside of such word domains. Put differently, some grammaticalized elements may be affixes and others rather clitics,⁸ some eligible to be assigned prominence for example and some that are not, and yet others that are variable in this regard. A task for our future research on Santali is to determine what the prosodic and morphotactic characteristics of the full range of inflectional elements are that may constitute Santali morphological word complexes and how exactly these interact (or do not interact) with the system of prominence assignment. Only once such a study is complete will we be able to definitively determine the entire system of prominence that characterizes this important Kherwarian Munda language.

Santali is far from alone within Munda in presenting a confusing picture asserted in different publications about what the system of prominence might be. A wide variety of other, often conflicting, claims about the intonational structure of individual Munda languages have appeared in print. With few exceptions (e.g., Rehberg 2003 for Kharia) instrumental data are not used as the basis for the analysis offered, so the interpretations remain largely impressionistic. With respect to Rehberg (2003), she proposed that low pitch on the initial syllable in a disyllabic word of Kharia followed by a high pitch is what signalled prominence, which while of course possible, seems largely motivated by a desire to conform the attested data in Kharia with the standard view that Munda languages are trochaic, such that it must be low pitch that signals prominence in Kharia if this trochaic pattern is true. We reserve further consideration of the Kharia data until we have had a chance to subject the data to our own instrumental analysis, but simply comment that the language seems to have a five-vowel system of phonemes and the realization of the name of the language has a schwa in the initial and allegedly prominent syllable, even while the schwa does not appear to be a phoneme in the language. Furthermore, intensity may increase concomitant with the raising of pitch in second “unaccented” syllables in Kharia (Peterson 2011) further underscoring that Kharia likely has second syllable prominence in disyllables, not first-syllable prominence.

Mundari, a language closely related to Santali, embodies the lack of clarity about the system of prominence that typifies our present understanding of Munda languages as expressed in print. Even whether ‘stress’ exists per se is debated, as Osada (1992:36) considers Mundari to be a pitch accent language, while Cook (1965:100), Langendoen (1963:14-15), and Sinha (1975:39), consider Mundari to be a stress language. But these latter three researchers do not agree on what the system of stress is. Sinha considers the language to have a quantity sensitive system whereby disyllabic words of the shape $C^1V^1C^2V^2$ or $C^1V^1C^2V^2C^3$ (where C^2 can also be a homo-organic nasal+stop sequence) stress the *second* syllable, but in disyllabic words of the shape $C^1V^1C^2C^3V^2$, stress falls on the initial syllable and in trisyllabic words, stress falls on the 2nd syllable regardless of the shape. Cook (1965) states that only if the final syllable is closed, it is accented, otherwise it is the initial syllable in disyllabic words. Osada (2008:104) states that if a word is trisyllabic, stress can *only* be on the second or the third syllable: on the third syllable if that is not a suffix, otherwise it falls on the second syllable in Mundari trisyllabic words, but *never* on the first syllable, regardless of syllable weight.

⁷ This was suggested about Munda and Khasian language to explain some observed differences between apparent phrases vs. lexemes in Khasian Pnar and Sora (Ring and Anderson 2018).

⁸ Santali clearly reflects such a system with subject markers which function as clitics in both imperfective and perfective series of inflections. Subject clitics in Santali do not even preferentially target the verbal ‘word’ as their host, but rather target the word immediately preceding the verb, or they may appear at the end of the morphological verbal word (or on occasion both places simultaneously). See Anderson (2007, 2015a, 2015b, 2020) for more details.

Table 2: Munda language prominence patterns and acoustic cues (based on Hildebrandt and Anderson 2021)

Language	Prominent Syllable	Acoustic Cues/Notes
Sora	Second syllable in disyllables (Horo 2017)	Duration, Intensity, Fundamental Frequency
Gorum	Penult/final syllable (Anderson & Rau 2008)	Not yet formally investigated
Gutob	Final/second syllable (Voß p.c.)	H-Pitch, other cues unclear. Not yet formally investigated
Remo	Second syllable (Bhattacharya 1968)	Intensity, Pitch, Duration? Not yet formally investigated
Gta?	Second/final syllable (Anderson in preparation)	Not yet formally investigated
Kharia	Conflicting: Initial/second/final? (Rehberg 2003, Peterson 2011)	L-Pitch (Initial) H-Pitch (Second/final) Intensity on non-initial syllables. Not yet formally investigated
Juang	Conflicting: Second vs. initial (Patnaik 2008, Dasgupta 1978)	Not yet formally investigated
Koṛowa	Second/last syllable of stem (Barker 1953)	Duration? Not yet formally investigated
Ho	Conflicting: QS/initial or Morpholexical (Nottrott 1882, Pucilowski 2013)	Not yet formally investigated
Mundari	Conflicting: Pitch accent vs. Stress accent Conflicting: Initial vs second vs. Morpholexical (Osada 1992, 2008; Cook 1965; Sinha 1975; Hoffmann 2001; Langendoen 1963)	Pitch, other cues? Not yet formally investigated
Santali	Conflicting: Second syllable vs. QS/initial (Bodding 1922, Neukom 2001, Ghosh 2008)	Intensity, Duration, Pitch to some degree
Korku	Second/QS? (Zide 2008)	Not yet formally investigated

However, all this aside, our data clearly shows that in disyllables, prominence is found on the second syllable in the Assam Santali lect we discuss here. Therefore, Santali appears to reflect the same pattern that has previously been identified for both Assam and Odisha lects of the distantly related Sora language, also of the Munda family (Horo and Sarmah 2015, Horo 2017a, Horo et al. 2020). Given the typological, geographic and genetic distance between Santali and Sora, and given strong areal tendencies against second syllable prominence in South Asia, and the fact that most of language groups that Munda is related to phylogenetically within the Austroasiatic phylum show a similar (and seemingly) cognate system of prominence assignment, one might be tempted to suggest that these data may point us in the direction of assuming that proto-Munda may well have been second syllable prominent in disyllables as the simplest explanation for the observed parallels between Sora and Santali, and impressionistically, other Munda languages as well, such as Gta?, in addition to their demonstrable similarity to other Austroasiatic groups.

8 Future Research Goals: Towards an intonational typology of the Munda languages

In this paper we are beginning the first step on a long journey to compare the interface between phonological structure and prosodic features and the morphosyntax of the Munda languages. First, we must determine what the patterns and cues of prominence for each of the languages are, or at least a representative set of the languages. From basic uninflected words, we expand this typology for inflected

forms of words and phrases and see what patterns emerge and how these different elements combine and whether such combinations exhibit distinct phono-prosodic patterns. This larger work is underway for Sora currently and will expand to Santali in the next year. We will extend this similar investigation to other Munda languages representing different branches of the family tree with an eye to not only grounding future discussions of Munda prosodic structures in instrumental phonetics analysis supported by acoustic evidence, but also how these structures interface with the complex morphosyntax of the languages.

9 Summary

None of the previously mentioned scholars relied on instrumental data for their analyses, which therefore remain impressionistic. Our study is based on instrumental analysis and suggests that Santali (at least as spoken in Assam) *always* shows prominence on the second syllable of disyllables, cued by intensity, f_0 and duration. This suggests that Assam Santali shows a pattern similar to that attested in Assam and Odisha lects of Sora (Horo and Sarmah 2015, Horo 2017, Horo et al. 2020), contra Donegan and Stampe (2004). In other words, it is an iambic pattern and not quantity sensitive, at least in disyllables.

However, resolving what possible historical development the Munda languages have undergone both individually and at the proto-language level remains several steps away as we must first engage in the systematic synchronic analysis of word, phrase, and utterance level prominence systems (and subsets thereof) for the attested languages of the family before we have an adequate empirical basis to engage in such far reaching but important questions. The present study is just a preliminary step on this journey.

References

- Anderson, Gregory D. S. 2007. *The Munda Verb. Typological Perspectives*. Berlin: Mouton de Gruyter.
- Anderson, Gregory D. S. 2015a. Overview of the Munda languages. In M. Jenny and P. Sidwell (eds.) *The Handbook of Austroasiatic Languages*. Volume 1. Grammar Sketches of the World's Languages. Mainland and Insular South East Asia. Amsterdam: Brill. 364-414.
- Anderson, Gregory D. S. 2015b. 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. 2020. Proto-Munda Prosody, Morphotactics and Morphosyntax in South Asian and Austroasiatic Contexts. In M Jenny, P. Sidwell and M. Alves (eds.) *Austroasiatic Syntax in Areal and Diachronic Perspective*. Leiden: Brill, 157-197.
- Anderson, Gregory D. S. in preparation. The Gta? Language. *Plains Gta? and Hill Gta? Texts, Lexicon and Grammar*.
- Anderson, Gregory D. S., Luke Horo and K. David Harrison. 2022. Vowel Harmony in the Munda Languages. Submitted to H. van der Hulst and N. Ritter (eds.) *Oxford Handbook of Vowel Harmony*. Oxford: OUP.
- Anderson, Gregory D. S. and Felix Rau. 2008. Gorum. In Gregory D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge (Taylor and Francis), 381-433.
- Barker, Phillip R. 1953?/nd. *The phonemes of Korowa*. Unpublished manuscript. Seattle, Washington.
- Bhattacharya, Sudhibhushan. 1968. *A Bonda Dictionary*. Building Centenary and Silver Jubilee Series 18. Poona [Pune]: Deccan College.
- Bodding, P. O. 1922. *Materials for a Santali Grammar, mostly phonetic*. Dumka: Santal Mission.
- Cook, Walter A. 1965. *A Descriptive Analysis of Mundari: A study of the structure of the Mundari language according to the methods of linguistic science*. Ph.D. Dissertation, Georgetown University.
- Dasgupta, Dipankar. 1978. *Linguistic Studies studies in Juang, Kharia Thar, Lodha Mal-Pahariya, Ghatoali Pahariya*. Calcutta: Anthropological Survey.

- 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., & Stampe, David. 1983. Rhythm and the Holistic Organization of Language Structure. In J. F. Richardson, M. L. Marks, & A. Chukerman (eds.), *Papers from the parasession on the Interplay of Phonology, Morphology and Syntax*. Chicago Linguistic Society, Chicago: 337-353.
- Donegan, Patricia J., & Stampe, David. 2004. Rhythm and the synthetic drift of Munda. In R. Singh (ed.), *The Yearbook of South Asian Languages and Linguistics*. Mouton de Gruyter, Berlin, Germany: 3-36.
- Fry, Dennis B. 1955. Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America* 27, no. 4: 765-768.
- Fry, Dennis B. 1958. Experiments in the perception of stress. *Language and speech* 1, no. 2: 126-152.
- Ghosh, Arun. 2008. Santali. In Gregory D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge (Taylor and Francis), 11-98.
- Gordon, Matthew K. and Timo. B. Roettger 2017. *Acoustic Correlates of Word Stress: A Cross-Linguistic Survey*, *Linguistics Vanguard*, 3(1). DOI: <https://doi.org/10.1515/lingvan-2017-0007>. Accessed 4 October, 2021.
- Hildebrandt, Kristine A. and Gregory D. S. Anderson. 2021. Word Prominence in Languages of Southern Asia. To appear in Ksenia Bogomolets and Harry van der Hulst (eds.), *Word prominence in polysynthetic languages*. Oxford: Oxford University Press.
- Horo, Luke. 2017. *A Phonetic Description of Assam Sora*. Ph.D. Dissertation, Indian Institute of Technology, Guwahati.
- Horo, Luke & Priyankoo Sarmah. 2015. Acoustic analysis of vowels in Assam Sora. In L. Konnerth et al. (eds.), *Northeast Indian Linguistics* 7:69-86. Canberra: ANU.
- Horo, Luke, Priyankoo Sarmah and G. D. S. Anderson. 2020. Acoustic Phonetic Study of the Sora Vowel System. *Journal of the Acoustical Society of America* 147.4:3000-3011.
- Langendoen, Terry. 1963. *Mundari Phonology*. Unpublished manuscript. Cambridge, Massachusetts.
- Neukom, Lukas. 2001. Santali. *Languages of the World/Materials* 323. München: Lincom.
- Nottrott, Rev. A. 1882. *Grammatik der Kolh-Sprache*. Gütersloh: C. Bertelsmann.
- Osada, Toshiaki. 1992. *A reference grammar of Mundari*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.
- Osada, Toshiaki. 2008. Mundari. In Gregory D. S. Anderson (ed.), *The Munda Languages*. Routledge Language Family Series. London: Routledge (Taylor and Francis), 99-164.
- Patnaik, Manideepa. 2008. Juang. In Gregory D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge (Taylor and Francis), 508-556.
- Peterson, John M. 2011. *Grammar of Kharia*. Leiden: Brill.
- Pucilowski, Anna. 2013. *Ho Morphology and Morphosyntax*. Ph.D. Dissertation, University of Oregon.
- Registrar General of India. 2011. *Distribution of the 100 Non-Scheduled Languages – India, States & Union Territories – 2011 Census*. New Delhi: Office of the Registrar General and Census Commissioner, India.
- Rehberg, Kerstin. 2003. *Phonologie des Kharia–Prosodische Strukturen und segmentales Inventar*. MA Thesis, Universität Osnabrück.
- Sinha, N. K. 1975. *Mundari Grammar*. Mysore: Central Institute of Indian Languages.
- Wickham H., Chang W., Wickham M.H. 2016. *Package ‘ggplot2’*. Create Elegant Data Visualisations Using the Grammar of Graphics. Version 2: 1-189.
- Zide, Norman H. 2008. Korku. In Gregory D.S. Anderson (ed.), *The Munda Languages*. Routledge Language Family Series. London/New York: Routledge (Taylor and Francis), 256-298.

EFFECTS OF VOICELESS AND PREGLOTTALIZED NASALS ON F0 IN EASTERN KHMU (KMHMU' AM)

James KIRBY
University of Edinburgh
J.Kirby@ed.ac.uk

Abstract

This paper presents an acoustic analysis of fundamental frequency (F0) perturbations conditioned by voiceless and preglottalized nasals in the speech of 20 speakers of a phonologically conservative, non-tonal variety of Eastern Khmu (Kmhmu' Am). Broadly speaking, F0 of vowels following voiced nasals is similar to F0 following voiced plosives, but during the closure phase, F0 is much lower for the voiced obstruents than for the voiced sonorants. F0 following voiceless obstruents is initially perturbed upwards, but quickly converges to the intonational baseline. The effect of voiceless nasals on F0 is comparable to, or even greater than, that of voiceless obstruents. The effect of preglottalized nasals on F0 is similar that of the voiced nasals, but individual speakers vary considerably in this regard. These findings clearly illustrate the phonetic basis for the patterning of voiceless sonorants in tonogenesis and tone splits.

Keywords: Khmu, voiceless sonorants, preglottalized sonorants, F0, tonogenesis, phonetics

ISO 639-3 codes: kjg

1 Introduction

The term *sonorant* subsumes vowels, nasals, liquids, and glides, i.e., those sounds that are produced with a continuous, non-turbulent airflow in the vocal tract. Sonorants are typically voiced, with one study estimating phonologically contrastive voiceless sonorants occur in just 5% of the world's languages (Maddieson 1984a). At least within East and Southeast Asia, however, voiceless sonorants, or at least voiceless nasals, are more common: of the 61 languages with voiceless nasals in the PHOIBLE database (Moran & McCloy 2019), 23 are spoken in East or Southeast Asia. Voiceless nasals in Southeast Asia typically consist of two distinct phonetic phases: a period of voicelessness accompanied by nasal airflow (aspiration), followed by a short, sonorous voiced portion (Dantsuji 1984; Bhaskararao & Ladefoged 1991).

Sonorants may also be realized with an accompanying glottal constriction. When this constriction precedes the sonorous portion (as opposed to being coextensive with it), these may be referred to as *preglottalized* sonorants. According to surveys (Ruhlen 1975; Maddieson 1984a; Moran & McCloy 2019), they are similarly rare as phonologically contrastive segments; PHOIBLE lists just 11 instances.

