Paul Sidwell[†], *Mark Alves*[‡]
[†] University of Sydney; paulsidwell@gmail.com
[‡] Montgomery College; Mark.Alves@montgomerycollege.edu

The Vietic languages: a phylogenetic analysis

We present a new internal classification of the Vietic languages, covering all recognized sub-groups and languages for which we could source suitably comparable data. The analysis reconciles results from two distinct methodologies: (1) computational phylogenetics based on a 116-item basic word list, and (2) historical phonological changes among syllable codas. The analysis identifies five principal sub-groups: Thavung-Malieng (TM), Chut-Arem, Pong-Toum, Cuoi-Tho, and Viet-Muong (VM). While the identification of these subgroups is not original, aspects of the branching structure are. We find that the Vietic tree has a binary structure at the root, splitting between TM and the rest of the branch, which we may call the *Eastern Vietic* clade. Within that Eastern clade, we further find a north-south division, in which the northern VM group is most innovative while those to the south are more conservative. Previously, scholars have tended to group TM and Chut together as archaic lects based on broad typological features of word structure without suggesting specific common phonological innovations to link them. Our findings suggest that syllable restructuring and tonogenesis, which are hallmarks of Vietic languages, have developed largely independently among subgroups, notwithstanding very broad commonalities in phonological change which have contributed to strong areal convergence. Additionally, the Vietic homeland question is briefly discussed, with a northern locus of dispersal arguably supported by old loanword evidence and archaeological data.[1]

Keywords: Vietic languages; Austroasiatic languages; linguistic classification; computational phylogenetics.

1. Background and Scope

Our understanding of the branching structure of the Vietic languages, and the position of Vietnamese in particular (as the focus of most scholarly attention), has developed at a frustratingly slow rate, with published studies reporting results that are inconsistent, incomplete, and lacking in transparency. This sits oddly alongside the rest of Austroasiatic[2], which has seen tremendous progress in classification across most branches in recent years (see the various discussions in Jenny & Sidwell eds. 2014). This progress follows decades of comparative historical work on Austroasiatic, and Vietic has not especially lacked attention in this regard (see references below). The slow development of Vietic classification occurred in part because – for the first half of the 20th century – there was real division among scholars as to whether Vietnamese was even an Austroasiatic language. In part, the uncertainty was a result of the tonal typology of Vietic that resembles Tai and Sinitic languages in its broad character; confusion regarding shared Chinese loanwords in both Tai and Vietic but which were regarded as evi-

[1] We would like to thank Michel Ferlus, Irina Samarina, and Nguyen Huu Hoanh for sharing of significant unpublished Vietic data. Without such data, we would not have been able to reach the extent of analysis and insight provided in this study.

[2] The term Austroasiatic is used here in preference to *Mon-Khmer*, as we do not regard the latter as a true taxon. It is generally recognized these days that Mon-Khmer is merely a label for Austroasiatic languages other than Munda, the latter tending towards more synthetic typology than most other Austroasiatic branches.

dence of common origins of the latter; and the poor knowledge of closely related Vietic languages that are more Austroasiatic in typology[3]. In this study, we have put those issues behind us and have pursued a fresh attack at the problem of Vietic internal classification achieved by combining computational phylogenetics and historical phonology. The result includes a family tree model of Vietic which we assert approximates an accurate phylogenetic structure of nested relations among Vietic sub-groups.

The idea that Vietnamese is ultimately related to Austroasiatic, primarily identified through identification of shared basic vocabulary, was first indicated in the mid-1850s (Logan 1854) under the rubric of the *Mon-Annam* language family, and enjoyed broad support until the early 20th century. Strikingly, Vietnamese was left out of discussion of Austroasiatic in Schmidt's (1906) influential study, and through the early decades of the 20th century, some scholars (e.g. Maspero 1912, Blagden 1913) treated Vietnamese as having origins more closely linked to Tai and Chinese. However, sympathy for the Austroasiatic position was maintained in various quarters (e.g. Pryluski 1924), although views differed over the relative priority of lexicon versus typology for several decades. In hindsight, it is understandable that the strong typological convergence of Viet-Muong, Tai, Hmong-Mien, and Sinitic languages was difficult to recognize for what it was. The theory of areality in linguistics was not well developed, and the idea of languages sharing so much that they converge structurally seemed to be just too much to accept in some quarters. Consequently, the robust isolating morphology, monosyllabicity, and complex tone systems, combined with numerous Chinese loanwords in both Tai and Viet-Muong made it seem more likely that Vietnamese shared a common origin with those languages than the Austroasiatic family with its derivational morphology, largely non-tonal character, and robustly sesquisyllabic word structure.

Nevertheless, in the second half of the 20th century, scholars converged on general recognition of Vietnamese and other Vietic tongues as Austroasiatic languages that have been affected structurally by contact with Tai and Chinese. Scholars were persuaded by multiple converging lines of evidence:

- the persistent identification of Austroasiatic core vocabulary in Vietnamese and the consistent results of lexicostatistical studies (e.g. Thomas and Headley 1970, Huffman 1978, Alves 2017);
- the hypothesis of tonogenesis proposed by Haudricourt (1953, 1954a) and numerous subsequent publications supporting it;
- a clearer understanding of the shared early Sinitic loanwords in Tai and Vietnamese (e.g. Haudricourt 1954b), including many different loanwords, thereby demonstrating such words do not demonstrate Tai-Vietic language contact;
- the recognition that Vietnamese must not be considered in isolation, but rather as one of a number of Vietic languages forming a distinct language branch (cf. various studies cited in this article).

It is in view of the typologically transitional Vietic languages spoken by small, isolated groups that the Austroasiatic affiliation of Vietnamese becomes particularly evident (cf. Alves 2003 for discussion). Beginning with initial observations at the beginning of the 20th century (mainly wordlists gathered by French linguists such as Cadière 1905, Chéon 1907, Guignard 1911, and others), scholars began to become aware of very small highland communities in north-central Vietnam and the Vietnam-Laos border lands, speaking languages that have strong lexical affinities with Vietnamese and Muong, but with strikingly different phonologi-

[3] See also Gage (1985) and Alves (2006) for historical reviews of the place of Vietnamese in Austroasiatic and Sidwell (2009) and Sidwell (2014) for a summary of Vietic phylogenetic studies.

cal and morphological typological features. Some of these languages are non-tonal, while others have simple tone systems. Some have a complex range of onset clusters, while Muong lects have fewer clusters, and Vietnamese has only a single type with medial -w-. All preserve presyllables, which are completely absent from Vietnamese and Muong. In addition, those languages have evidence of Austroasiatic prefixes and infixes, which are mostly fossilized and with little to no productivity. These languages tended to be classified together with the Muong lects as minor languages standing alongside Vietnamese, essentially regarded as uncultivated rural dialects unworthy of the attention properly given to the national tongue. One notable exception is Maspero's (1912) major work on Vietnamese language history, a monograph containing dozens of tables of comparative Vietic, Austroasiatic, Tai, and Chinese lexical data and historical phonological observations. The Vietic data he presented was rich, with Vietnamese, multiple Muong lects, Nguon, Cuoi-Tho lects, and the Sach lect of the Chut group. However, he referred to all of these as «dialects», grouped only by three geographic areas. His primary focus was on Vietnamese history, Sino-Vietnamese, and establishing a connection between Vietnamese and Tai, not addressing the issues of the relationships among the «dialects». Nevertheless, the data he amassed and his insights were influential in later works, including those of Ferlus described below.