The importance of voiceless (and to a lesser extent, preglottalized) sonorants in the processes of tonogenesis and tone splitting has been remarked on by numerous scholars (Haudricourt 1961; Matisoff 1973; Chen 1992; L-Thongkum 1992, 1997; Hyslop 2009; Pittayaporn & Kirby 2017; Michaud & Sands 2020). It is well established that it is almost always the historical voicing status of a segment, rather than whether it is an obstruent or sonorant, that predicts how onsets behave in tone splits. A classic example is the evolution of Sgaw Karen (Haudricourt 1961), in which two tones later split into four under the influence of the laryngeal specification of the initial consonant: modal voiced stops and voiced sonorants conditioned a low register, while preglottalized voiced stops, voiceless aspirated and unaspirated stops, and voiceless sonorants conditioned a high register. Slightly more complex is the example of Dong (Kra-Dai), where an earlier system of 3 tones (A, B, C) was split three ways, but here again, the behavior is clearly conditioned by onset voicing: preglottalized sonorants and plain voiceless stops conditioned high register tones, voiceless continuants and aspirated stops conditioned mid register tones, and plain voiced sonorants, plosives, and fricatives conditioned low register tones. A

contemporary example is that of Cao Bằng Tai, in which the historical contrast between voiced and voiceless sonorants has been lost (neutralized to the voiced versions), and it is only following sonorants that all six lexical tones are found (Hoàng Văn Ma 1997; Pittayaporn 2009; Pittayaporn & Kirby 2017). Many similar examples are documented in Haudricourt (1961).

Data like these are what led Pittayaporn (2009) to propose a model of tone splitting in which voiceless sonorants play a critical role. In Pittayaporn's model, there is a stage of categorical but redundant pitch registers based solely on voicing (Stage II), followed by the development of phonemic register in the sonorants when the voicing contrast is lost in the sonorant sub-system only (Stage III). A phonetic corollary of this model would seem to be that voiceless and preglottalized sonorants should have pitch-perturbing properties similar to those of the obstruents they pattern with in historical tone splits.

It is well established that phonologically voiceless obstruents tend to raise F₀ on the following vowel, while phonologically voiced obstruents either have no effect or, in some instances, can result in lowered F₀ (House & Fairbanks 1953; Lehiste & Peterson 1961; Hombert 1978; Kohler 1982; Ohde 1984; Kingston & Diehl 1994; Hanson 2009; Kirby & Ladd 2016; Coetzee et al. 2018; Kirby 2018; Gao & Arai 2019; Kirby et al. 2020). However, there is much less work on the pitch-perturbing properties of voiceless and preglottalized sonorants. What little phonetic work exists on voiceless nasals has focused almost exclusively on Tibeto-Burman languages, primarily Burmese (Dantsuji 1984; Maddieson 1984b; Bhaskararao & Ladefoged 1991; Chirkova, Basset & Amelot 2019). Those studies that have looked at F₀ (Dantsuji 1984; Maddieson 1984b) have found a dichotomy similar to that seen in plosives in other languages (i.e., F₀ is higher following voiceless sonorants compared to voiced sonorants), but these studies did not compare the effects of voiced and voiceless sonorants on F₀ with those of obstruents, only to one another. Moreover, since Burmese is a tonal language, the co-intrinsic F₀ effects may well be attenuated (Hombert 1978; Francis et al. 2006; Kirby 2018). What is ideally required is a *non-tonal* language which has the full complement of onset types: voiced, voiceless unaspirated, and voiceless aspirated plosives, along with voiced, voiceless, and preglottalized sonorants.

2 Eastern Khmu (Kmhmu' Am)

Luckily, there exists at least one such language: Kmhmu'. Kmhmu' (also Khmu, Kammu, etc., cf. Proschan 1997) is an Austroasiatic language with several distinct varieties, with a total of around 700,000 speakers primarily in Laos, Thailand, China and Vietnam. Kmhmu' is fairly well studied, due perhaps in part to the relative and sustained vitality of the language, but also because of its importance in understanding Austroasiatic more generally. Cheeseman et al. (2017) provide an extensive, although not exhaustive, bibliography.

Kmhmu' varieties are typically divided into two types, distinguished primarily in terms of lexicon and phonology. The most striking difference is the existence of one set of dialects, termed 'Southern' in the terminology of Lindell et al. (1980; 1981) and 'Eastern' in the terminology of Suwilai Preamsirat (1987; 1999; 2001; 2004), which retain a rich initial consonant inventory including voicing oppositions for both stops and sonorants. In the other main type ('Northern' for Lindell¹, 'Western' for Preamsirat), the laryngeal contrast of the initial consonants has been lost and restructured as a contrast of vowel phonation, tonality or a combination of the two (although no acoustic evidence of a variety incorporating phonation type contrasts, primarily or secondarily, is known to exist). The so-called 'tonal' varieties of Kmhmu', in which onset F₀ perturbations have been phonologized, have received more attention from phoneticians (Suwilai Preamsirat 1999; 2001; 2004; Svantesson & House 2006;

¹ Svantesson and colleagues (Svantesson 1983; 1989; Svantesson & House 2006) sometimes use the term 'Northern' to differentiate between allegedly tonal dialects which retain an aspiration contrast (e.g., *yùan*) versus those which apparently do not (e.g., *ròk*). In Svantesson et al.'s 'Western Khmu', the voiced stops of the Eastern varieties merged with the voiceless unaspirated series; in 'Northern Khmu', they merged with the aspirated series. In both cases, however, the old voiced series is distinguished by a lower vowel pitch (with supposedly redundant aspiration in the Western varieties).

Abramson, Nye & Luangthongkum 2007); however, the acoustic properties of the voiceless and glottalized sonorants, retained in the more conservative varieties, have not been carefully studied.

This work focuses on a phonologically conservative Southern/Eastern variety spoken in and around Vientiane, Luang Prabang, Xiang Khouang, and Bolikhamsay provinces in Laos, referred to by our consultants as *Kmhmu' Am* /kəm̥uʔ ʔam/ (/ʔam/ being the negative particle shibboleth). In other sources, this variety is sometimes known as *Kmhmu' Uu* or *Kmhmu' Cwang*. Osborne (2018) provides a thorough and detailed overview of the phonology of this variety, which features 36 consonants /p^h p b t^h t d c^h c ɟ k^h k g ʔ m̥ m̥ m̥ n̥ n̥ n̥ n̥ ŋ^h ŋ^h ŋ^h w̥ w̥ w̥ j̥ j̥ j̥ s h l l̥ r̥ r̥/, of which only 15 /p t c k ʔ m n ŋ w j̥ j̥ h l r̥/ occur as codas; 10 monophthongs /i e ε a ə ə i u o ə/, all of which contrast for length; and 3 diphthongs /ia ia ua/.

The conservative phonology of *Kmhmu' Am* provides a natural laboratory in which to study how the F0 differences between the voiced, voiceless, and preglottalized sonorants compare to the magnitude of the differences between voiced and voiceless obstruents. Eastern Khmu thus affords us a window into the phonetic structures of a language that may reflect the ancestral state of many of the tonal and registral languages of modern Southeast Asia.

3 Methods

The current study compares the F0 profiles of voiceless and preglottalized sonorants with those of voiced sonorants and voiced and voiceless obstruents. Although the primary focus is on the sonorants, it is important to provide an analysis of the obstruents to understand whether, and how, the different segment types affect F0 in different ways. We expect to see lowering of F0 during the closure for voiced obstruents, but tracking the trajectory of voiced sonorants thereafter, as well as raising of F0 following the release of voiceless (aspirated and unaspirated) obstruents, relative to the voiced sonorant baseline. Based on their typical diachronic patterning, we expect both voiceless and preglottalized sonorants will raise F0 in a manner similar to that of voiceless obstruents.

3.1 Participants

This study is based on recordings of 25 speakers (14 female and 11 male, ages 21-69) made in a *Kmhmu'* village in Vientiane in January 2020. As 4 of our consultants were primarily speakers of a slightly different variety (*Kmhmu' Pee*), and the recordings of one older male speaker were of insufficient quality, the findings reported here are based on a subset 20 speakers (12 females and 8 males). All consultants were also fluent in Lao to varying degrees, but all were native speakers of *Kmhmu'* and spoke *Kmhmu'* daily as their primary language.

3.2 Procedure

Speakers were recorded reading a list of 125 words, which they produced four times: twice in isolation and twice in a carrier phrase /ʔoʔ cə law _____ ʔan kləh/ (1sg IRR speak _____ SBJV clearly) ‘‘I will say _____ clearly’’. Participants produced the *Kmhmu'* form in response to an oral prompt of the Lao gloss by an experimenter; some participants who were literate in Lao were able to read the glosses themselves. Prior to recording, *Kmhmu'* assistants went over the Lao glosses with each participant, so they were comfortable with the procedure, and familiar with the *Kmhmu'* lexical items of interest. Recordings were made direct to disk using the SpeechRecorder software (Draxler & Jansch 2004) with a headset condenser microphone in a quiet, sound-treated booth. A simultaneous EGG signal was also recorded from most speakers, and used to assist in the segmentation, but is not analyzed here.

The present paper focuses on a subset of the full list, consisting of 59 items with long vowel nuclei and one of 15 onsets (see Appendix): voiceless plosives /p t k/, (pre)voiced plosives /b d g/, aspirated plosives /p^h t^h k^h/, voiced nasals /m n/, voiceless nasals /m̥ n̥/ and preglottalized nasals /ʔm̥ ʔn̥/. We also recorded examples of voiceless laterals and approximants, and of the velar nasal /ŋ/, but the lexical items with these onsets have either short vowels or, in the case of the lateral, we could not find a suitable lexical item with the corresponding voiced onset for comparison.

3.3 Segmentation

Target syllables were manually segmented and stored as an EMU speech database (Winkelmann, Harrington & Jänsch 2017). Annotations were made on two tiers. The first tier was used to indicate the presence or absence of a constriction in the supraglottal vocal tract. In the carrier phrase contexts, this was straightforward: for plosives, the closure phase was the period of silence preceding the release burst; for sonorants, this was either the sonorous nasal portion (for voiced nasals) or a period containing both silence/frication noise followed by a periodic nasal portion (for voiceless and preglottalized nasals). For utterances produced in isolation, no closure phase could be annotated for the voiceless plosives; here, the visible duration of prevoicing was taken as an indicator of the closure phase. Similarly, for preglottalized nasals in isolation context, the beginning of periodic vibration was taken as the onset of the closure as measured (Figure 1). For voiceless nasals in isolation context, the beginning of the closure phase was deemed to be the onset of visible high-frequency frication in the spectrogram (Figure 2). In all cases, the onset of the open phase was assessed as either the plosive release burst, if present, or the onset of periodic formant structure with a clear second formant.

The second tier was used to indicate the onset of periodic vocal fold vibration (for plosives) or the onset of periodic nasal murmur (for nasals). Particularly for /m̥ n̥/ in carrier phrase contexts, vocal fold vibration would often be visible throughout the closure phase, but this was almost always acoustically distinct from a ‘true’ nasalized portion, identifiable by increased waveform amplitude and the presence of faint formant structure (Figure 3).

For voiced plosives, voicing would sometimes be present during the closure, but either die off prior to the release burst, or be cut off by the burst, and there could then follow a brief period of voicelessness before the onset of the vowel. In these instances, three points were annotated: voicing onset during the closure phase; voicing offset during the closure phase; and voicing onset during the open phase (Figure 4). This permitted measurement of both the duration of prevoicing and the post-release voicing lag time within the same syllable.

Sonorant codas were segmented when present, but in the following analysis, vowel and sonorant codas are treated as a single unit, since exploratory data analysis suggested that F0 excursions on syllables like /da:ŋ/ were timed similarly to those on syllables like /da:/.

3.4 Analysis

After segmentation, F0 in the target items was measured at 5 msec intervals using the `ksvf0` estimator in the `wrassp` package (Bombien, Winkelmann & Scheffers 2021). Raw Hertz values were transformed to semitones for each speaker by scaling each value x by $12 \log_2 x/\mu_s$, where μ_s is that speaker’s mean F0 value. These were used to estimate the mean F0 for the 50 milliseconds preceding the closure release (for segments with at least some closure voicing) as well as the first 50 milliseconds immediately following the release.

Figure 7: Example of segmentation of a preglottalized nasal in /ʔmɔ:n/ *ገጠጠ* ‘place, area’, citation form context, speaker F13.

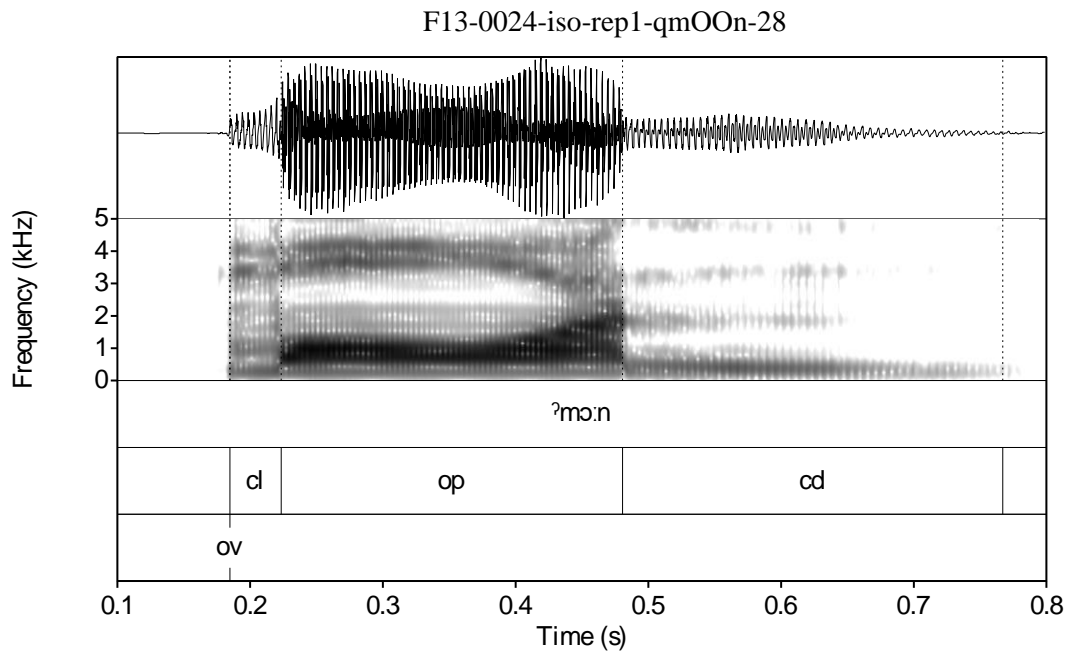


Figure 8: Example of segmentation of a voiceless nasal in /m̥a:n/ *ፊጎ* ‘to bury’, citation form context, speaker F12.

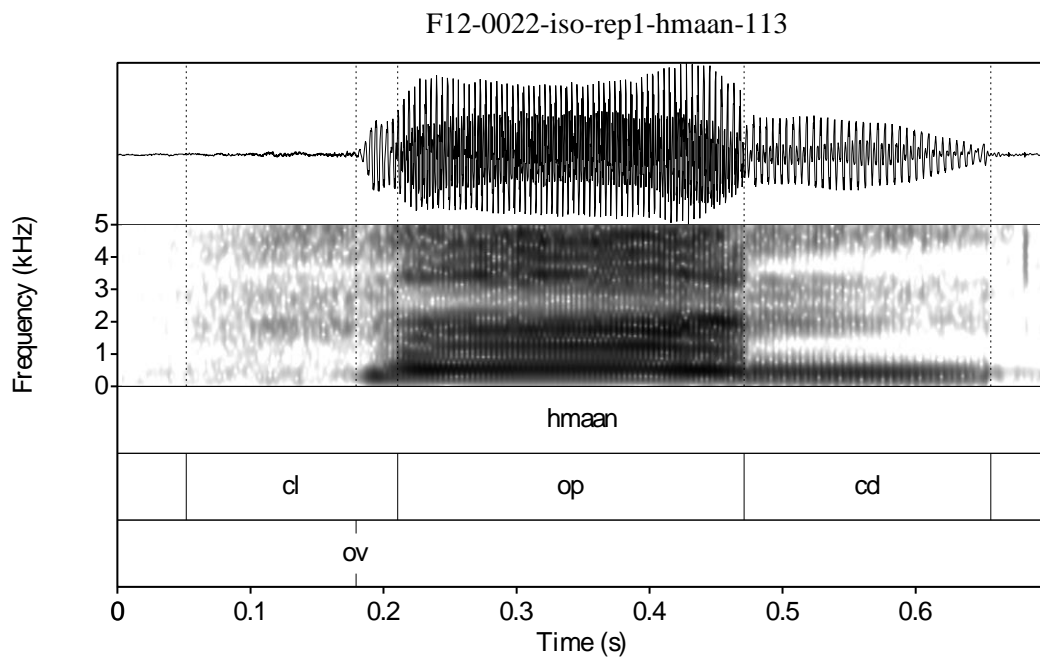


Figure 9: Example of segmentation of a voiceless nasal in /ŋɔːŋ/ ၵ်ႉ ‘still, yet’, carrier phrase context, speaker F13.