In the second half of the 20th century, Vietnamese scholars (e.g. Mạc 1964, Phạm 1975, 1979, etc.) noted that these minor upland languages have some typological traits that are more similar to other Austroasiatic languages than Vietnamese or Muong. Consequently, there emerged a better recognition of the extent and complexity of *Viet-Muong* languages as they were primarily known and referred to in linguistic studies and reference works (e.g. Barker 1963, Barker and Barker 1970, Thompson 1976, Ferlus 1974, and others), or *Vietnamuong* in a few publications.

With the 1973 International Conference on Austroasiatic Linguistics (or ICAAL, hosted by the University of Hawaii), new focus and effort fell upon issues of language phylogeny. In particular, Ferlus provided one of the first substantive lists of Vietic language subgroups (Ferlus 1974: 70–71) and the first phylogenetic proposal for Vietic (Ferlus 1979: 81), under the *Viet-Muong* rubric. A little later Hayes (1982) coined the name *Vietic* for the branch, reserving Viet-Muong for just the northern sub-grouping of Vietnamese, Muong and Nguon. The new name was arguably needed as with increasing documentation and understanding of the smaller upland languages of the group, the Viet-Muong label had become increasingly problematic. Hayes (1992) also offered his own proposal for the internal structure of Vietic based on lexicostatistics and the historical phonology of coda *-h, although his proposals did not find wider support. Other scholars have also offered classificatory schemes (e.g. Chamberlain 2003, Trần 2011, Sidwell 2014), although these studies, and the classification by Ferlus (1979, discussed in more detail below) apply different methodologies, making them difficult to compare or assess.

In this context, we (Sidwell and Alves) decided to take a fresh look at the problem from first principles. The primary challenge previously was the limited amount of directly comparable data which could be brought to bear on the problem. The past century of research on Vietic followed a familiar pattern in language documentation: there were few overriding principles guiding investigators, who opportunistically and/or idiosyncratically collect lexical and textual materials. Each had his own research priorities which overrode elicitation of standard lists or concern for common glossing of basic vocabulary items. The problem arose in part because researchers came from different national backgrounds and worked in different historical periods. For example, word lists in early French language publications have somewhat limited value, such as the Harème lexicon gathered by Riviére (Malglaive 1902: 285–290) which com-

piles words he collected in the field before the École française d'Extrême-Orient devised a standard list. Additionally, Malglaive and others used French-based impressionistic transcription which has to be interpreted with knowledge of both French orthography and the linguistic typology of Southeast Asian languages. More recently, Chamberlain (1998, 2018) has compiled zoological terminology in Vietic languages; while this is especially useful for specific etymological issues, it is difficult to find matching lexical items in sources for related languages compiled by other researchers. These and other issues lead to practical difficulties in the aggregation and semantic alignment of wordlists for comparative purposes, although we are fortunate in that these days, investigators do more or less consistently transcribe their data in broad IPA such that phonetic values can be assumed with reasonable confidence.

Given the real difficulties of aggregating and dealing with the available data, it is understandable that efforts to date have been limited, with scholars generally attempting aggregations based on their own data sets. An example of the latter is Babaev and Samarina (2018) who aggregate 100-word lists for Ruc, Sach, Maleng, Arem, and Kri based on the data compiled over decades of Soviet/Russian-Vietnamese joint expeditions to inform a lexicostatistical matrix. The impressive data aggregation of Ferlus (2007) presents another aspect of the problem: being compiled specifically to identify historical cognates, that collection ignores many early loanwords and low-level lexical innovations that are crucial for statistical analysis, artificially producing gaps in the data that we labor to restore by reference to Ferlus' original source materials. To deal with these issues and make the study manageable, we have attempted to exhaustively aggregate lexicons from across the full diversity of Vietic languages while keeping to a modest list of basic vocabulary items. We assume that this approach maximizes the likelihood of finding items in any list to fill the limited semantic categories, minimizing overall data gaps.

In addition to the aggregation of published and unpublished lexicons, we have also taken advantage of existing comparative studies for recognizing cognates and regularity in correspondences. In this respect, we rely principally on the analytical framework of Ferlus' phonological and lexical reconstruction of proto-Vietic which emerged in a series of works (1991, 1997, 1998, 2007, etc.). Additionally, the wider Austroasiatic context is rendered largely accessible by the publication of Shorto's (2006) *Mon-Khmer Comparative Dictionary*, and the online resources and search tools of the *Mon-Khmer Etymological Dictionary*[4]. Given the access we now have to relevant tools and resources, we took the view that it is now worth pursuing a comprehensive phylogenetic assessment of Vietic by applying parallel computational phylogenetic methods, with related matters of etymology (i.e., retentions, innovations, and loanwords), and analyses of shared historical sound-changes in Vietic, integrating these into a coherent model of phylogeny.

2. Previous studies

Through the first half of the 20th century, researchers were aware of the relationship of Vietnamese with Muong and languages such as Nguon and Sach (e.g. Chéon 1907, Maspero 1912, etc.). In a largely ethnographic book, Vương (1963) assembled core prior studies and raised major ethnohistorical points about Austroasiatic broadly. As part of his supporting data, he presented a substantial list of comparative lexical data (about 100 items in the Appendix to his book) for several major lect groups of Vietic, including Vietnamese, Muong, varieties of Tho, Arem, May, Ruc, and Sach. His discussion did further connect Vietnamese with Austroasiatic,

[4] http://sealang.net/monkhmer

but he did not address the degree of relatedness among these languages. Indeed, sufficient data and analytical tools to properly assess the degree of relatedness among these languages were lacking. Into the 1960s, scholars had no effective sense of nested branching among Vietic languages, and Vietnamese and Muong were conventionally treated by linguists as fair representatives of the branch for comparative and typological purposes. An influential example of such is Thomas and Headley's (1970) lexicostatistical study of Mon-Khmer languages. That study included Vietnamese and Hoabinh Muong lexicons as representatives of Vietic and mentioned as an after-thought that "Arem, Mày (Ruc), and Taypong [...] clearly belong in the Viet-Mương branch" (Ibid. 404). This was consistent with the received view that the branch essentially split between Vietnamese on one side and Muong plus various small languages on the other.

It was not until the 1970s that in a series of articles by Ferlus (1974, 1975, 1979), Vietnamese and Muong were recognised as a sub-group distinct from the other Vietic languages, with the latter forming one or more other clades within Vietic. Subsequently, the first published proposal for a full phylogenetic tree of Vietic with nested sub-branching appeared at the end of that decade (Ferlus 1979, Figure 1).

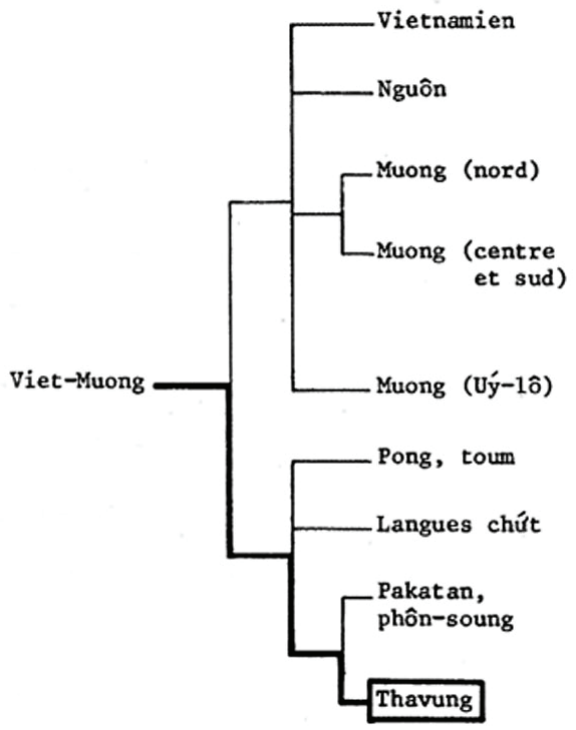

Figure 1. Vietic classification by Ferlus (1979: 81)

Ferlus' (1979) phylogeny split Vietic (*Viet-Muong* in his terminology) into two sub-branches: a northern group of Vietnamese, Nguon, and Muong lects, versus a southern group encompassing all other Vietic languages. The main significance of Ferlus' tree is to make clear that the subgrouping of Vietnamese-Nguon-Muong (or *Viet-Muong* in present day terms) was solidly recognized. This followed nearly two decades of comparative reconstruction focused solely on Viet-Muong by various scholars, in which the relationship with other Vietic lects was left unclear. This separation of Viet-Muong from all other groups lingered in later phylogenetic claims. It is striking that the southern group in this scheme was divided into three equidistant clades (i.e. Pong-Toum, Chut, Pakatan-Thavung), hinting that this was effectively an assumed classification rather than a well justified phylogeny.

2.1. Comparative Reconstruction

Comparative reconstruction is closely related to classification. The method models ancestral states of language at various levels of representation (e.g., segments, morphemes, etc.) and may be dependent on the modelling of tree structure while also informing the tree structure. While this is potentially circular, in practice, one is expected to test various models of branching against hypotheses of shared innovations which are assumed to indicate nodes in the tree. The tree structure that best accommodates the reconstructed changes between the proto-language and daughter languages is then proposed as approximating the real phylogeny of the family.

Serious comparative Austroasiatic studies began with the work of Schmidt in the early 20th century (1903, 1904, 1905), laying serious groundwork for understanding the historical de-

velopment of Austroasiatic phonology, morphology and lexicon. However, Vietnamese or other Vietic data did not figure in that work until Haudricourt (1953, 1954a) showed the regular correspondences between Vietnamese tones and Khmu syllable structure. This permitted the formulation of sound laws that map the development of Vietnamese morphemes from originally Austroasiatic roots. By the 1970s, the historical lexicon and phonology was known well enough by concerned scholars that many Vietnamese roots were uncontroversially linked to proto-Austroasiatic forms (e.g. *nước* 'water' < *ɗa:k, *chó* 'dog' < *cɔʔ, etc.), although this was not reflected well in formal publications of the time.[5]

From the early 20th century, the École française d'Extrême-Orient in Hanoi conducted language survey work, with scholars collecting lexicons on the bases of standard elicitation lists, but these lists were limited in extent, and data was not transcribed in structuralist-linguistic terms. The situation began to change from the late 1950s onward as linguists trained in phonemics and structuralist theory began collecting field data on Muong and other minority languages of Vietnam, yielding lexicons readily useful for comparative analyses. This included foreign scholars in South Vietnam who worked with Muong speakers who had relocated from the north after partition, in addition to a modest field linguistic tradition within the Socialist Republic of Vietnam. Consequently, comparative studies of Viet-Muong became viable, and results began to appear from the mid-1960s (Barker 1966, Barker & Barker 1970). Significantly, many of the smaller Vietic speaker groups, such as the Chut, live in highland and border areas which were mostly inaccessible to scholars due to geographical and war-time conditions, a situation that favored the focus on Viet-Muong.

By the mid-1970s, scholars were incorporating data from other Vietic groups, variously using both older and newer sources, working towards a comprehensive Vietic reconstruction. The most important in this tradition is the work of Ferlus extending from the 1970s to the present day. Ferlus (1975) effectively established the paradigm for comparative Vietic in a 33-page paper that reconstructs Vietic phonology using comparative, philological, and loanword evidence. That work is weighted heavily towards Vietnamese and Muong, and it incorporates limited data from Vietic minority languages (mainly Thavung, Sach, and Nguon). The 1970s also saw the beginning of a long-term cooperation between Soviet and Vietnamese scholars, and in particular, Sokolovskaya took an interest in Vietic reconstruction. She marshalled 27 published sources to offer a reconstruction of 600-plus proto-Vietic roots (Sokolovskaya 1978) and later prepared a more substantial comparative dataset, although this was lost after her untimely passing in the 1980s.[6]

Other works of comparative-historical reconstruction related directly to Vietic have appeared, and more than a dozen are referenced and characterized in Table 1. One can see from the table that these are quite diverse, with works focusing on different aspects of phonology or level of structure, with a slow progression towards a coherent and comprehensive view of Vietic history only emerging from decades of investigations. The scholar to whom we are most indebted is Michel Ferlus, who has done most to aggregate, integrate and analytically assess the scholarly history and the ongoing emerging breadth of data to reconstruct proto-Vietic.