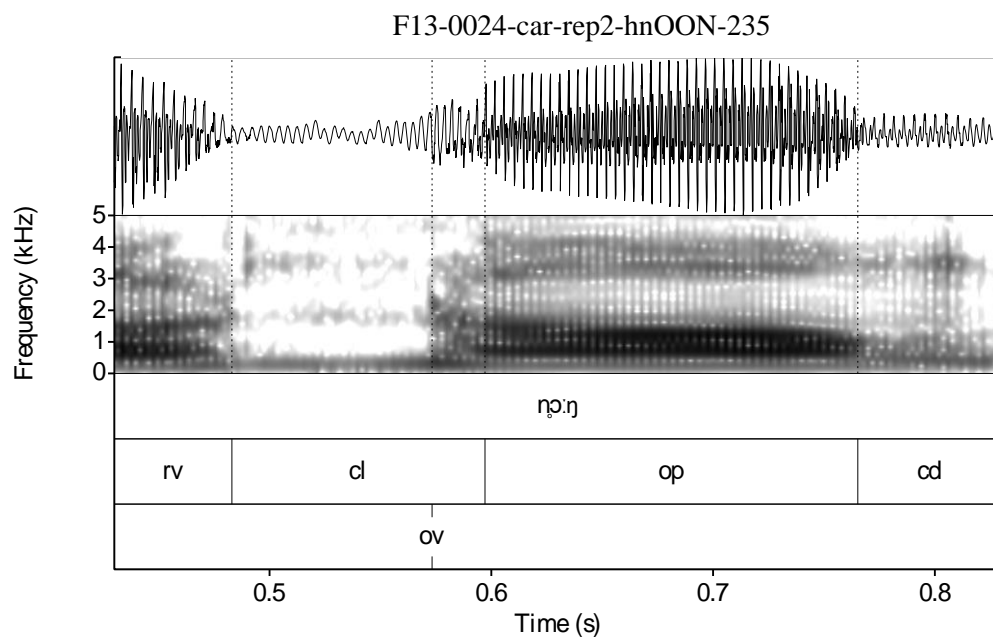
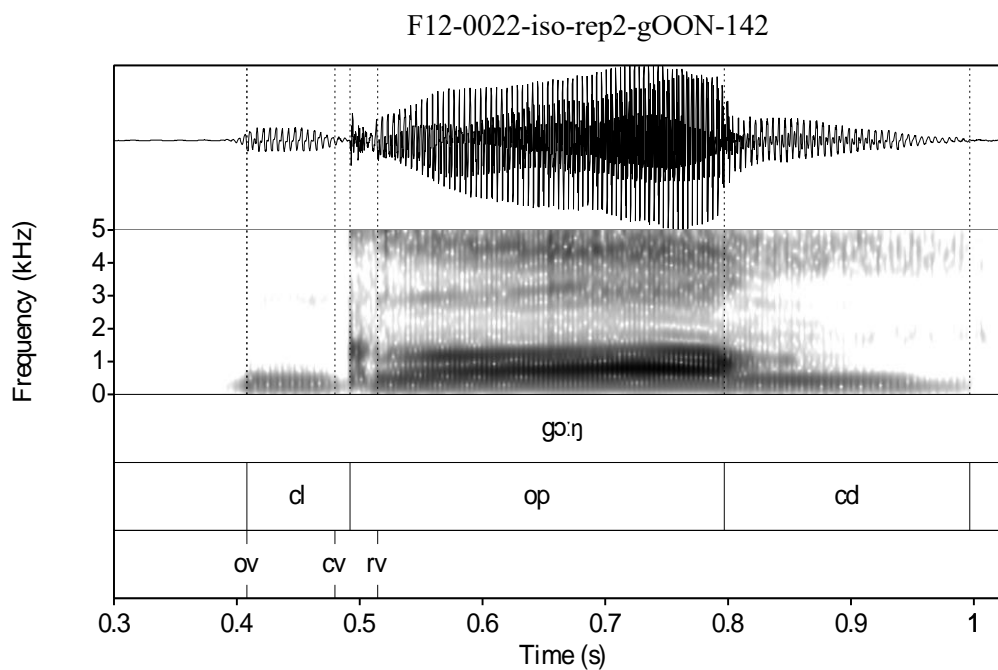


Figure 10: Example of segmentation of a voiced plosive in /gɔːŋ/ ငၢ်ႉ ‘soup’, citation form context, speaker F12.



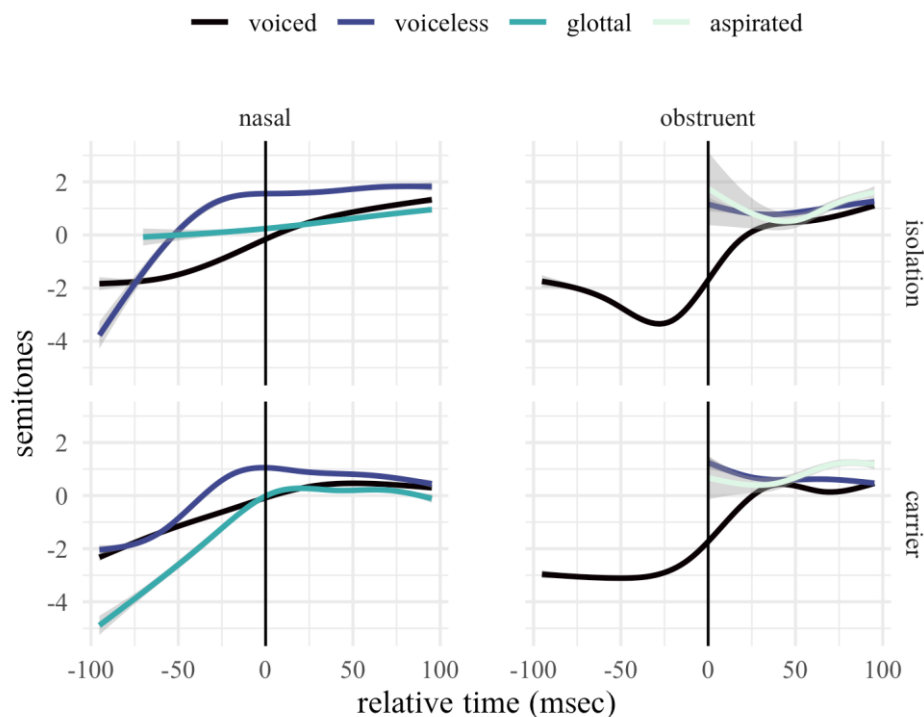
4 Results

4.1 General patterns

The smoothed conditional mean F0 trajectories² (in semitones) for the 100 milliseconds preceding the closure and the 100 milliseconds following are shown in Figure 5, faceted by context (isolation versus carrier) and manner (nasal versus obstruent). Trajectories are plotted in relative time, rather than normalized time, because the normalized time comparison would mask the considerable durational differences between the sonorous portions of the voiceless and preglottalized sonorants compared to the voiced nasals. Estimates of F0 (in semitones) averaged over the last 50 milliseconds of the closure and the first 50 milliseconds following closure release, where the differences are most pronounced, are given in Tables 1 and 2.

Several major trends are apparent from these figures. First, there is considerable lowering of F0 during the closure phase of the voiced obstruents, relative to voiced nasals (see also Figure 6). Second, F0 is raised in voiceless nasals both during the nasalized closure portion as well as during the following vowel. This difference persists for at least the first 100 msec in isolation forms; in the carrier phrase context, the effect is clearly attenuated, but there is still a mean difference of around 1 semitone compared to the voiced and preglottalized nasals in the region around the closure release. As highlighted in Figure 6, in the citation forms, F0 following voiceless nasals is slightly higher than following voiceless obstruents, but this effect disappears in the carrier phrase context.

Figure 11: Mean F0 trajectories (GAM smooths) over last 100 msec of closure and first 100 msec post-release for nasals and obstruents by utterance context.



² F0 trajectories were smoothed by fitting a generalized additive model (GAM) using a cubic spline basis with shrinkage and 6 evenly spread knots; see Wood (2006) for further details.

Figure 12: Mean F0 trajectories (GAM smooths) over last 100 msec of closure and first 100 msec post-release for voiced plosives vs. voice nasals (left column) and voiceless plosives vs. voiceless nasals (right column) by utterance context.

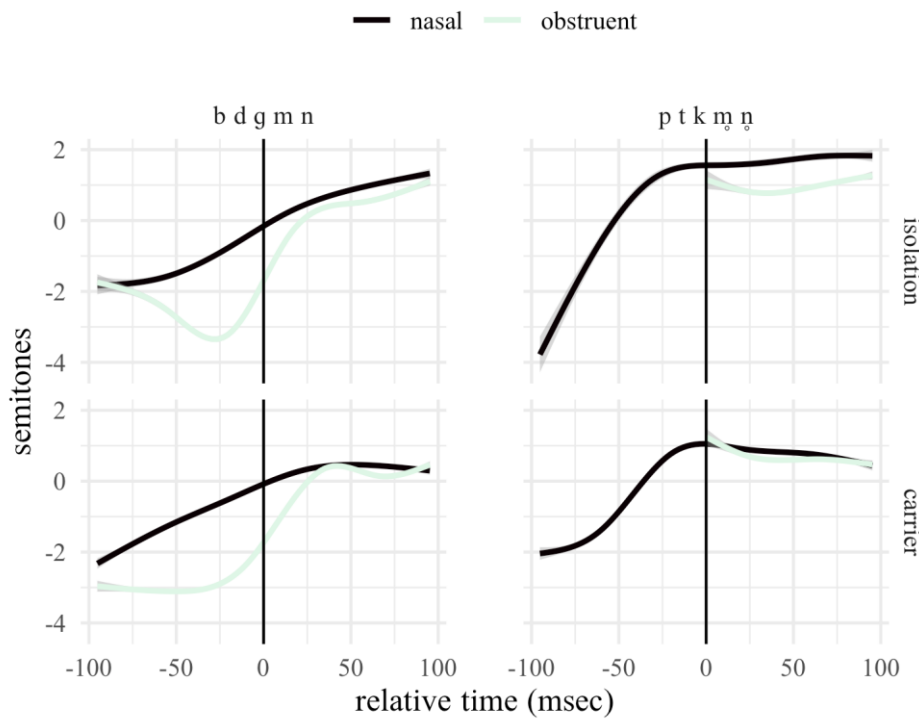


Table 7: Mean and standard deviation of F0 (in semitones) averaged over last 50 milliseconds preceding closure release.

		<i>isolation</i>		<i>carrier phrase</i>	
place	onset	F0 (mean)	F0 (SD)	F0 (mean)	F0 (SD)
<i>bilabial</i>	b	-2.64	2.05	-2.78	1.72
	m̥	0.92	1.64	0.35	1.60
	m	-1.05	1.69	-0.73	1.27
	ʹm	0.23	1.46	-0.72	1.32
<i>coronal</i>	d	-3.24	2.20	-3.12	1.65
	n̥	1.56	1.79	0.55	1.76
	n	-0.59	1.75	-0.47	1.46
<i>velar</i>		0.02	1.50	-0.76	1.76
	g	-3.19	2.48	-2.89	1.84

Table 8: Mean and standard deviation of F0 (in semitones) averaged over first 50 milliseconds following closure release.

place	onset	<i>isolation</i>		<i>carrier phrase</i>	
		F0 (mean)	F0 (SD)	F0 (mean)	F0 (SD)
<i>bilabial</i>	b	0.62	2.02	0.24	1.67
	m̥	1.04	1.28	0.30	1.16
	m	0.28	1.71	0.21	1.47
	p	0.88	2.06	0.80	1.49
	p ^h	-0.04	2.81	0.05	1.08
	ʔm	0.66	1.23	0.49	1.10
<i>coronal</i>	d	-0.20	2.01	-0.30	1.62
	n̥	1.73	1.64	1.04	1.27
	n	0.60	1.79	0.38	1.64
	ʔn	0.11	1.12	-0.15	1.28
	t	1.16	2.26	0.93	1.74
	t ^h	-0.12	2.94	-0.06	2.05
<i>velar</i>	g	-0.29	2.09	-0.16	1.50
	k	0.05	1.80	0.01	1.35
	k ^h	1.39	0.76	0.98	1.57

4.2 Individual differences

While space does not permit a full exploration of the individual patterns, a few examples should make it clear that not all speakers are homogeneous in terms of how their voiceless and preglottalized nasals perturb the F0 trajectory.

For most ($n=15$) speakers, preglottalized sonorants do not generally raise pitch during the closure or following it, relative to /m n/. For some speakers, such as M3 (Figure 7), pitch following /ʔm ʔn/ is lower than that of the corresponding voiced nasals, at least in carrier phrase context; female speaker F7 (Figure 8) shows a similar pattern. However, for the remaining 5 speakers, exemplified here by M2 and F4 (Figures 9 and 10), both voiceless and preglottalized nasals condition higher F0 compared to voiced nasals, in isolation and in the carrier phrase, both in the (short) sonorous period during the closure, but in M2's case extending well into the following vowel.

The existence of these two types of speakers appears to be what gives the impression of the preglottalized sonorants being “intermediate” in terms of their effects on F0 when we look at the group average in Figure 6. But actually, it seems there are two main groups of speakers – those for whom preglottalized sonorants pattern with the voiceless sonorants (and obstruents), and those for whom they pattern with the voiced sonorants (and obstruents). Thus, the claim that “voiceless unaspirated stops [in Eastern Kmhmu’] are phonetically stiff voiced, bringing them in-line with the glottalised sonorants” (Osborne 2018:71) may bear further scrutiny.

Figure 7: Smoothed mean F0 trajectories for speaker M3 (male, age 34).

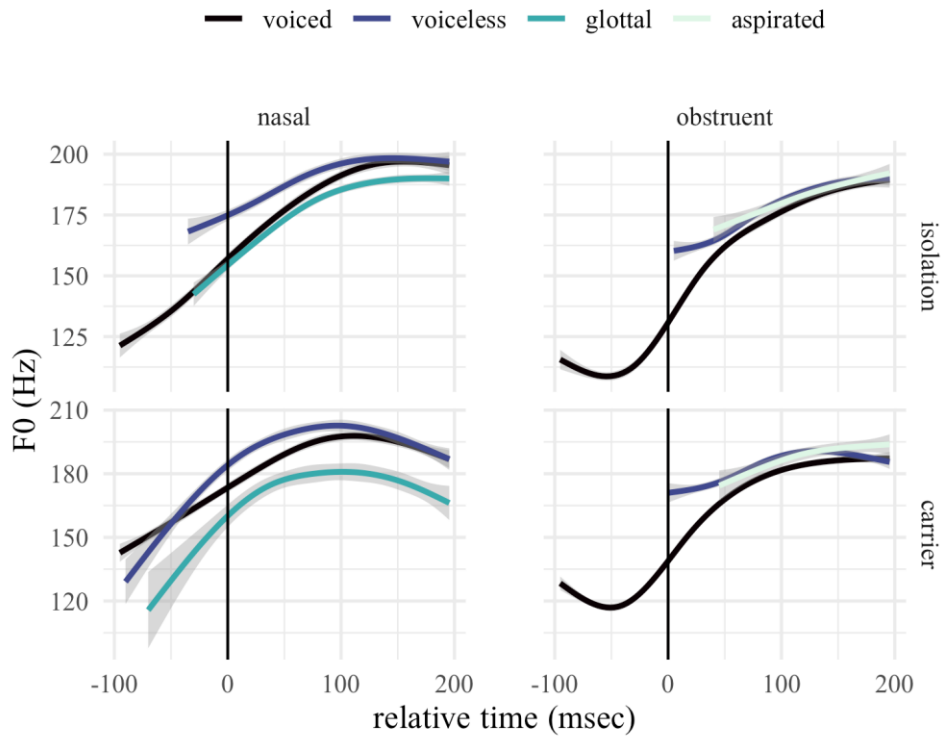


Figure 8: Smoothed mean F0 trajectories for speaker F7 (female, age 32).

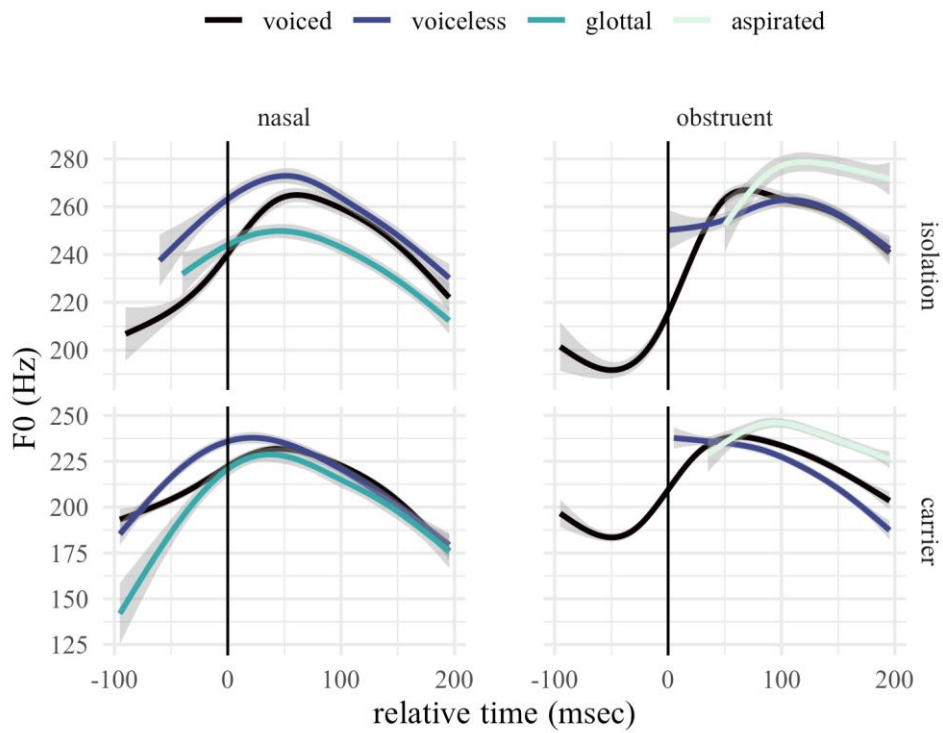


Figure 9: Smoothed mean F0 trajectories for speaker M2 (male, age 38).

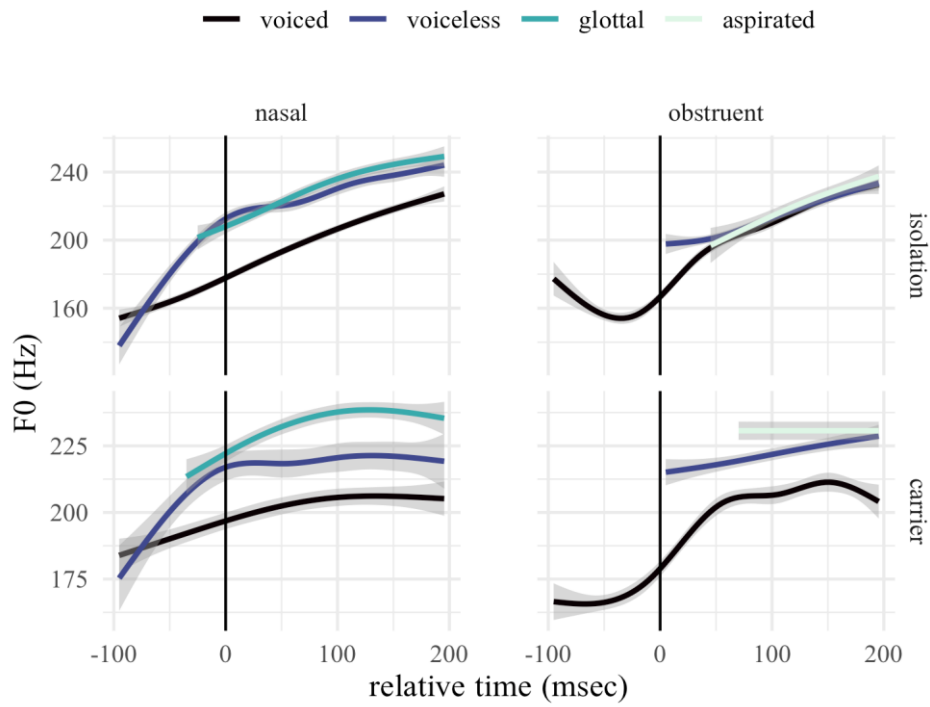
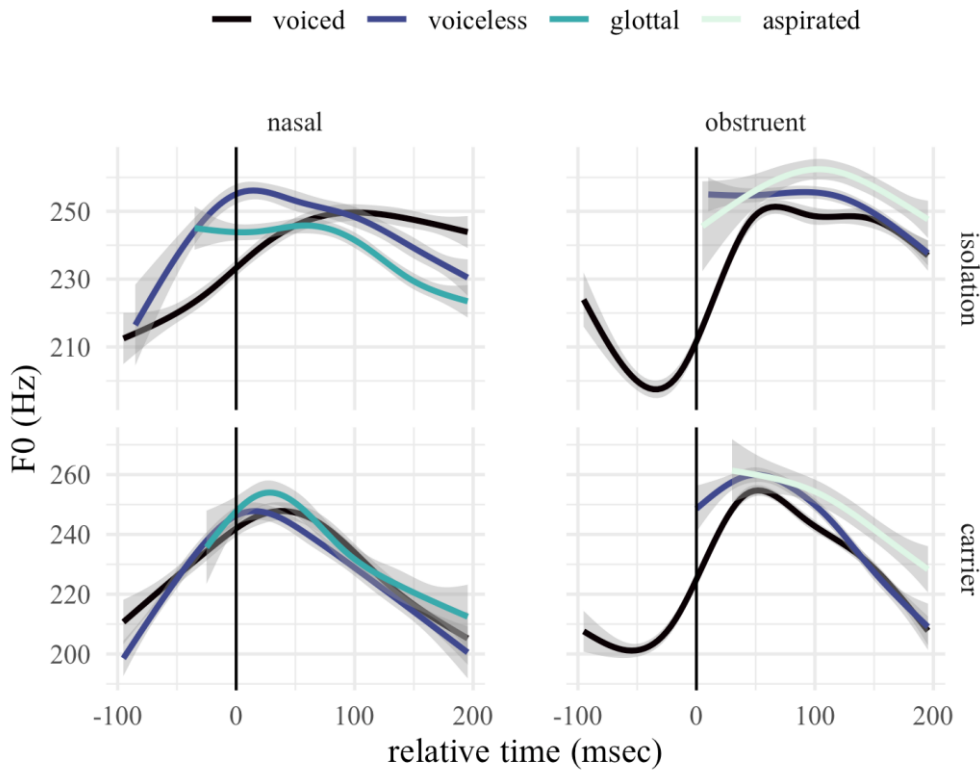


Figure 10: Smoothed mean F0 trajectories for speaker F4 (female, age 37).



5 Discussion

In this sample of speech from 20 speakers of Eastern Khmu (Khmhu' Am), voiceless nasals condition F0 raising relative to voiced nasals, both during the oral closure as well as following the release. This finding is consistent with earlier work on Burmese (Dantsuji 1984; Maddieson 1984b). In addition, by comparing trajectories (rather than just point estimates) and by analyzing obstruents as well as sonorants, the present work shows that the magnitude of the F0 perturbation is at least on par with that conditioned by voiceless plosives, and in general, the temporal extent is greater, at least in isolation forms.

The behavior of the preglottalized nasals with respect to F0 was found to be more variable. For some speakers, F0 seems to be raised following these segments, but for many, preglottalized nasals appear to pattern with voiced nasals in terms of their (lack of) effects on F0. It is worth mentioning that the preglottalized nasals in Eastern Khmu varieties correspond to “voiced, slightly implosive” stops in other Khmu varieties (Svantesson & Holmer 2014:963), and implosives are known to exhibit variable patterning with respect to tone both synchronically (Tang 2008) and diachronically (Haudricourt 1961; Gedney 1972; Hombert 1978).

The present findings provide phonetic support for the proposal that tone splits might well begin with voiceless sonorants, at least in languages which have them. The effect of this class of segments on F0 is at least as great as that of voiceless obstruents, and because there is no voice break between the (admittedly short) nasal and the following vowel, the F0 contour is potentially even more audible compared to voiceless obstruents. Coupled with the fact that, in carrier phrase contexts, the aspiration of the voiceless nasal is likely to be difficult to perceive, this seems to provide the perfect conditions for the actuation of sound change (Ohala 1993; Janda & Joseph 2003).

An important question to address in future work is to see whether the magnitude of the F0 perturbation varies with the duration of the voiceless nasal. If F0 is high even when there is voicing throughout the closure, this would suggest that the primary difference is one of laryngeal tension setting, not of the perceptually rather fragile presence versus absence of nasal frication. Indeed, it may well be the presence of a shared articulatory posture, rather than any particular acoustic effect, which drives voiceless (and in some cases, preglottalized) sounds to pattern together in tone splits.

Also interesting is the extent to which F0 is lowered *during* the closure for voiced plosives. This effect was very strong in the data examined here, suggesting a phonetic basis whereby prevoiced plosives might diverge from other segment types. Yet historically, voiced plosives usually pattern with voiced nasals in terms of tone splits. This patterning may be related to the fact that prevoiced plosives are often partially or completely devoiced, as seen in e.g., Chru (Brunelle et al. 2020) or European Portuguese (Pape & Jesus 2015), and when this happens, their post-release F0 profiles are often indistinguishable from those of voiced nasals. The devoicing rate is one of several other acoustic properties, such as spectral balance, which bear on this question, and which should be examined in future work.

Acknowledgments

This study was made possible by funding from the Arts and Humanities Research Council (AH/P014879/1) and the European Research Council (ERC StG 758605). Many thanks to Ryan Gehrmann for planning and assistance in the field, and for helpful comments on an earlier version of this manuscript; I remain solely responsible for any errors of fact or interpretation. Thanks also to Sue-Anne Richer for assistance with data annotation, and to Mark and Anne Osborne for warm hospitality and essential logistic support in Vientiane. Last but certainly not least, a hearty ຂອບໃຈ to all of the Khmhu' consultants who traveled from near and far to enthusiastically take part in this study.

Appendix: Wordlist analyzed in present study

IPA	Lao	Gloss
ba:	ບ້າ (ບ້າຍິງ)	you (female)
ba:r	ສອງ	two
bi:t	ມອດ(ໄຟ)	to put out fire, extinguish

bu:	ໂພງ, ພອງ. (ແກ້ມ) ບຸ່ງ	puffy; swollen
bu:c	ເຫຼົ້າສ່າ	liquor
de:r	ຄວ່າງ, ຫວ່າງ (ແຫ)	spread out (fishing net)
do:m	ອອນຊອນ	beautiful, sweet, natural (sound)
da:	ທາ	to apply, paint
da:n	ໂຕລີລັງມ/ໂຕລີກງມ	lizard
da:l	ບໍ່ຄົມ	dull
(hn)du:m	ສຸກ	ripe
go:n	ແກງ	soup
ga:	ປີນ	climb
ga:n	ເຮືອນ	house
ge:t	ເທ	to pour (water)
gi:	ນີ້	here, this
gu:m	ຜັດເຂົ້າ	to winnow the paddy
ma:n	ຜັງ	to bury
ma:r	ເກືອ	salt
ne:n	ແຂງ, ແໜ້ນ	hard, tight
no:n	ຍັງ	still, yet, remain
na:j	ນັ້ນ	that (dem.)
ni:	ຫົນ	debt
ke:n	ມັດຕິດແອວ	tighten at the waist
ko:k	ບັງ	caterpillar
ko:l	ບັກ	to cut down
ko:n	ລູກ	offspring, child
ka:l	ກ່ອນ	before
ka:p	ຄາງ	chin
ku:p	ຫໍ່ (ຜັບ)	to cook food by wrapping in a banana leaf and roasting in the fire
khə:l	ຜິວປາກ	whistle
khə:l	ເສັ້ນຕອກ	thin bamboo strip used for making baskets
khɪ:	ທີ່ນີ້	here (loc. Adv.)
khɪ:l	ຂົນ	body hair
ma:m	ເລືອດ	blood
mo:j	ຫົ່ງ	one (numeral)
mu:m	ອາບນ້ຳ	to take a bath
mə:j	ນ້ຳມັນ	grease, oil
(hr)no:m	ຕອກ	thin bamboo strip used for tying things
na:	ລາວ (ແມ່ຍິງ)	she (pronoun); rice paddy?
nu:m	ຢຽວ	urine
pe:l	ຈັບເບິ່ງ	catch and watch
po:k	ຈີ່	burn, grill over fire
pa:l	ບານ	birthmark
pu:	ລົບ (ແກ່ນລົບ)	empty rice husk
pu:c	ແກ້ໄສ້ງອອກ, ເສື້ອ	to take off clothes
pu:l	ສີ່	four
po:j	ວີ	fan, blow

p ^h a:n	ຂ້າ	to kill
ʔmɔ:k	ບອກ, ສອນ	to tell, teach
ʔmɔ:n	ສະຖານທ	place, area
ʔna:ŋ	ດ່າງ	clf. for nets
ʔna:m	ເທົ່າກັນ	same size
ta:j	ອ້າຍ	older sibling
ta:k	ຄາຍອອກ	to spit out
ti:ŋ	ລົ້ມ (ໄມ້)	to fall down (house, tree)
tu:	ໃສ່ຮ້າຍປ້າຍສີ	to falsely accuse
tu:t	ຕົ້ນ, ກົກ	plant
t ^h a:k	ບອກເປືອກຕົ້ນໄມ້	to strip bark off of tree

References

- Abramson, Arthur S., Patrick W. Nye & Theraphan Luangthongkum. 2007. Voice register in Khmu': experiments in production and perception. *Phonetica* 64.2–3:80–104. DOI: <https://doi.org/10.1159/000107911>
- Bhaskararao, P. & Peter Ladefoged. 1991. Two types of voiceless nasals. *Journal of the International Phonetic Association* 21.2:80–88. DOI: <https://doi.org/10.1017/S0025100300004424>.
- Bombien, Lasse, Raphael Winkelmann & Michel Scheffers. 2021. wrassp: an R wrapper to the ASSP Library (Version 1.0.1).
- Brunelle, Marc, Tạ Thành Tấn, James Kirby & Đinh Lư Giang. 2020. Transphonologization of voicing in Chru: Studies in production and perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11.1:15. DOI: <https://doi.org/10.5334/labphon.278>
- Cheeseman, Nathaniel, Paul Sidwell & R. Anne Osborne. 2017. Khmuic linguistic bibliography with selected annotations. *Journal of the Southeast Asian Linguistics Society* 10.1:i–xlvi. DOI: <http://hdl.handle.net/10524/52401>
- Chen, Matthew Y. 1992. Competing sound changes: evidence from Kam-Tai, Miao-Yao and Tibeto-Burman. *Journal of Chinese Linguistics* 20.2:193–210.
- Chirkova, Katia, Patricia Basset & Angélique Amelot. 2019. Voiceless nasal sounds in three Tibeto-Burman languages. *Journal of the International Phonetic Association* 49.1:1–32. DOI: <https://doi.org/10.1017/S0025100317000615>
- Coetzee, Andries W., Patrice Speeter Beddor, Kerby Shedden, Will Styler & Daan Wissing. 2018. Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics* 66:185–216. DOI: <https://doi.org/10.1016/j.wocn.2017.09.009>
- Dantsuji, Masatake. 1984. A study on voiceless nasals in Burmese. *Linguistics* 7.6:59–81.
- Draxler, Christoph & Klaus Jänsch. 2004. SpeechRecorder - a universal platform independent multi-channel audio recording software. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 559–562. Lisbon.
- Francis, Alexander L., Valter Ciocca, Virginia Ka Man Wong & Jess Ka Lam Chan. 2006. Is fundamental frequency a cue to aspiration in initial stops? *The Journal of the Acoustical Society of America* 120.5:2884–2895. DOI: <https://doi.org/10.1121/1.2346131>
- Gao, Jiayin & Takayuki Arai. 2019. Plosive (de-)voicing and f₀ perturbations in Tokyo Japanese: Positional variation, cue enhancement, and contrast recovery. *Journal of Phonetics* 77:100932. DOI: <https://doi.org/10.1016/j.wocn.2019.100932>
- Gedney, William J. 1972. A checklist for determining tones in Tai dialects. In S. M. Estellie (Ed.), *Studies in Linguistics in honor of George L. Trager*, 423–438. The Hague: Mouton.