By the early 2000s, Ferlus had released an extensive and well justified proto-Vietic reconstruction which establishes phonological correspondences that allow us to account for all ap-

[5] From the 1973 ICAAL meeting onward scholars were circulating manuscripts and sharing conference presentations which featured well developed models of Austroasiatic phonology and many reconstructed forms. However, much of this work was not published, or only published decades later. Hints of this can be seen in publications such as Shorto (1976), which foreshadowed his posthumous (2006) *Mon-Khmer Comparative Dictionary*.

[6] A substantial manuscript was passed to the Mon-Khmer Studies editors but did not proceed further. The authors would appreciate any leads that would assist in tracking down a copy of this missing manuscript.

parent sub-groups within the branch. However, this reconstruction is not dependent on a model of nested branching within Vietic; morphemes in each language or sub-group are directly derived from proto-Vietic forms. Thus, the reconstruction is not reliant on sequencing shared innovations among sub-groups. Effectively, each subgroup is treated as direct descendants of proto-Vietic, with their various patterns of lexicon and phonology reflecting coincidental parallels. While this may imply a flat rake-like tree structure, actually this is not the case, and nested relations were simply not treated as analytically relevant above the sub-group level. We have no problem as such with Ferlus' approach as a means of coming to a phonological reconstruction, but it was clearly not his priority that the reconstruction yield a detailed family tree. That said, Ferlus' organization of lexicon and phonological correspondences provides a solid basis for informing our methods and results.

Authors	Years	Level	Notes
Barker	1966	VM	Tonal correspondences demonstrated with some 600 VM lexical comparisons
Barker & Barker	1970	VM	Coda and vowel reconstructions supported by 210 VM lexical comparisons
Ferlus	1975, 1982	Vietic	Reconstructs proto-consonants and tones, and their development to modern Vietnamese values
Thompson	1976	VM	Reconstructs approximately 700 Proto-VM lexemes based on Vietnamese and three varieties of Muong
Sokolovskaya	1978	Vietic	Reconstruction of over 600 Proto-Vietic lexemes supported by data from 27 sources. Proto-Vietic treated as non-tonal, non-glottalized
Ferlus	1991, 1997	Vietic	Reconstructs Proto-Vietic vowels
Ferlus	1992	Vietic	Reconstructs Proto-Vietic onset consonants, connecting modern Vietnamese to Proto-Vietic and Chinese loanwords
Nguyen T. C.	1995	Vietic	Reconstructs Proto-Vietic inventories of onsets, vowels, and codas (but non-tonal), tracing modern Vietnamese segments to Proto-Vietic and two stages of Chinese loanwords
Ferlus	1998, 1999, 2004	Vietic	Reconstructs Vietic tonal history
Nguyen V. T.	2005	VM	Reconstructs VM onsets, vowels, and tones, in addition to one hundred lexemes based on Hanoi Vietnamese, Nghe Tinh (Northern) Vietnamese, and 30 Muong doculects
Shorto	2006	AA	Incorporated Viet. and Muong in his reconstruction (including tone categories), tabling consonant correspondences.
Ferlus	2007	Vietic	Reconstructs more than 1,200 Proto-Vietic lexemes
Ferlus	2014	Vietic	Provides reconstructions of onsets, vowels, and codas in Proto-Vietic (which he calls Proto-Viet-Muong), notes on tonogenesis, and proto-language vowels for his proposed Proto-Pong-Cuoi

Table 1. Comparative reconstructions related to Vietic

A specific example of this problem can be seen in Hayes' (1992: 216) attempt to invoke the loss and/or retention of coda -h for subgrouping among Thavung, Ruc, Muong and Vietnamese. Relying solely on the presence or absence such a segment is quite problematic; in the context of the areal trend to tonicity, it is particularly important to recognise that one cannot just treat coda [-h] as straightforward segment to be retained or lost as is, but whether it remains

present rephonologized as a tonal feature (e.g. phonation features of a tone), and this may or may not be readily recognized in data transcription. It is our considered view that the problem of justifying clustering/splitting arguments in this genetic and areal context is best addressed by identifying multiple types of evidence of linguistic innovations and retentions that align consistently to support a model of phylogeny. In the history of Vietic classifications, authors have relied upon diverse methods and forms of evidence, but none appear to have been able to assemble a sufficient range of types of data to achieve a coherent synthesis.

2.2. Classification

As we have noted above, comparative studies of Vietic are now quite advanced, yet the results have not been successfully applied to the classification problem. This is not to say that classification has been ignored, on the contrary, but our reading of the situation is that scholars have rather relied on a range of rather different, and to some extent incompatible, methodologies when offering their classifications. These include phonological/typological considerations, geographic distribution, and/or lexical data (i.e. lexical isoglosses or lexicostatistics). The approaches are sometimes explicitly stated, while in some cases, the methods must be inferred from the method of presentation. Table 2 shows our characterization of how scholars have prioritized different aspects of language to inform their classifications of Vietic and their resulting hypotheses.

Author	Year	Approach	Key Insights
Ferlus	1979: 81	Typological	Two main branches 1. Viet-Muong (neutral regarding Nguon) 2. Pong & Chut & Thavung-Malieng
Ferlus	1989–90: 52–53	Typological and geographical	Several main groups listed without tree structure 1. Vietnamese 2. Muong 3. Pong-Cuoi 4. Western archaic (Thavung, Pakatan) 5. Eastern archaic (Chut, Arem, Malieng)
Diffloth	1989 (via Chazée 1999)	Typological	Two main branches 1. Viet-Muong & Pong-Cuoi 2. Archaic languages
Hayes	1992: 220–221	Geographical organization, lexicostatistics, and sound change (final *-h)	Three main branches 1. West Vietic (Thavung lects, Pakatan, others) 2. Central Vietic (Pong-Cuoi) 3. East Vietic: Viet-Muong (Nguon with Muong) & Chut
Nguyễn, T. C.	1995: 12	Typological	Two main branches 1. Viet-Muong 2. Pong-Cuoi & Chut & Thavung-Malieng (adopting Ferlus's 1979 model)
Chamberlain	1998: 106	Ethnozoological lexical data and inferrable geographical organization	Two main branches 1. Viet-Muong & Pong-Cuoi & Chut 2. Thavung-Malieng

Table 2. Previous classification of Vietic languages

Table 2. Previous classification of Vietic languages (continued)

Author	Year	Approach	Key Insights
Nguyễn H. H.	1999	Lexicostatistical	Several inferable lect groups 1. Viet-Nguon-Muong 2. Phong-Poong-Tum-Ly Ha 3. Chut 4. Arem 5. Malieng-Kri-Pakatan 6. Thavung-Phon Sung
Chamberlain	2003	Ethnozoological lexical data and inferable geographical organization	Several branches (citing Diffloth 2001) 1. Viet-Muong 2. Pong-Cuoi 3. Thavung 4. Chut-Malieng 5. Malang 6. Kri
Peiros	2004	Lexicostatistical	Several main branches (branching downward from Thavung) 1. Thavung 2. Malieng 3. Ruc-Arem 4. Pong-Cuoi 5. Viet-Muong
Sidwell	2015: 205	Synthesis of multiple studies	Three main branches 1. Viet-Muong 2. Pong-Cuoi 3. Chut (archaic languages divided into groups to the East and West)
Chamberlain	2018	Lexical (fauna terms)	Two main branches 1. Viet-Muong & Pong-Chut 2. Thavung-Malieng lects & Kri-Mlengbrou
Trần	2018: 61	Typological	Two main branches 1. (monosyllabic) Viet-Muong & Cuoi 2. (bisyllabic) All others (Pong, Arem, Chut, Malieng, Aheu)

It is clear from all the literature that the Viet-Muong sub-branch within Vietic is uncontroversial. It consists of Vietnamese, multiple varieties of what are collectively called Muong (30 Muong lects have been documented in Nguyễn V. T. 2005), and Nguon, of which there are at least two documented varieties. Significant phonological diversity is found among Muong lects, although lexically they are rather homogenous, so it is not quite clear to what extent they reflect distinct languages or are better regarded as a close dialect chain or linkage. Recently, Phan (2012) presents phonological evidence and arguments that Muong is a paraphyletic taxon: based on a lack of clear phonological patterning, the Muong lects do not sit unified in opposition to Vietnamese within the tree, but all of them and Vietnamese (and potentially Nguon) descend from proto-VM more or less equally. Consequently, we take as a given that Vietnamese, Nguon, and the Muong lects form a unified *Viet-Muong* node somewhere within Vietic, and do not investigate the specific question of Viet-Muong internal branching further

as it is assumed that this is not relevant to the higher tree structure. We have utilized Vietnamese, Muong, and Nguon data in our analysis, and our results also confirm this grouping. Finer clarification of the historical linguistic relationships within the Viet-Muong sub-branch will require an assembling and analysis of phonological and lexical data of the few dozen Muong doculects and thus is beyond the capacity of the current study.

The approaches that have been used to group/divide the Vietic languages can be summarized as follows, with our brief assessments of their consequences:

- Phonological: Hayes (1992: 216), based on Ferlus (1974, 1975), posits limited sub-branching based on the retention or loss of proto-Vietic *-h. Noting various shared phonological retentions, Nguyễn T. C. (1995: 257) acknowledged that shared phonological innovations to determine sub-branching among the smaller Vietic lect groups were insufficient to provide finer phylogenetic distinctions.
- Typological: Restructured monosyllabic Viet-Muong languages are most often considered to be sub-grouped in contrast with the more conservative sesquisyllabic Chut, Arem, Thavung, Malieng/Pakatan, Kri, and others. Also, both the Viet-Muong and Cuoi groups have complex systems of tonal phonemes, while others have smaller tonal inventories or none. However, this aspect is challenging to implement for phylogenetic purposes as the timing of the development of tonogenesis may vary among the sub-branches.
- Lexical: Lexical isoglosses can be referred to in considering sub-grouping membership. Chamberlain (2003, 2018) refers solely to fauna terms, utilizing a type of isogloss approach, to suggest detailed phylogenetic trees of Vietic. However, his methodology and assumptions in use of such data to make these determinations are not clear, making them difficult to assess.
- Geographic: Hayes (1992) relies at least partly on geographic grouping as suggested by regional labels (i.e. east, west, and central). Ferlus in his various works refers to a broad north-south division. While geography does sometimes correspond to phonological, lexical, and/or typological factors, we assume linguistic factors must be considered first, with geography as a secondary indicator of language history (e.g. population movements and language contact).

The twelve publications in Table 2 each show different versions of a Vietic Phylogenetic structure, with overlap in various cases, but otherwise distinct claims in each proposal. A recurring hypothesized division is between Viet-Muong versus all else (e.g. Ferlus 1979, Nguyễn T. C. 1995). This classificatory division appears to be influenced by the typological distinction of the strictly monosyllabic Viet-Muong languages versus the sesquisyllabic structure of other Vietic languages. However, the degree of variation of the dozen approaches makes it clear that no consistent position has dominated. Key to the discussion is the position of Pong-Toum and Cuoi, which have variously been grouped with VM, with sesquisyllabic lects, or as a distinct Pong-Cuoi clade.

Contradictory claims are evident: some proposals put the sesquisyllabic Chut sub-group into a closer relationship to Viet-Muong than other archaic Vietic lects, while other proposals suggest grouping Chut more directly with other archaic varieties. A north-south division seems to emerge at the extremes of Viet-Muong in contrast with the archaic Vietic languages, but there is otherwise little consistency among the posited nested branchings. All things being equal, we might expect that classifications based on different data types and methodologies would converge on common points that underlie the real shape of the family tree. However, beyond identifying the Viet-Muong node, studies have offered strongly diverse results, indicating fundamental problems with the nature of the evidence used and/or the assump-

tions underlying its analysis. The variation in methodologies also means that proposals lack comparability, which we attempt to overcome in this study by the methods outlined in Section 3.

3. Methods and data

Given the above, we decided to proceed with a fresh analysis on first principles, rather than attempt to revise or repair any existing analysis. For this, we compiled a 116-item basic word list (described below), extracting the data from the most complete aggregation of sources available to us in some kind of readily processed form. The initial method is the same as applied in classical lexicostatistical or glottochronological studies, with all items scored for their cognate values, and apparent and likely loans (primarily from Chinese and Tai) identified and scored as zero items, according to our assessment as experienced comparative Austroasiatic specialists with understanding of phonology and etymology in the region. SplitsTree (4.14.2 for Windows) was used to generate neighbor nets and phylograms, and these were further assessed by comparing the computed pattern of branching nodes with shared phonological changes implied by comparative reconstruction. The data we considered was not limited to the 116-word list already mentioned; a more extensive lexical database was built as follows: the data from Ferlus (2007, available online at sealang.net/monkhmer) was laid out in a spreadsheet, and then augmented by integrating additional wordlists, aligning items with Ferlus' etymologies. The advantage of using Ferlus' analysis as an organising principle is that we could rapidly identify reflexes of specific etyma, without having to accept all of Ferlus' reconstructions. Columns with our spreadsheet were created for sorting etymologies by onsets, nuclei, and codas, so we could rapidly identify multiple members and reflexes of specific proto-segments and thus discuss shared changes/innovations. The result is a dynamic working file of Vietic lexical data and phonological analyses which not only informed this project but is also the basis for a larger effort to revise and extend Vietic comparative reconstruction.