- Hanson, Helen M. 2009. Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America* 125.1:425–441. DOI: <https://doi.org/10.1121/1.3021306>
- Haudricourt, André-G. 1961. Bipartition et tripartition dans les systèmes de tons. *Bulletin de La Société de Linguistique de Paris* 56.1:163–180.
- Hoàng Văn Ma. 1997. The sound system of the Tày language of Cao Bằng province, Vietnam. In Jerold A. Edmondson & David B. Solnit (eds.), *Comparative Kadai: the Tai branch*, 221–231. Dallas, TX: Summer Institute of Linguistics.
- Hombert, Jean-Marie. 1978. Consonant types, vowel quality, and tone. In Victoria Fromkin (ed.), *Tone: a linguistic survey*, 77–111. San Diego: Academic Press.
- House, Arthur S. & Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustic characteristics of vowels. *The Journal of the Acoustical Society of America* 25.1:105–113.
- Hyslop, Gwendolyn. 2009. Kurtöp tone: A tonogenetic case study. *Lingua* 119.6:827–845. DOI: <https://doi.org/10.1016/j.lingua.2007.11.012>
- Janda, Richard D. & Brian D. Joseph. 2003. Reconsidering the Canons of Sound-Change: Towards a “Big Bang” Theory. In Barry Blake & Kate Burridge (eds.), *Historical Linguistics 2001: Selected Papers from the 15th International Conference on Historical Linguistics, Melbourne*, 205–219. Amsterdam: John Benjamins Publishing Company.
- Kingston, John & Randy L. Diehl. 1994. Phonetic knowledge. *Language* 70.3:419–454. DOI: <https://doi.org/10.2307/416481>
- Kirby, James. 2018. Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages. *Journal of Phonetics* 71:326–354. DOI: <https://doi.org/10.1016/j.wocn.2018.09.009>
- Kirby, James & D. Robert Ladd. 2016. Effects of obstruent voicing on vowel F0: evidence from ‘true voicing’ languages. *The Journal of the Acoustical Society of America* 140.4:2400–2411.
- Kirby, James, D. Robert Ladd, Jiayin Gao & Zuzana Elliott. 2020. Elicitation context does not drive F0 lowering following voiced stops: Evidence from French and Italian. *The Journal of the Acoustical Society of America* 148.2:EL147–EL152. DOI: <https://doi.org/10.1121/10.0001698>
- Kohler, Klaus J. 1982. F0 in the production of fortis and lenis plosives. *Phonetica* 39:199–218.
- Lehiste, Ilse & Gordon E. Peterson. 1961. Some basic considerations in the analysis of intonation. *The Journal of the Acoustical Society of America* 33.4:419–425.
- Lindell, Kristina, Jan-Olof Svantesson & Damrong Tayanin. 1980. Phonology of Kammu dialects. *Cahiers de Linguistique Asie Orientale* 7.1:59–67. DOI: <https://doi.org/10.3406/clao.1980.1073>
- Lindell, Kristina, Jan-Olof Svantesson & Damrong Tayanin. 1981. Phonology of Kammu dialects. *Cahiers de Linguistique Asie Orientale* 9.1:45–71. DOI: <https://doi.org/10.3406/clao.1981.1089>
- L-Thongkum, Theraphan. 1992. The raising and lowering of pitch caused by a voicing distinction in sonorants (nasals and approximants): an epidemic disease in SEA languages. In *The Third International Symposium on Language and Linguistics*, 1079–1088. Bangkok: Chulalongkorn University.
- L-Thongkum, Theraphan. 1997. Implications of the retention of proto-voiced plosives and fricatives in the Dai Tho language of Yunnan province for a theory of tonal development and tai language classification. In Jerold A. Edmondson & David B. Solnit (eds.), *Comparative Kadai: the Tai branch*, 191–219. Dallas, TX: Summer Institute of Linguistics.
- Maddieson, Ian. 1984a. *Patterns of sounds*. Cambridge: Cambridge University Press.
- Maddieson, Ian. 1984b. The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. *Journal of Phonetics* 12:9–15.
- Matisoff, James A. 1973. Tonogenesis in Southeast Asia. In Larry M. Hyman (ed.), *Consonant types and tone*, 73–95. Los Angeles: University of Southern California.

- Michaud, Alexis & Bonny Sands. 2020. Tonogenesis. In Mark Aronoff (ed.), *The Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. Retrieved from <https://doi.org/10.1093/acrefore/9780199384655.013.748>
- Moran, Steven & Daniel McCloy (eds.). 2019. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. Retrieved from <http://phoible.org>
- Ohala, John J. 1993. The phonetics of sound change. In Charles Jones (ed.), *Historical linguistics: Problems and perspectives*, 237–278. London: Longman.
- Ohde, Ralph N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America* 76.1:224–230.
- Osborne, R. Anne. 2018. A phonology of Eastern Kmhmu' with special reference to palatal continuant codas and neutralisation of vowel length contrast. *Journal of the Southeast Asian Linguistics Society* 11.2: 67–86.
- Pape, Daniel & Luis MT Jesus. 2015. Stop and fricative devoicing in European Portuguese, Italian and German. *Language and Speech* 58.2:224–246. DOI: <https://doi.org/10.1177/0023830914530604>
- Pittayaporn, Pittayawat. 2009. *The phonology of proto-Tai*. Cornell University PhD dissertation. Retrieved from <http://ecommons.cornell.edu/handle/1813/13855>
- Pittayaporn, Pittayawat & James Kirby. 2017. Laryngeal contrasts in the Tai dialect of Cao Bằng. *Journal of the International Phonetic Association* 47.1:65–85. DOI: <https://doi.org/10.1017/S0025100316000293>
- Proschan, Frank. 1997. 'We Are All Kmhmu, Just the Same': Ethnonyms, Ethnic Identities, and Ethnic Groups. *American Ethnologist* 24.1:91–113.
- Ruhlen, Merrit. 1975. *A guide to the languages of the world*. Stanford.
- Suwilai Premsrirat. 1987. A Khmu grammar. In Suwilai Premsrirat (ed.), *Papers in South-East Asian Linguistics No. 10: Khmu, a minority language of Thailand*, 1–144. The Australian National University: Pacific Linguistics.
- Suwilai Premsrirat. 1999. Phonological variation and change in the Khmu dialects of northern Thailand. *Mon-Khmer Studies* 29:57–69.
- Suwilai Premsrirat. 2001. Tonogenesis in Khmu dialects of Southeast Asia. In Graham Thurgood (ed.), *Papers from the Ninth Annual Meeting of the Southeast Asian Linguistics Society*, 121–134. Arizona State University, Program for Southeast Asian Studies.
- Suwilai Premsrirat. 2004. Register complex and tonogenesis in Khmu dialects. *Mon-Khmer Studies* 34:1–17.
- Svantesson, Jan-Olof. 1983. *Kammu phonology and morphology*. Malmö: CWK Gleerup.
- Svantesson, Jan-Olof. 1989. Tonogenetic mechanisms in Northern Mon-Khmer. *Phonetica* 46:60–79.
- Svantesson, Jan-Olof & David House. 2006. Tone production, tone perception and Kammu tonogenesis. *Phonology* 23.2:309–333. DOI: <https://doi.org/10.1017/S0952675706000923>
- Svantesson, Jan-Olof & Arthur Holmer. Kammu. In Mathias Jenny & Paul Sidwell (eds.), *The handbook of Austroasiatic languages, Vol. 1*, 957–1002. Leiden: Brill.
- Tang, Elizabeth K. 2008. *The phonology and phonetics of consonant-tone interactions*. UCLA PhD dissertation.
- Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45:392–410. DOI: <https://doi.org/10.1016/j.csl.2017.01.002>
- Wood, Simon N. 2006. *Generalized Additive Models: an introduction with R*. Chapman and Hall/CRC.

PRELIMINARY ACOUSTIC ANALYSIS OF MINOR SYLLABLES IN KANISE KHUMI

Wen Hao TAN
Payap University
wenhao_1218@hotmail.com

Abstract

The Kanise Khumi language of the Khomic branch of Kuki-Chin is a typologically interesting case in exhibiting distinctive surface pitch patterns and vowel variation [i~ɜ~a]. This paper provides a descriptive account of minor syllable in Kanise Khumi with acoustic analysis of a set of words that correspond to minor syllables in seven other Khomic varieties. The study aims to elucidate the phonetic properties of minor syllables in the language and investigate how they fit into the description of this syllable type in the Khomic context. Archived data (Bryant 2020) is analyzed in terms of the vowel duration, vowel formants (F_1 and F_2), and pitch (F_0). The results showed that minor syllables in Kanise Khumi resemble minor syllables of other Khomic varieties to a considerable extent.

Keywords: Kuki-Chin, Khomic branch, Khumi, minor syllable
ISO 639-3 codes: cek

1 Introduction

Minor syllables typically have a reduced number of contrastive segmental features compared to major syllables, which have a more complex range. Moreover, the degree of phonotactic restrictions of minor syllable varies cross-linguistically, and the phonetic description of elements that fit this syllable type differs between languages in consequence (Thomas 1992; Butler 2014). As the non-final syllable of a sesquisyllable, this syllable type is prevalent in the Khomic family [Kuki-Chin, Peripheral Group (Peterson 2017)], which belongs to the Tibeto-Burman stock, in which sesquisyllables are well-attested (e.g., Matisoff 1973; Thomas 1992; Michaud 2012). Languages in the Khomic family take a more restrictive definition of minor syllable than the rest of sesquisyllabic languages. In previous descriptions of Khomic languages, minor syllables are described as underlyingly vowelless syllables, consisting of a single consonant. This interpretation implies that pitch and vowel quality are surface-level realizations. It is hardly surprising since pitch and vowel quality carry a low functional load and are not pertinent to phonological descriptions of minor syllables in Khomic languages (Herr 2011; Hornéy 2012; Peterson 2019).

In Kanise Khumi (hereafter, Kanise), a Khomic variety spoken by approximately 5,776 people in Chin State, Myanmar (Ikeda 2021), 1,124 out of 2,000 words in a collected wordlist contain purported minor syllable. Baleno (2020) conducted an impressionistic study on sesquisyllabic nouns in Kanise. She revealed that minor syllables exhibit high surface pitch variation from the conventional mid pitch. Vowel quality of minor syllables also vary as far as [i~ɜ~a] from the conventional [ə]. However, studies of related Khomic varieties do not report vowel and pitch contrasts in minor syllables (Herr 2011; Hornéy 2012; Peterson 2019). This discovery makes Kanise a typologically interesting case. Furthermore, it begs the question of to what extent does minor syllable in Kanise fits into the descriptions of minor syllables in the language family.

With this in mind, this paper presents the first acoustic analysis of Kanise's minor syllables. To avoid L1 and auditory-impression bias, a set of 113 words were chosen that are sesquisyllabic cognates with at least one of seven related Khomic varieties. Minor syllables were analyzed in terms of the vowel duration, vowel formants (F_1 and F_2), and pitch (F_0).

This paper serves as a pilot study and aims to provide a descriptive acoustic account of Kanise's minor syllable. The results of the study give more direction for ongoing research on morphotonemic studies of the language. §2 presents basic information about Kanise (language classification and

phonology). §3 reviews previous descriptions of minor syllable in related varieties and on Kanise. §4 proposes the research questions of this study. §5 discusses the methodology of the acoustic analysis. §6 presents the results, while §7 provides a qualitative discussion of the results. Finally, §8 ends the paper with concluding remarks on the results, limitations and future directions.

2 Language overview

2.1 Phonology

Kanise has a canonical word structure of $[C_1(V_1)(T_1)]C_2(C_3)V_2(N)T_2$. The parentheses represent optional elements. The first three elements, $[C_1(V_1)(T_1)]$, make up a minor syllable in a sesquisyllabic word. In minor syllables, only the obstruents / p t k ʔ s n m r ts / are attested in the C_1 slot. The phonological status of the minor syllable vowel (V_1) and tone (T_1) is currently under study. There is no constraint on the type of consonant that can occur as an onset in a full syllable (See Table 1 for the onset inventory). However, if an onset cluster is present, the first consonant (C_2) is restricted to a voiceless plosive, and the second consonant (C_3) a liquid: /r/ or /l/.

Kanise has nine simple vowels and four diphthong vowels, as in Table 2. All open-mid vowels are slightly diphthongized with vowel height raising towards the end of their articulation. The language has three phonemic tones, as listed in Table 3, with one checked tone.

Table 1: Onset inventory of Kanise

p ^h		t ^h		k ^h	
p		t		k	ʔ
b		d			
		s			h
	v				
		ts			
m		n		ŋ	
m̥		n̥		ŋ̥	
		r	j	ɰ	
		ɾ			
		l			
		ɭ			
		tɕ			

Table 2: Vowel inventory of Kanise

Monothongs			Diphthongs		
i	ĩ	u		ui	au
e		o			
ɛ [ɛ ^e]	ɜ [ɜ ^o]	ɔ [ɔ ^o]		oe	ai
	a				

Table 3: Tonal inventory of Kanise (Ikeda 2021)

No.	Description	Notation
1	High Level Modal tone	[55]
2	Falling Low Breathy tone	[31]
3	Mid Glottalized tone	[33ʔ]

A Kanise word is minimally composed of a CVT structure. A coda is optional and is restricted to the placeless nasal (N) only. The nasal coda may impart nasalization onto the preceding vowel and may be optionally deleted. Table 4 lists all attested word structures in Kanise. On the whole, the phoneme inventory of Kanise can be considered as relatively symmetrical and resembles related Kuki-Chin varieties such as Mara (Luce 1985), and Kaang (So-Hartmann 1988), Lemi Chin (Herr 2011).

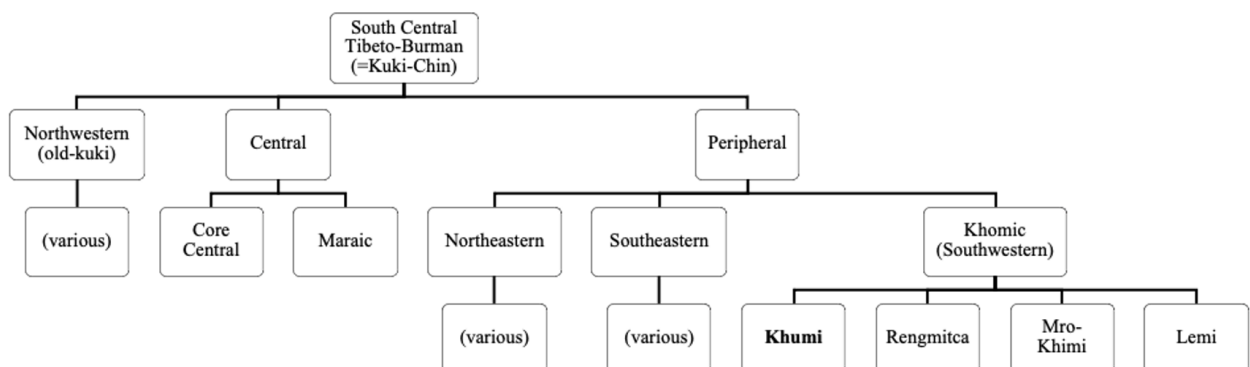
Table 4: Attested word types in Kanise

Word type	Example	Gloss	Word type	Example	Gloss
CV	buʔ	cooked rice	CVN.CV	louNɿ ta:ɿ	grindstone
CCV	krəʔ	deadfall	CVN.CVN	pɛNɿ pɛNɿ	dibble
CCVN	krouNɿ	trap	N.CV	ŋɿ ŋaʔ	one
Cə.CV	təʔpo:ɿ	cave	Cə.CVN.CV	ʔəʔ deNɿ təʔ	to pound rice
Cə.CVN	kəʔleNɿ	cliff	CV.CV.CV	ʔɿ hɿ li:ɿ	single
CV.CV	ʔaʔ si:ɿ	star	CV.CVN.CV	ʔaʔ pʰoNɿ təʔ	to pull out
CV.CCV	pəʔ tlɛ:ɿ	basket	CV.CCV.CV	ʔa:ɿ tlɿ təʔ	to weed
CV.CVN	tsɿ baŋɿ	sheaf of grain	N.CV.CV	ŋɿ.ʔɿ ʔaʔ	to give birth

2.2 Language classification

Kanise Khumi speakers live in villages south of the Sami subtownship in Chin State, Myanmar (Joseph Bryant, p.c.). The language is associated with the language code [cek] “Eastern Khumi Chin,” which is an overarching code shared by other highly mutually intelligible dialects such as Asang, Khenlak, Khongtu, Lemi, Likhy (Eberhard et al. 2020). Alternative names for this language include Nideun (Eberhard et al. 2020), Tahaensae (or Taheunso) (So-Hartmann 1988), or Uiphaw, which are names of the Kanise Khumi-speaking clans (Nathan Statezni, p.c.). As the name suggests, the language may very well belong to the Khumi cluster group “Khumi” within the Kuki-Chin branch of the Tibeto-Burman family (Peterson 2017). With similar considerations of linguistic similarities, various scholars have proposed Khumi as being under the Southern Chin group (Grierson 1927; Benedict 1972; Shafer 1974; Peiros 1998) or other Chin groups (Bradley 1997). In light of phonological and morphosyntactic evidence, Peterson (2000) formulated the Center/Periphery model in classifying the Khumi cluster group. He adopted the term Khomic group (formerly known as the Southwestern group) as a first-order branching within the Peripheral group alongside the Southern and Northeastern groups (Peterson 2017), as shown in Figure 1. There has been no further classification proposed within the Khomic group.