The SplitsTree software used is widely applied in phylogenetic studies in fields such as genetics, and is one of a number of programs that linguists have applied in recent years to basic vocabulary lists to test phylogenies (see Greenhill et al. 2020 for a broad discussion about about uses and limits of phylogenetics in respect of cultural traits). We selected splitstree for its ease of use: we do not have expertise in computer programming, but we have ensured that the analysis is transparent and repeatable. We invite other researchers to download our 116-word dataset and experiment with other software packages as they feel appropriate. Our use of lexical data is strictly predicated on our understanding of inheritance versus borrowing, based on the comparative reconstruction literature reviewed above, and our own knowledge gained over years of research on languages in the region. Lexical matches are scored as cognate (i.e. inherited from a common proto-form) and not for phonological similarity (as in approaches such as the ASJP (Automated Similarity Judgment Program)[7] and allied methods). The splitsTree approach does not simply reproduce traditional lexicostatistics. Rather, it is Bayesian, running an algorithm to identify all patterns of shared cognates, and recalculating repeatedly through all 116 entries, eventually generating tree structure on the basis of bottom-up hierarchical clustering. The effect is to force grouping based on lexical retention, while division is indicated by semantic change / lexical replacement.

[7] http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm

3.1. Vietic Lects Used

Lexical data for 29 lects were assembled, 12 of which were already aggregated in the Ferlus 2007 file and the later Ferlus 2017 aggregation of Pong lects, and the remainder added from other published and unpublished sources.[8] The full listing follows:

Languages/Lectal groups	**Data Sources**
Vietnamese (Hanoi)	Standard dictionaries
Muong lects	
Muong Hoa Binh	Ferlus 2007
Muong Son La	Ferlus 2007
Muong Muong Thanh Hoa	Ferlus 2007
Muong Bi	Nguyễn V. K. et al. 2002
Nguon lects	
Nguon Co Liem	Nguyễn P. P. 1997
Nguon Yen Tho	Nguyễn P. P. 1997
Cuoi lects	
Tho (Cuoi Cham)	Ferlus 2007
Tho (Lang Lo)	Ferlus 2007
Cuoi Thai Hoa	Nguyễn H. H., n.d.
Cuoi Tan Hop	Nguyễn H. H., n.d.
Phong lects	
Phong	Ferlus 2007, 2017
Toum	Ferlus 2007, 2017
Liha	Ferlus 2007, 2017
Liha	Peiros 2004 (data sourced from Nguyễn Văn Lợi)
Chut lects	
Sach	Ferlus 2007
Ruc	Ferlus 2007; Nguyen, Tran & Ferlus 1998; Nguyễn V. L. 1993
May	Babaev & Samarina 2018
Arem	Kasuga 2008
Other archaic lects	
Thavung lects	
Thavung	Ferlus 1979
Thavung Phon Soung	Ferlus 2007
So Thavung	Premsrirat 2000
Ahoe	Enfield 2011
Malieng lects	
Malieng (Quang Binh province)	Samarina n.d.
Malieng (Ha Tinh province)	Samarina n.d.
Malang	Ferlus 1997
Malang Pakatan	Nguyễn V. L. and Nguyễn H. H. (2001)
Malieng Bro	Ferlus 1992
Kri (Ha Tinh Province)	Samarina n.d.
Kri Phoong	Nguyễn V. L. and Nguyễn H. H. (2019)

[8] We take note that an anonymous reviewer correctly points out that extensive relevant lexicons were collected by the late Nguyễn Văn Lợi, and use of these would materially contribute to this study. Indeed, we were able to obtain two such word lists of Malang Pakatan and Kri Phoong, but other materials were not made available to us.

Some of the sources (particularly for Aheu-Ahlao and Harème) lack sufficient lexical data to provide confidence in their use for statistical analyses (i.e. provide less than 80% coverage of the lexical entries). Their membership in their respective lect groups in such cases is determined by key isoglosses, as described below.

1) Harème (listed in Ferlus (1996: 12) as part of the Malieng group) data includes only 50 items in our 116-word list. Nonetheless, within the Harème list, there are notable alignments with lexical items seen uniquely in Malieng, Malang, and Kri ('big/large', 'black,' 'cloud,' 'drink,' 'earth/soil,' 'mountain,' 'rain', etc.), thereby supporting the original claim of affiliation in this group. Ferlus (2014: 1) notes earlier confusion of Harème with Arem, and similarly provides evidence in the numerals 1 to 10 related it with Maleng Bro.

2) Available Aheu-Ahlao data is particularly limited: Chamberlain's study focuses on faunal terminology so there are hardly any useful lexical alignments with our 116-item list. However, Chamberlain's data do include etyma noted in other studies to be members of the Thavung group, indicating that this lect is closely related to the Thavung lects in our database. Our working assumption is that Aheu-Ahlao straightforwardly pairs with Thavung. Selected examples of Aheu-Ahlao-Thavung isoglosses are listed below:

- 'mouse' [ʔiik] occurs in all Thavung varieties, and 'porcupine' is seen in So Thavung [ɲiː^{43}] and Ahoe-Ahlao [jiːj], with a shared loss of Vietic *-m (Proto-Vietic *k-ɲiːmʔ).
- 'buffalo' So Thavung [k^{h}uaj^{43}] and Ahoe [khwaay] are like borrowed from Lao k^{h}wáːj 'buffalo'.
- 'duck' So Thavung [ʔa^{33}|tʌ̲ː^{32}] and Ahoe [ʔatʌː] and Ahlao [ʔatɛː] are likely shared loanwords from Katuic (e.g. Proto-Katuic *ʔadaa).

Beyond the issue of Harème and Aheu-Ahlao being absent from the computational analysis, we have taken the view that the data we assembled for this experiment are reasonably representative of the diversity within Vietic, sufficient to justify proceeding in the manner we have done, and any limitations arising from these circumstances do not constitute fair reason not to proceed or report as we have done. On the contrary, we assert that the present study sets a standard for transparency and testability for Austroasiatic phylogenetic studies that is unsurpassed in print.