Figure 1: South Central Tibeto-Burman (Kuki-Chin) subgroupings (Peterson 2017)



3 Previous studies of minor syllables in the Khomic family

3.1 Previous studies in related varieties

There have been three descriptions done on minor syllables in the Khomic sub-branch: in Lemi Chin (Herr 2011), Mro Khimi (Hornéy 2012), and Bangladesh Khumi (Peterson 2010; Peterson 2011; Peterson 2019). In this language family, minor syllables function either as fossilized lexical morphemes from historical compounds or as grammatical morphemes. These formatives can be lexicalized in some cases (Peterson 2013).

Minor syllables typically have a much shorter duration than major syllables, and the same holds for syllables in Khomic languages. Herr (2011) measured the vowel length of minor and major syllables across ten sesquisyllables in Lemi Chin. She found minor syllables to have an average duration of 0.0565 seconds compared to 0.2152 seconds in major syllables, which is about four times the duration. As for vowel quality, all documented Khomic varieties congruously exhibit non-contrastive or predictable minor syllable vowels, even though phonemic central vowels are present in these languages and can be auditorily indistinguishable from minor syllable vowels. Lemi Chin has vowel phonemes /i, ə, ø, a/, but among them [ə and ø] are also found to be surface realizations of minor syllable vowels. Moreover, all surface vowels [ɜ, ə, e, v, ə, ø] of minor syllables have similar vowel quality compared to the phonemic vowels in the central vowel range. Despite having phonemic centralized back vowels /uu/ [i ~ u], /ɜ/ [ɣ ~ ə], minor syllables in Mro Khimi are described as carrying a “surface vowel” that “has a much shorter duration than vowels in other syllables.” In Bangladesh Khumi, Peterson noted a short vowel associated with the sesquisyllabic structure. He labeled the vowel as “optional” in terms of well-formedness (Peterson 2011). This “optional vowel” is distinctive from the phonemic close-mid /ə/ vowel. As a whole, Khomic varieties conform with the conventional view that the vowel quality of minor syllables is not phonologically specified.

As observed above, the syllable structure of minor syllables is generally understood as an underlyingly vowel-less /C/. That is, “/C-/” in Lemi Chin, “/Cə-/” with “inserted schwa” in Mro Khimi, and “/C(v)/” with “optional vowel” in Bangladesh Khumi. Since there is no phonological nucleus present, minor syllables are considered non-tone-bearing or tonally irrelevant, and any surface pitch found is non-contrastive. It has been reported that Lemi Chin has a predictable mid-tone as the default tone of minor syllables (Herr 2011). Bangladesh Khumi, on the other hand, restricts minor syllables from occurring with all five lexical tones. In Mro Khimi, tone assignment in minor syllables is post-lexical and comes after vowel [ə] insertion as a two-part rule (Hornéy 2012). This means vowel and tonal quality are as-yet-unspecified before rule assignment.

3.2 Previous studies in Kanise Khumi

Baleno (2020) conducted a preliminary study on purported sesquisyllabic nouns in Kanise Khumi. Her study revealed that surface minor syllables vowels span as far as [i~ɜ~a] but are primarily realized as schwa [ə]. Syllabic nasal minor syllables [ʔŋ] were also present in her study. In addition, she found that minor syllables with surface [a] vowel are longer in duration than other minor syllables. She concluded that this [Ca] form requires further study to elucidate its status as a minor syllable. In the author’s observation, the surface [ɜ] vowel repeatedly occurs with minor syllables of the same morpheme. This morpheme has hitherto been speculated to be the (inalienable) possessive morpheme (Peterson 2011). It is also noteworthy that the [ɜ] vowel is exceptionally similar in phonetic quality to the conventional [ə] vowel and showed up in major/full syllables as a phonemic vowel (/ɜ/).

As for minor syllable pitch, Baleno suggested that minor syllables primarily carry a surface mid-pitch but show high pitch variations. Both surface pitches align well with the description of lexical mid-glottalized tone (M) and high-level modal (H) tone of Kanise (Ikeda 2021), that is, 143-186Hz of high surface pitch (averaged 162Hz) compared to 150-170Hz of lexical H tone, and 128-164Hz of surface mid-pitch (averaged 146Hz) to 133-158Hz of M tone. Thus, there are two possibilities: minor syllables are restricted to a default mid-tone, as with Lemi Chin, or that minor syllables may indeed have phonological variation, as with Mro Khimi. In sum, minor syllables in Kanise are a typologically

interesting case for an acoustic analysis. All Khomic varieties show slight variations in the phonetic properties of their minor syllables.

4 Research Questions and Methodology

The rest of this study addresses the following research questions:

- i. How does the duration of minor syllable compare to the duration of major syllables?
- ii. Does the tone of the major syllable affect the duration of the minor syllable?
- iii. Does the minor syllable surface vowel [ə] share the same vowel space with the surface vowel [ɜ]?
- iv. Is the pitch of the minor syllable affected by the pitch of the major syllable?

Having presented the research questions, the following sub-section outlines the research methodology of the study as follows: §4.1 discusses the data source (collection method and subjects), §4.2 on the materials used (data selection and process), and §4.3 on the research procedure.

4.1 Data collection

The Kanise data used in this study is sourced from a 2,076-item wordlist provided by Joseph Bryant. The wordlist is designed based on the *EFEO-CNRS-SOAS Word List for Linguistic Fieldwork in Southeast Asia* (Pain et al. 2019). Bryant assembled this wordlist during a one-week elicitation session back in 2019. The informant for the wordlist was a native male speaker in his 50s, with a younger male speaker furnishing another 193 words. After the elicitation, Bryant has transcribed all 2,076 items in the wordlist and supplemented Kanise orthography. He also made recordings of every entry in the wordlist in WAV format with a sampling rate of 44.1 kHz. It is essential to point out that the variety elicited in the data cannot represent all Kanise-speaking clans, only the one spoken by the subject and his village. Also, background noises can be heard in some of the recordings but do not have a significant effect in causing speech features to be unrecoverable or heavily distorted. The 2,076-item wordlist data is archived in Zenodo and accessible upon request (Bryant 2020).

4.2 Materials

As Kanise is in an early documentation stage, there is no adequate knowledge of the semantic and phonological make-up of minor syllables in the language. Hence, a wordlist consisting of a set of sesquisyllabic cognates (n=113) was prepared. The wordlist is sourced from seven related Khomic varieties in which the sesquisyllabic word type has already been established. This minimizes the likelihood of using irrelevant study material for this paper. Each item in the wordlist is cognate with at least one of the seven related Khomic varieties (cf. Table 9). The details of the related varieties are in Table 5.

Table 5: Eight Khomic varieties related to Kanise Khumi chosen in this study

Given language name	Sourced from	Spoken in	Notes
1. Bangladesh Khumi	Peterson 2010, 2011, 2019	Ruma Bazaar, Bandarban Hill Tracts, Bangladesh	-
2. Lemi Chin	Herr 2011	Paletwa township	-
3. Mro Khimi	Hornéy 2012	Paletwa and Rakhine townships	-
4. Awa Khumi	Luce 1985	Tamantha, west-southwest of Paletwa	Based on Saptawka (1934)
5. Khomi	So-Hartmann 1988	Paletwa township	-
6. Ahraing Khumi	Luce 1985	Kaletwa township	-
7. Khumi	VanBik 2009	Ruma Bazaar, Bandarban Hill Tracts, Bangladesh	Based on Peterson (2013)
8. Khimi	Shafer 1944	unknown, possibly Southern Paletwa	-

Further interval segmentation and labeling of the sound files of wordlist were done using *TextGrid creator and editor* scripts ran on *Praat version 6.1.39* (Boersma and Weenink 2021). All sound files were downsampled to 10,000 Hz when exported into *TextGrid*. The rhymes of major and minor syllables were segmented individually on a labeled tier with boundaries set of the beginning and end of the rhyme - the onset and offset of the second formant, respectively, where possible. The acoustic measures of the labeled segments were then further extracted in *Praat* using the *PraatSauce* script (Kirby 2020). Finally, extracted results were gathered into several ensemble files by rerunning the script, and graphical representations of the data were produced using the *plotnine* package (Kibirige et al. 2021). It is to be noted that only the second repetition of the three was analyzed throughout the study. Besides, maximum formant values were set to 5,000 Hz for males and 5,500 Hz for females.

4.3 Procedure

This pilot study is designed to demonstrate the phonetic properties of Kanise minor syllable and investigate how they fit into the description in the Khomic context. Hence, the study looked at the acoustic output of duration, vowel formants (F_1 and F_2) and pitch (F_0).

The first experiment examined the vowel duration of minor syllables as compared to that of major syllables. In this experiment, the tone carried by the corresponding major syllable is being held as a control variable since tone can affect the pitch and duration of neighboring syllables (Wang and Chen 1994). Therefore, the wordlist is grouped into three categories: category H (high-level modal tone, $n=36$), category M (mid-glottalized tone, $n=32$), and category L (low falling breathy tone, $n=47$). In addition, the potential effect of the tonal categorization on the duration of the minor syllable was kept under observation. However, the primary interest of this study was the durational difference between minor and major syllable.

The second experiment compared the vowel quality of the conventional surface vowel [ə] versus the surface [ɜ] vowel of the possessive morpheme to see whether they share the same vowel space or are acoustically distinct. Out of the 113 tokens in the wordlist, there were 17 tokens found of this morpheme (cf. Table 8 for wordlist of [ɜ] tokens). Since the rate of syllable reduction differs between languages, it remains unclear whether the sesquisyllable in Kanise still allows vowel contrast in its minor syllable. Suppose [ɜ] is a phonologically specified vowel quality. In that case, we would expect to see a higher F_1 and narrower range of F_2 values. On the contrary, we would expect [ə] to exhibit a wider range of formant values (Davidson 2006). It is worth mentioning that five minor syllable tokens in the wordlist contain vocalic elements that are distinctive in terms of vowel quality: three tokens of [a], one of [i], one of [o]. There are also three syllabic nasals [ŋ], which are not a vowel quality altogether. This finding corroborates Baleno's (2020) description of [i], [a], and [ŋ] surface variation on minor syllables. Though it is not conventional to see minor syllables that carry such distinctive vocalic elements, they are indeed passable minor syllable materials in that: i) they form a weak-strong pattern (disyllabic iamb) with the following syllable and are shorter in duration (Butler 2014), and ii) they could be minor syllables that carry contrastive phonemes in their vocalic elements as per Thomas' (1992) type III and IV minor syllables.

Since this paper does not address the working definition of minor syllable in Kanise, they were included in the experiment for this study. However, they were kept in sight for being potential outliers in terms of vowel quality. In measuring vowel formants, only data from the middle 50% intervals of the minor syllable rime were measured then averaged. This is to rule out any potential extreme variance conditioned by neighboring consonants, especially glides and nasals, or glitches in formant approximation.

The third experiment measured the average F_0 contour of minor and major syllables across three tonal categories. We are interested to see how the F_0 contour patterns over minor syllables and the stability of the contour. Suppose the F_0 contour merely follows the trajectory of the major syllable pitch. In that case, this suggests that the minor syllable pitch is interpolated from context and might be phonologically unspecified (no pitch specification). On the other hand, a stable tone pattern that is not explicable by contextual interpolation may suggest a contrastive pitch target (West 2014) or a tonally irrelevant default pitch. In measuring pitch, the duration of the rimes of the minor syllable was

normalized relative to that of the major syllable. Therefore, the minor syllable is fitted between $t = 0$ and $t = 0.2$ and major syllable between $t = 0.2$ and $t = 1$.

5. Results

5.1 Vowel duration

When it comes to vowel duration, the mean vowel duration of minor syllables in all three tonal categories are not significantly different from each other, as evidenced by the boxplot in Figure 2 and the durational measurements in Table 6.

Figure 2: Vowel duration of major and minor syllables in Kanise across three tonal categories

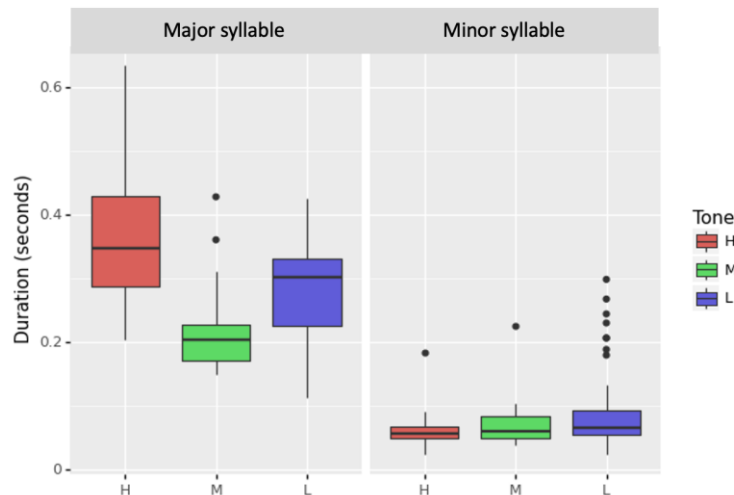


Table 6: Mean vowel duration of minor and major syllables in Kanise across three tonal categories and their durational ratio

Tone of major syllable	Syllable type	Mean duration	Minor/ major
H	Minor	59.4ms	0.1638/1
	Major	362.6ms	
M	Minor	71.6ms	0.3212/1
	Major	222.9ms	
L	Minor	92.4ms	0.3344/1
	Major	276.3ms	
Average	Minor	74.5ms	0.2593/1
	Major	287.3ms	

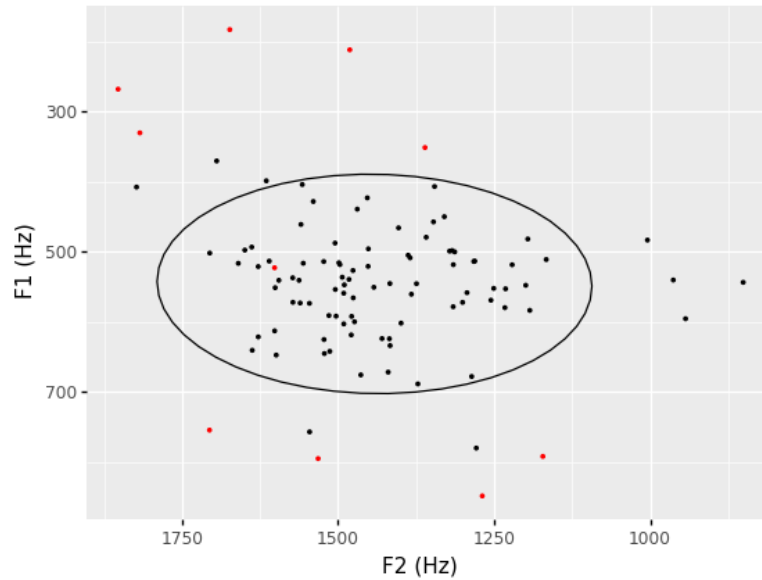
The tone carried by the major syllable seems to have a high correlation with the duration of the major syllable. For example, major syllables in the H category, which carry an unchecked tone (high modal), are the longest. On the other hand, major syllables in the M category are the shortest, having checked tone (mid-glottalized). However, in terms of the relationship between major syllable tone and the duration of minor syllable, they appear not to correlate. There is merely a 33ms difference between the longest (L category) and shortest (H category) minor syllable on average.

In addition, as for the durational difference between minor syllables and corresponding major syllables, sesquisyllables in category H has the shortest minor syllable and longest major syllable among the three categories. This results in their greater difference in length (59.4ms/ 362.6ms) compared to category L, which has the least durational difference (0.0924/0.2763). As a whole, minor syllables have about one-fourth (~26%) of the duration of major syllables (0.2593:1).

5.2 Vowel quality

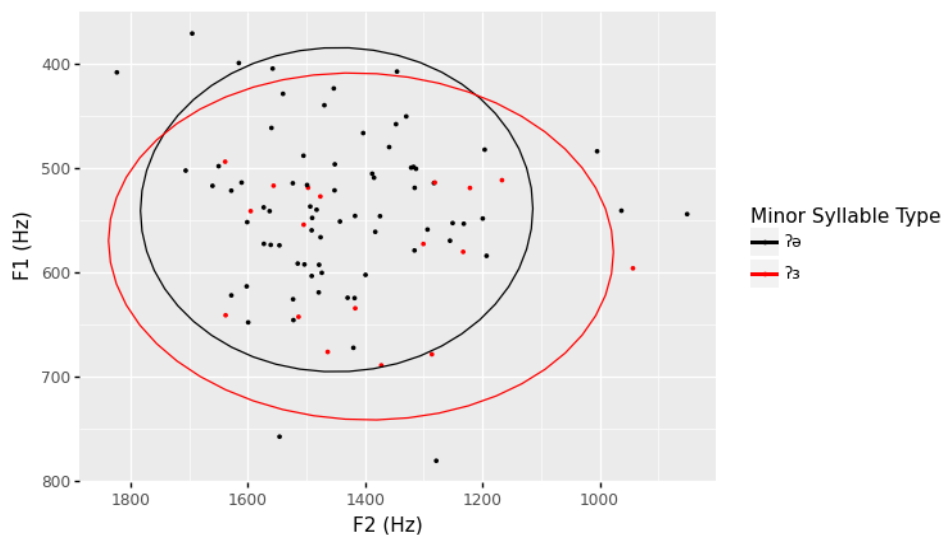
This part of the study is concerned with the vowel quality of the Kanise minor syllable. As a whole, minor syllable vowels in Kanise have their F_1 and F_2 values distributed around the mid-central range, with the centroid positioned around 1400, 550, as demonstrated in Figure 3.

Figure 3: F_1 and F_2 distribution plot of the Kanise minor syllable vowel



Moreover, most of the extreme variants in the distribution plot are identified as tokens with distinct vowel qualities (the aforementioned [i, a, o] vowels and syllabic nasals in §4.2). These outliers are plotted in red in the diagram, with their values falling far away from the mid-central vowel range. In Figure 4, minor syllables that consist of the possessive morpheme [ʔɜ] (n=17) are contrasted with all other minor syllables [ʔə] (n=96). The result (Figure 4) suggests that, generally speaking, [ʔɜ] has slightly higher F_1 values, but the difference is not striking.

Figure 4: F_1 and F_2 distribution plot for surface [ɜ] vs. surface [ə] vowel in Kanise

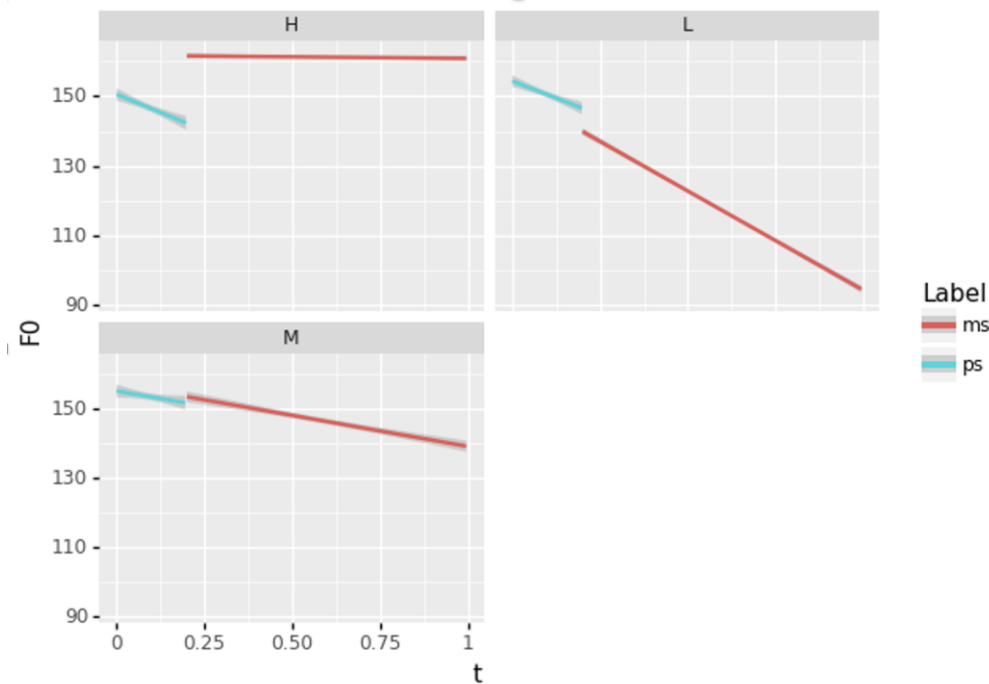


The F_1 values of the surface [ɜ] vowel can be seen scattered around the lower range of the surface [ə] vowels. Furthermore, the centroid of the [ɜ] vowel nearly approaches the average F_1 and F_2 values measured for the phonemic [ɜ] vowel in major syllables (see §11.1 for measurements). Nonetheless, the distribution range is not significant enough to be drawn separately from the surface [ə] vowels. This begs the question of whether the two categories are distinctive and whether [ɛ] and [ə] should be treated as two separate vowel categories.

5.3 Pitch

The last experiment was conducted to determine the predictability of pitch contours of minor syllables, and the result points toward two possible interpretations. Figure 5 illustrates the normalized pitch of minor syllables (labeled “ps”) and major syllables (labeled “ms”) across three tonal categories. In categories M and L, the surface F_0 patterns suggest an interpolation of F_0 values from context. The minor syllable F_0 in these two categories shows a falling slope contour, with the degree of slope close to that of the following full tone. This continuation of contour suggests that the F_0 patterns of minor syllables in these two categories may be an interpolation from the pitch specification of the following major syllable.

Figure 5: normalized F_0 contours of minor and major syllables of the three tonal categories in Kanise



Category H, on the contrary, does not reinforce this interpretation. When preceding H tone, the pitch pattern of minor syllables remains largely the same, exhibiting a mid-tone-like downtrend contour. In other words, the minor syllable pitch in category H does not anticipate the trajectory of the following full tone. When the three tone categories are collapsed together, the minor syllable pitch has a fairly consistent shape. The surface falling pitch starts from 150-155Hz (a difference of 0.568ST) and ends at 151-143Hz (a difference of 0.942ST). This fits well within the contour range of lexical M tone (133-158Hz) (Ikeda 2021).

6 Discussion

The study results are consistent with the observations on minor syllables in other Khomic languages.

First of all, the duration measurements of minor and major syllables indicate that the length of the minor syllables are roughly a quarter of that of the major syllables in Kanise. This finding conforms with Herr’s (2011) measurements of Lemi Chin’s minor versus major syllables and the general description of sesquisyllables being disyllabic iambs where the minor syllable is shorter in duration than the major syllable (Butler 2014). In addition, major syllable tones have no significant effect on the duration of the minor syllable, and the durational difference of the minor syllable across the three tonal categories seems trivial.

As for vowel quality, acoustic evidence points to the fact that surface vowels [ə] and [ɜ] are not acoustically distinct. Though the values for [ɜ] distribute in the lower area of the vowel space of [ə], both vowels appear to share substantial vowel space. Therefore, it may be more plausible to assume the minor syllable vowel in Kanise to be phonologically unspecified or underspecified in vowel quality, as

reported in all related Khomic varieties (Peterson 2010; Herr 2011; Hornéy 2012). Nonetheless, a further token-by-token study can lend itself to validate this explanation.

Lastly, it is found that the surface pitch patterns of Kanise minor syllables, generally speaking, are not subject to environmental influence and remain consistent across the three tonal categories. Following the same line of argument, the minor syllable pitch contours could be a “default” phonetic mid-tone, as per West’s (2014) underspecified or neutral tone. This type of tone is not derived from any lexical tone and lacks significant contrast. In such a case, the tone is not phonologically specified but inexplicable without any surface specification, as commonly attested in Chin varieties Lemi Chin (Herr 2011) and Hkongso (Wright 2009) and more generally, among Tibeto-Burman languages (Morse 1963).

An alternate interpretation could be the surface [ɜ] vowel and mid-pitch of minor syllables as phonologically specified properties. Since the average F_1 values for [ɜ] are slightly higher than [ə], one could posit [ɜ] as a vowel with distinct phonemic status, not just an allophone. There exists a possibility that the short duration of the minor syllable, which correlates with the lack of stress, is preventing the [ɜ] vowel from reaching full target attainment or distinctive vowel space from [ə]. It is universally true that when a vowel occurs in weakly stressed syllables, the formant undershoots as the vowel duration decreases (Lindblom 1963). However, because minor syllable vowels are often neutralized due to vowel reduction, it makes the mid-central range a poor area for vowel quality contrasts. We would expect contrasting sounds to be perceptually distinct instead (Flemming 2004). Hence, this puts the [ɜ] vowel in a disadvantaged position in functioning as a phonological vowel considering there is “no motivation to resist the pressure to assimilate to context” (Flemming 2009). Since the current results only touch upon this possibility, a closer examination is needed to elucidate its status. For example, surface [ɜ] vowels, or minor syllable vowels in general, should be compared with the [ɜ] vowel that has phoneme status in major/ main syllables in a more carefully designed experiment. In doing so, we could see if they are phonologically the same vowel placed in two different contexts.

Furthermore, the minor syllable pitch may suggest a phonetic pitch target and thence a phonemic tone. One argument would be that the surface mid-pitch in minor syllable is the shorter stimulus of the lexical mid-tone. Ikeda (2021) discovered the lexical mid-tone to have an average F_0 of 132Hz at the end of the rhyme. In contrast, in our experiment, the minor syllable’s mid-pitch ended with a 151-143Hz range. Therefore, we can propose that the minor syllable mid-pitch does not fully reach the end of the fall due to the limited time to complete the full approximation of tone, as with the minor syllable’s vowel. Having said that, the falling contour may not even be a distinctive cue of the lexical mid-tone, since not all entries of this tone occur with a “short-fall ending” (Ikeda 2021). This makes the minor syllable pitch more plausible to be a tonal phoneme if the falling ending is optional. However, an in-depth study is needed to confirm such a hypothesis.

7 Conclusion

The present study has provided a brief description of minor syllables in Kanise based on acoustic correlates of vowel duration, vowel quality, and pitch. The results have shown that minor syllables in Kanise fit into the description of minor syllables in the Khomic context to a considerable extent. However, it should be stressed that looking at these phonetic properties alone do not allow us to draw conclusions about the phonological status of minor syllables. This is because many factors can be influential, such as etymological origin, prosodic weight, and the morphological content of minor syllables.

That said, the results of the study have demonstrated the need to provide a more restrictive working definition of the minor syllable in Kanise. We have observed several tokens of minor syllables with extreme variants of vocalic elements in the experiment on vowel quality. This has raised questions about their status as minor syllables, or more explicitly, the non-final syllables of sesquisyllables. Since sesquisyllables are a type of disyllabic iamb among other word types (e.g., extended monosyllables, near-disyllables), there can potentially be an extensive range of possible durations, vowel, and pitch distribution in a non-final syllable. Furthermore, these word types appear to lie on a continuum instead of discrete categories (Butler 2014). This motivates the need to explicate the exclusivity of minor syllables in this language.

It should be noted that this study has several limitations. The set of sesquisyllables that was chosen might not be representative enough for this study. A bigger set of data may shed new light. However, the study is by nature exploratory as it is based on only one speaker. Furthermore, the paradigm created for this study can also be improved, specifically in controlling the variables that have a potential effect on the experiment, such as the segmental context (lexical tone, potential boundary tone, and surrounding consonants).

Although the research paradigm was not designed to address these factors, the study helps us better understand the phonetic properties of minor syllables in Kanise.

References

- Baleno, Grace. 2020. *Surface pitch patterns on Kanise Khumi sesquisyllabic nouns*. Research Report. Trinity Western University: Canada Institute of Linguistics.
- Benedict, Paul K. 1972. The Sino-Tibetan tonal system. In *Langues et techniques, nature et société*, ed. Jacques Barrau and et al, 25–34. Paris: Klincksieck.
- Boersma, Paul, and David Weenink. 2021. Praat: doing phonetics by computer (Version 6.1.39). Computer program.
- Bradley, David. 1997. Tibeto-Burman languages and classification. In *Papers in Southeast Asian Linguistics 14: Tibeto-Burman Languages of the Himalayas*, ed. D. Bradley, 86:1–72. Pacific Linguistics, Series A. Canberra: Australian National University.
- Bryant, Joseph. 2020. Kanise Khumi Wordlist Data. Archive. Retrieved 15 September 2020.
- Butler, Becky Ann. 2014. Deconstructing the Southeast Asian sesquisyllable: a gestural account. PhD dissertation, Ithaca, New York: Cornell University.
- Davidson, Lisa. 2006. Schwa Elision in Fast Speech: Segmental Deletion or Gestural Overlap? *Phonetica* 63: 79–112. <https://doi.org/10.1159/000095304>.
- Eberhard, David M., Gary F. Simons, and Charles D. Fennig, ed. 2020. *Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: Twenty-third*. Dallas, TX: SIL International.
- Flemming, Edward. 2004. Contrast and Perceptual Distinctiveness. In *Phonetically-Based Phonology*, Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge, MA: Cambridge University Press.
- Flemming, Edward. 2009. The phonetics of schwa vowels. In *Phonological Weakness in English: From Old to Present-day English*, Donka Minkova, 78–98. Basingstoke: Palgrave Macmillan.
- Grierson, George A. 1927. Linguistic Survey of India, Vol. I, Part I, (Introductory). *Calcutta*: 32-38).
- Herr, Kristen Elizabeth. 2011. The phonological interpretation of minor syllables, applied to Lemi Chin. MA thesis, Chiang Mai, Thailand: Payap University.
- Hornéy, Christina Scotte. 2012. A Phonological Analysis of Mro Khimi. MA thesis, Grand Forks, ND: University of North Dakota.
- Ikeda, Elissa. 2021. Tone patterns in Kanise Khumi monosyllables. In *Proceedings of the 11th Payap University and UNRN Research Symposium 2021*, 1223–1234. Chiang Mai, Thailand: Research and Academic Service Affairs.
- Kibirige, Hassan, Greg Lamp, Jan Katins, gdowding, austin, matthias-k, Tyler Funnell, et al. 2021. *Plotnine* (version v0.8.0). Zenodo. <https://doi.org/10.5281/zenodo.4636791>.
- Kirby, James. 2020. Praatsauce: Praat-based tools for spectral analysis. Computer program.
- Lindblom, Björn. 1963. Spectrographic study of vowel reduction. *The journal of the Acoustical society of America* 35. Acoustical Society of America: 1773–1781.
- Luce, Gordon Hannington. 1985. *Phases of pre-Pagan Burma: Languages and history*. Vol. 2. School of Oriental and African Studies. Oxford: Oxford University Press.
- Matisoff, James A. 1973. Tonogenesis in Southeast Asia. In *Consonant Types and Tones*, ed. Larry M. Hyman, 71–95.