The creation of the 116-item list was the outcome of beginning with the Swadesh 100 and 200 lists and reconciling these with the available data with the aim of achieving at least 80% coverage for each lect in the analysis. Procedurally, we went forward as follows: sources were selected and lexicons aggregated in a spreadsheet, with rows identified with Swadesh 100 and 200 items, subject to semantic and phonological adjustments as we judged necessary. For most of the languages, full coverage of the Swadesh 100 categories was not possible, with 20 or more gaps being common. Some 40 additional categories were added from the Swadesh 200 list, based on the 40 best represented items in the aggregated data, seeking to achieve a 120-item list with at least 100 items coverage for all lects. Ultimately, achieving a 120-item list was quite problematic given internal problems with the Swadesh categories, which do not always map neatly to the semantics of mainland Southeast Asian languages, and we settled on 116 items. A number of items that we investigated had to be excluded as the data included confounding and/or limited evidence; these items and reasons for their exclusion are listed in Appendix 1.

The spreadsheet we created, with cognate assignments, and some etymological notes, and the derived nexus file, were converted into a PDF file and archived with zenodo.org.[9] The nexus file was input to SplitsTree running default settings for generating neighbor-nets and phylograms. Phylograms were created in both UPGMA (unweighted pair group method with

[9] See: https://zenodo.org/record/5263195 (DOI: 10.5281/zenodo.5263195).

arithmetic mean) and NJ (neighbor-joining) analyses, producing similar branching results (discussed below). Jahai (Aslian) and Khmu (Khmuic) were included as out-groups to root the Vietic tree, with Jahai set in SplitsTree as the first out-group, consistent with received views on Austroasiatic classification.

4. Primary results

4.1. Neighbor-Net

The neighbor-net (Figure 2) results are regarded as important for identifying low-level subgroups and to provide indications of whether the tree-structure is strongly branched or ambiguous in some way. The latter may point to contact interference or errors in coding cognates. In this case our interpretation of the neighbor-net is that the results are fairly clean, pointing to six groupings as follows: (1) Viet-Muong, (2) Cuoi-Tho, (3) Pong-Toum, (4) Chut (Ruc, May, Sach), (5) Arem, and (6) Thavung-Malieng. The precise status of Arem is unclear; while there are weak indications that it groups with the Chut lects, this could reflect a couple of unknown borrowings in the dataset, or could be a statistical artifact of missing items. Otherwise, the apparent groupings are firm.

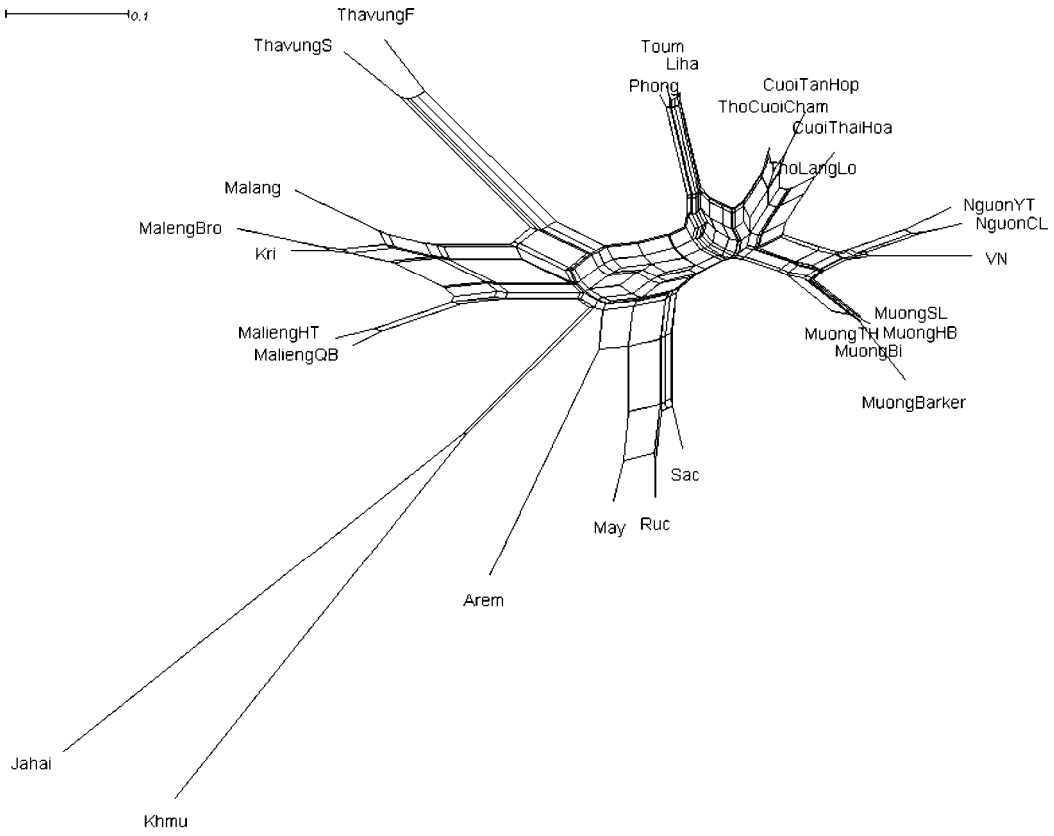

Figure 2. Neighbor-Net results for 28 Vietic doculects.

4.2. Phylogram

The UPGMA phylogram generated with SplitsTree is reproduced in Figure 3; it is strongly consistent with the neighbor-net results, while adding some clarity in terms of the apparent

nested branching. We also generated NJ and BioNJ phylograms, but effectively the same results were replicated, so those trees are not given here for space considerations.