- Michaud, Alexis. 2012. Monosyllabicization: patterns of evolution in Asian languages. *Monosyllables: from phonology to typology*. Berlin: Akademie Verlag: 115–130.
- Morse, Robert H. 1963. Phonology of Rawang. *Anthropological Linguistics* 5. [Anthropological Linguistics, Trustees of Indiana University]: 17–41. JSTOR.
- Pain, Frederic, Michel Ferlus, Alexis Michaud, Thị Thu Hà Phạm, and Ryan Gehrman. 2019. EFEO-CNRS-SOAS Word List for Linguistic Fieldwork in Southeast Asia. Identifier: oai:halshs.archives-ouvertes.fr:halshs-01068533.
- Peiros, Iliá. 1998. *Comparative Linguistics in Southeast Asia*. Books Series. Pacific Linguistics, Research School of Pacific and Asian Studies, National University of Australia.
- Peterson, David A. 2000. On the status of the Southern Chin subgroup. In . Bangkok, Thailand: Trang.
- Peterson, David A. 2010. Khumi elaborate expressions. *Himalayan Linguistics* 9. <https://doi.org/10.5070/H99123402>.
- Peterson, David A. 2011. Core Participant Marking in Khumi. *Linguistics of the Tibeto-Burman Area* 34.2: 73–100.
- Peterson, David A. 2013. Affecting valence in Khumi. *Language Typology and Historical Contingency: In honor of Johanna Nichols*: 171–194.
- Peterson, David A. 2017. On Kuki-Chin subgrouping. In *Sociohistorical linguistics in Southeast Asia*, ed. Picus Sizhi Ding and Jamin Pelkey, 189–209. Leiden: Brill.
- Peterson, David A. 2019. Bangladesh Khumi. In *The Mainland Southeast Asia Linguistic Area*, ed. Alice Vittrant and Justin Watkins, 314:12–55. Trends in Linguistics Studies and Monographs. Berlin & Boston: Walter de Gruyter GmbH & Co KG.
- Saptawka. 1934. *Mro. Angthang akahawih Matthai tariteh*. Translated by St. Matthew. Yangon, Myanmar: The British and Foreign Bible Society.
- Shafer, Robert. 1944. Khimi Grammar and Vocabulary. *Bulletin of the School of Oriental and African Studies, University of London* 11. Cambridge University Press: 386–434. JSTOR.
- Shafer, Robert. 1974. *Introduction to Sino-Tibetan*. Wiesbaden, Germany: Otto Harrassowitz.
- So-Hartmann, Helga. 1988. Notes on the Southern Chin languages. *Linguistics of the Tibeto-Burman Area* 11: 98–119.
- Thomas, David. 1992. On sesquisyllabic structure. *Mon-Khmer Studies* 21: 206–210.
- VanBik, Kenneth. 2009. *Proto-Kuki-Chin: a reconstructed ancestor of the Kuki-Chin languages*. STEDT Monograph Series 8. Berkeley, Calif: Sino-Tibetan Etymological Dictionary and Thesaurus Project, Dept. of Linguistics research unit in Univ. of California, Berkeley.
- Wang, Yih-Ru, and Sin-Horng Chen. 1994. Tone Recognition of Continuous Mandarin Speech Assisted with Prosodic Information. *The Journal of the Acoustical Society of America* 96.
- West, Luke. 2014. Pitch specification of Sgaw minor syllables. PhD dissertation, University of Toronto.
- Wright, Jonathan M. 2009. Hkongso Grammar Sketch. Northgate Dr, Irving: Dallas International University.

Appendix

Table 7: *F₁ and F₂ values of the phonemic /ɜ/ vowel in Kanise word-initial full syllables*

No.	Token	Gloss	IPA	F1	F2
1	nayn hryh	sieve	ʒr̥ɜː ɸɥɿ	636.75Hz	1453.18Hz
2	may yang	to fish	pɜːɿ j̥ãɿ	634.07Hz	1349.07Hz
3	qayn thyhn	young (to be young)	ɟ̥ɜːɿ t̥h̥ɿ	556.79Hz	1438.75Hz
4	pay qui	warp (weaving)	pɜːɿ ɣuiɿ	558.11Hz	1406.04Hz
5	pay ryh	to draw in a casting net	pɜːɿ ɿɥɿ	522.25Hz	1286.04Hz
6	tlayn layn	to join	kl̥ɜːɿ l̥ɜːɿ	652.47Hz	1484.30Hz
7	dayn cayn	to stand on tiptoe	d̥ɜːɿ s̥ɜːɿ	612.80Hz	1601.54Hz
8	thayn kay	smaller than	t̥h̥ɜːɿ k̥ɜːɿ	650.44Hz	1567.03Hz
			Average:	602.93Hz	1400.72Hz

Note: Measurements were taken across 8 tokens, at the midpoint of the rhyme.

Table 8: *Tokens in Kanise containing the possessive morpheme with surface [ɜ] vowel in this study*

No.	Token	Gloss	IPA	No.	Token	Gloss	IPA
0183	y kOUNg	‘trunk’	ʔɜːɿ k̥õːɿ	1172	y ku	‘hand’	ʔɜːɿ k̥uɿ
0203	y pa	‘flower’	ʔɜːɿ p̥ãɿ	1197	y phai	‘thigh’	ʔɜːɿ p̥hãɿ
0211	y thai	‘fruit’	ʔɜːɿ t̥hãɿ	1219	y hou	‘bone’	ʔɜːɿ hõuɿ
0815	y hei	‘head louse’	ʔɜːɿ h̥eɿ	1228	y thi	‘blood’	ʔɜːɿ t̥h̥iːɿ
1098	y qau	‘corpse’	ʔɜːɿ ɣauɿ	1231	y nhan	‘flesh’	ʔɜːɿ n̥ãːɿ
1101	y lu	‘head’	ʔɜːɿ l̥uɿ	1256	y hmyhn	‘bile, gall’	ʔɜːɿ h̥m̥ɜːɿ
1105	y sang	‘hair of head’	ʔɜːɿ sãːɿ	2219	y hnai	‘pus’	ʔɜːɿ n̥ãɿ
1114	y mei	‘eye’	ʔɜːɿ m̥eɿ	2884	y ni	‘he, she, it’	ʔɜːɿ n̥iɿ
1135	y haw	‘tooth’	ʔɜːɿ h̥õɿ				

Table 9: *Experimental materials: sesquisyllabic cognates (n=113) across eight related Khomic varieties (White represents phonemic transcription; light grey represents phonetic transcription and dark grey represents orthography.)*

Bryant 2020		Peterson 2019	Herr 2011	Hornéy 2012	Luce 1985	Hartmann 1988	Luce 1985	VanBik 2009	Shafer 1944
Kanise Khumi		Bangladesh Khumi	Lemi Chin	Mro Khimi	Awa Khumi	Kho-mi	Ah-raing Khumi	Khumi	Khim
IPA	English								
pəɸloũɿ	heart	p’liwng ⁴	b.lũ ⁵		mlũ ²	p ^o le ^u ŋ	pǎleũ ²	pliíwng	blung
kəɸnõɿ	ear	k’no ⁴	k.nõ ¹	k-ná	kãna ²	k ^o nó?	kãno ²	knoó	kãnaw
kəɸniɿ	sun, day	k’ni ⁴	k.ni ¹	k-ní		kani		kníi	kãning
k ^h õɸɸ ¹ dãɿ	tomorrow	kh’ dang ¹⁰	khõi ⁵ dã ¹						kã dang
θ̥ɸtrouɿ	six	t’riw ¹	t.ru ⁵			t ^o ri ^u ?		triw	taruk
tsəɸpoːɿ	son	c’po ¹	s.põ ³						chãpaw
pəɸliɿ	four	p’lü ⁵	b.li ³			p ^o li		plúee	bili
təɸvõɿ	bear	t’võng ¹¹	t.võ ¹					tvóeeng	

tsəɫ nnouʔɬ	daughter	c'niw ¹⁰	s.nũ ⁵						chänu
pəɫtoːɫ	to teach	p'tiw ¹	m.tu ⁵						mätu
pəɫdeɫ	to kill	p'döy ²				p ^ə deɪʔ		p ^ə thueng	mädü
pəɫdajɬ	to smooth	p'lây ⁴				pəná ¹			
təɫhɔɫ	to pierce	t'hëng ²							
təɫhĩɬ	to shake								
pəɫ ʔaũɬ	to open	p'ewng ³	p.ʔɔ ¹					a(m)' éwng	päawng
pəɫsɔɫ	to empty intestine	p'kew ²							
pəɫtoɫ	to send	p'to ²							mätaw
pəɫloɫ	to wake sm up	p'lâ ²							
pəɫ t ^h oɛɬ	to spit	p'thây ¹	m. thoi ³			p ^ə t ^h ɔ ¹ ʔ			mä thawi
təɫbaũɫ	mouth	l'bewng ¹	t.bɔ ⁵			t ^ə bɛ ^ə ŋ	lɔ̃ bɛäũ ⁵		täbawng
təɫ praũɫ	cemetery		t.prũ ⁵						
təɫpoːɫ	cave		t.pu ³						täbu
ʔɛɫpaʔɬ	flower		ʔ.pa ³			pa ^{uʔ}		páw	
mɛɫɾɛɫ	town		m.ri ³						
kəɫɬɪɫ	to work		k.lu ³						
təɫmɔɫ	wound		t.mɔ ¹	t-m̩ a				tmoó	
ʔɛɫt ^h aiʔ	fruit		ʔ.thai ³			ət ^h ai ^{1ʔ}		atháy	äthai
ʔɛɫ vâɫ	price		ʔ.vâ ¹						ävang
təɫvoɫ	stream		t.vɔ ¹						tävaw
təɫvoːɫ	bird		t.vɔ ⁵	t-vá		təvó		tvoo	tävaw
sĩːɫ p ^h ɾĩɫ	needle		s ^h .p ^h rĩ					sprúeeng	
təɫŋaiɬ	to listen		t.ŋai ¹					tngáy	tängai
ʔnɬdãɫ	breakfast		k.dã ¹						
kəɫjauɫ	belly		ʔ.jo ³						
ʔɛːɫ mɛɬ	eye		ʔ.mi ³		ã miʔ ³	meiʔ	mei ⁵	móey	ämih
ʔɛɫ p ^h aiɬ	thigh		ʔ.p ^h ai ¹		ã p ^ə æ ⁴	p ^h aːi	p ^ə æ ⁵	pháy	
ʔɛːɫ lhɔɫ	tooth		ʔ.fɔ ¹		ã fɔ ²	hó	ho ²	hoó	äfaw
kəɫsɛɫ	ginger		k.thĩ ¹						
təɫkaiɫ	tiger		t.kai ¹			t ^ə kaːi		t káay	
kəɫt ^h ãɫ	fat (n.)		ʔ.t ^h ã ³			t ^h aːuʔ		tháw	
ʔnɬdãɫ	morning		k.dã ¹						kändang
səɫnɛɫ	year		s ^h .ni ¹			s ^h a niŋ			säning
pəɫjuɫ	rat		b.jɔ ⁵			p ^ə ju		búuy	
pəɫyuiɫ	snake		b.rui ¹			p ^ə vuːi		pvúuy	mägui
pəɫt ^h ãɬ	fly		m.t ^h ã ³	m-thɔ		p ^ə t ^h ã ^u		pthaw	
pəɫt ^h iːɫ	comb		m.t ^h i ³					pthi	

təŋgiːl	intestines		t.ri ¹					tuyviː	
təʔpoːl	mush- room		t.pə ³					t pho	
kəʔ t ^h auːl	door leaf		k.t ^h o ³						kāthoh
kəʔlaiːl	monkey		k.lai ³			kəla ⁱ			
səʔtāːl	crocodile		s ^h .tā ³						
kəʔsiːl	poison		ʔ.she ³						
təʔmaiːl	tail		t.mai ³		mæ ⁴	t ^ə máː ⁱ	tǎ mæ ¹	tmaay	
ʔɜːl hɛʔ	louse		ʔ.hi ¹			he ⁱ ʔ		hoey	
pəʔtsiːl	saliva		m.se ¹						
təʔriːʔ	to write		t.ri ³						tāri
kəʔduiːl	egg		k.dui ¹			k ^ə du ⁱ		kduúy	
pəʔgoːl	to steal		b.ru ³					pviw	māgu
ʔɜːl lɛːl	head		ʔ.lə ¹		ǎ lu ²	lu	lu ²	luú	
pəʔlaiːl	tongue		b.lai ¹		mle ²	p ^ə la ⁱ	pǎlə ⁴	plaáy	
pəʔlɛːl	shoulder		b.lɛ ³			p ^ə le ^{iŋ}	pǎlə ²	pleéng	
məʔləːl	net		b.lo ⁵						māloh
pəʔlɛːl	to vomit		b.lo ⁵			p ^ə le ^u		pleew	
pəʔloːl	salt		b.loi ⁵			p ^ə lo ⁱ			mālawi
pəʔlaũːl	canoe		b.lɔ̃ ⁵	m-lə ⁷				pleewng	
pəʔtɔ̃ːl	cotton		b.lo ⁵					plo	
pəʔʔɔːl	to bake		b.ʔo ³						
pəʔtɔ̃ːl	poor		m.tā ⁵						mātang
pəʔtɛːʔ	to taste		m.dɛ ¹					pte(ng)	māteng
pəʔnuːiːʔ	to laugh		m.nui ¹	m-n̄ ũ		pənu ⁱ ʔ		pnúy	mānui
pəʔsuːiːʔ	to whistle		m.s ^h ui			t ^ə pi			
pəʔ koɛʔ	to fill		m.koi ³						
təʔpiːl	squeeze		t.pi ¹						
təʔbaũːl	beak, bill		t.bɔ̃ ⁵						
təʔmɛːl	to forget		t.mā ⁵						tāmang
təʔ ɣɜːːl	war spear		t.ri ⁵			t ^ə vi			
səʔkoɛʔ	to hug		t.koi ³						tkawy
təʔʔaiːl	crab		t.ʔai ⁵					t ^ə aay	
təʔhɛːl	to stir up		t.he ¹						
təʔhɛːl	scratch		t.hə ⁵			t ^ə py ^h iŋ			
səʔ koɛʔ	fishhook		s.koi ¹					chākawi	
səʔbɔ̃ːl	to help		s ^h .bɔ̃ ³					bawng	
səʔ t ^h ɔ̃ːl	God		s ^h .t ^h ɔ̃ ¹						sāhaw
səʔraːl	teacher		s ^h .re ⁵						
səʔk ^h iːl	sambar		s ^h .k ^h i ¹					skhií	
kəʔdiːʔ	to empty		k.di ¹						
kəʔtuːːl	grandchild		k.də ⁵					ktuu	
kəʔŋniːl	blanket		k.ni ⁵	kñ î					

kəʔ sɿ˨˩˦	to polish		k.s ^h e ¹						
kəʔsaj˨˩˦	elephant		k.s ^h ai ¹			k ^ə s ^h á.i		ksáay	
ŋəʔsaj˨˩˦	prawn		k.s ^h ɔ̃ ¹						
kəʔlɛ˨˩˦	weight		k.li ³						
kəʔlɛ̃˨˩˦	cliff		k.lɛ̃ ¹						
kəʔ ləi˨˩˦	debt		k.lai ¹						kālai
kə̃ʔ ləi˨˩˦	dance		k.la ¹			lān		láng	
ʔa˨˩˦ ɱi˨˩˦	name		ʔ.mĩ ⁵			əmín		amueéng	ämüŋ
ʔəʔ ɱm̃ː˨˩˦	bile, gall		ʔ.mĩ ⁵				mu ¹		
pəʔto˨˩˦	lung		m.to ⁵					ptaaw	
pəʔke˨˩˦	kidney		m.kei ⁵					pkaawy	
ʔəʔ tɕi˨˩˦	echo		ʔ.tɕi ¹						
ʔəʔ tʰi˨˩˦	blood		ʔ.tʰi ³		ǎ t ^ʰ i ²	tʰi	t ^ʰ i ¹	thii	äthi
ʔəʔ ɱi˨˩˦	he, she, it		ʔ.ni ¹			ni			äni
ʔəʔ ɱaj˨˩˦	pus		ʔ.nai ¹		ǎne ²		næe ²	naáy	
ʔəʔɣau˨˩˦	corpse		ʔ.ro ⁵		kã ²		veu ¹	veew	
ʔəʔ ɱaũ˨˩˦	colors		ʔ.rɔ̃ ¹						
ʔəʔsã˨˩˦	hair		ʔ.s ^h ã ⁵		ǎ s ^ʰ ã ²	s ^h ãm	sam ¹	saang	äsang
ʔəʔkɯ˨˩˦	hand		ʔ.kə ³		ǎ kəe ³	ke ^u ?	kíw	keu ⁵	äkuh
ʔəʔʔ kɔ̃˨˩˦	trunk		ʔ.kũ ⁵					kiiwng	äkung
ʔəʔ ɱã˨˩˦	flesh		ʔ.ɱã ⁵		ǎna ²	ɱãn	ɱaẽ ¹		
ʔəʔ hou˨˩˦	bone		ʔ.hu ³		ǎɣu ³	he ^u ?	heu ⁵		ähuh
təʔvɔ̃˨˩˦	stomach			th-po/	ǎyo?		təpəm ⁵		
ʔəʔ si˨˩˦	drug			k-shí					
pəʔso˨˩˦	to wash			m-s̃ hɿ		p ^ə s ^h i?		psiiw	mäse