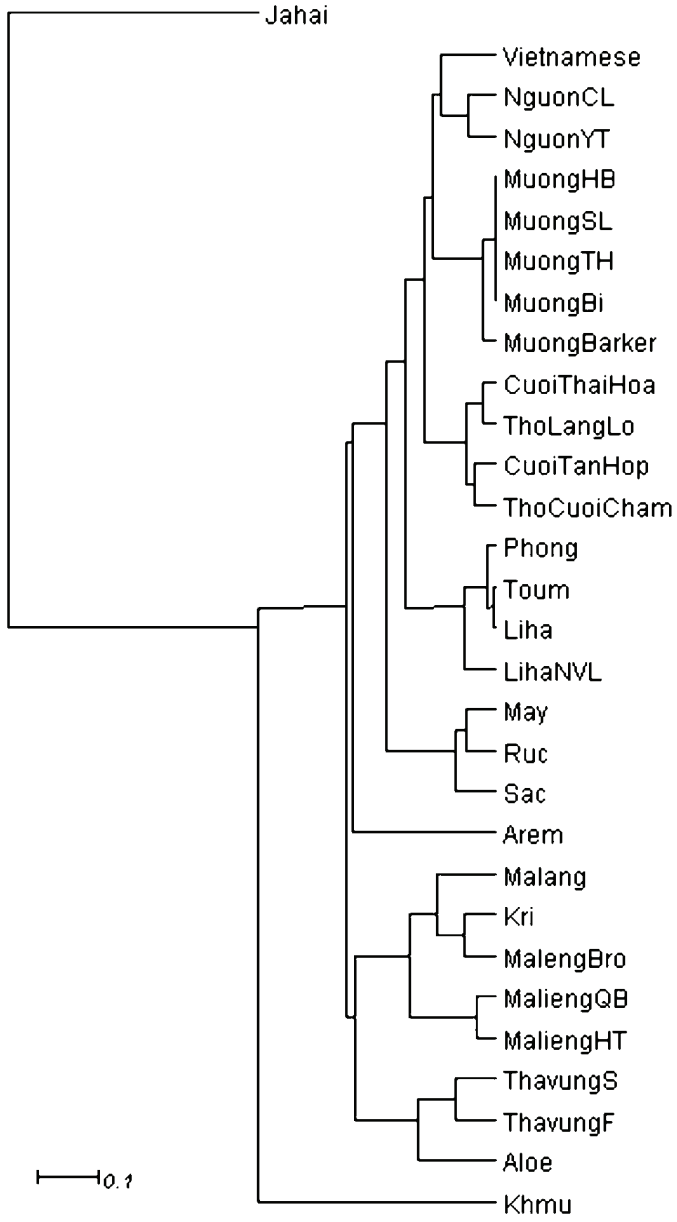

Figure 3. Phylogram results for 29 Vietic doculects, UPGMA.

4.3 Historical phonology and classification

If our computational results do approximate the real tree structure of Vietic, there ought to be correlations in terms of phonological developments reconstructable for the indicated branchings, in addition to typological features, such as syllable structure and suprasegmental features, that have been noted in specific groups, as discussed above. The distinction of the presence or absence of sesquisyllables among Vietic languages is a prominent feature by which relatedness among the sub-branches can potentially be evaluated, and previous approaches have used this as a key criterion. However, based on this feature alone, a finer analysis cannot be achieved, and segmental features must be considered.

Unfortunately, the issue of identifying shared or contrasting segmental developments in Vietic is quite problematic. Ferlus' (2007) reconstruction is a starting point, and we largely

concur with his assignment of proto-language values to segments. As is often the case when applying the comparative method, it is apparent that segmental correspondences fall broadly into two types: (1) a core of relatively regular correspondences each attested in multiple etymologies, and (2) a mass of unique and/or apparently irregular correspondences that are difficult to explain or have competing explanations which are difficult to discriminate.[10] Confining our consideration to the first type, we have found no strong bases to identify nested branching relations in terms of syllable onsets or vowels. Each of the six identified subgroups mostly show onset and vowel reflexes that are variously quite stable, or they show variation within each subgroup that does not link across multiple sub-groups.

proto-Vietic	*-**h**	*-**s**	*-**r**	*-**l**
Viet-Muong	-ø	***-lh >*** -j Viet -j/-c/-n Muong -t/-n Nguon	***-l** >* -j Viet -j/-l/-n[11] Muong -n Nguon	***-l** >* -j Viet -l/-n/-ɯ Muong -n Nguon
Cuoi-Tho	-ø	***-l*** > -l/-n/-j	***-l*** > -l/-n	***-l*** > -l/-n/-ø
Pong-Toum	-ø [-h only in loans]	***-c*** > -t/-c/-k	***-l*** > -l/-n	***-l*** > -l/-n
Chut	**-h**	***-lh*** > -h/-l/-lh/-rh	***-l*** > -l	***-l*** > -l
Arem	**-h**	***-h*** > -h	***-l*** > -l	***-l*** > -l
Thavung-Malieng	**-h/-ʔ**[12]	***-s*** > -s/-j/-jh/-ɥh/-jʔ	***-r*** < -r/-ɥ/-l/-n	***-l*** > -l/-n

Table 3. Vietic reflexes of codas *-h, *-s, *-r, *-l

However, close examination of coda correspondences allows us to identify the evolution of proto-Vietic *-h, *-s[13], *-r, *-l as potentially relevant to this question. There is considerable variety—and seeming instability—both among and within the lectal groups in the modern reflexes of these sounds, including a range of segments, phonation features, and tones. The data are summarised in Table 3; we include our branch-level reconstructions to highlight the common patterns.

The reconstruction indicates that Vietic *-r and *-l remained in contrast within T-M while merging to /-l/ in the rest of Vietic (notwithstanding some later mergers with -l in some T-M lects). This is consistent with the hypothesis that T-M branched directly from the proto-Vietic node and is the most conservative subgroup in respect of this segment.

Proto-Vietic *-h was already discussed as an indicator of branching structure by Hayes (1992), but we now have a clearer understanding of its significance. Hayes' phylogram is simple (see Figure 4), treating only Thavung (TV), Ruc (RU), Muong (MU), and Vietnamese (VN). He found that the loss of *-h is shared within Viet-Muong, distinguishing it from Ruc and

[10] Shorto (1976) discusses this problem at length in relation to Austroasiatic reconstruction, although we do not necessarily endorse specific proposals for dealing with these issues.

[11] One example of -n was found by us in one Muong word list.

[12] Some apparent instances of -ʔ reflexes of *-h are noted for Maleng. Otherwise, -h is general in TM lects.

[13] The reconstruction of *-s is Ferlus' reconstructed value. We take the view that this segment was probably a laminal fricative with post-alveolar or pre-palatal constriction given the tendency for palatal reflexes in some Vietic lects, but retain the *-s formalism for consistency and simplicity.

Thavung. The more complete data we have aggregated indicates that the loss of *-h is also shared with Pong-Toum and Cuoi-Tho. This is potentially a shared innovation at the VM-P-C node in the tree, yet it may not be so straightforward. Vietic *-h was not simply lost as a segment, but rephonologised as glottal tension in the syllable nucleus, yielding the *hỏi* and *ngã* (traditionally considered Category-C tones) tones of Vietnamese and their equivalents in other lects. While modern Vietnamese dialects vary in phonetic realizations of tones, in addition to contour, it is common throughout regions of Vietnam for these two tones to exhibit phonation features (i.e., glottalization or breathiness, again varying according to local varieties) that are likely residual evidence of the earlier segments. Such is the case as well in Cuoi in Tan Ki district, in which this tone category is glottalized (Nguyễn and Nguyễn 2019:lxii).

Effectively the VM, PT, and CT groups share the rephonologization of *-h as a Category-C tone. However, it is not presently determined whether these tonal developments were one or multiple independent events. This is a complex topic whose details are beyond the scope of this study, but for the present purposes, we will suppose that the single rephonologization of *-h at the VM-PT-CT node is likely and represents a relevant finding which is consistent with the identification of this node in the computational analysis.

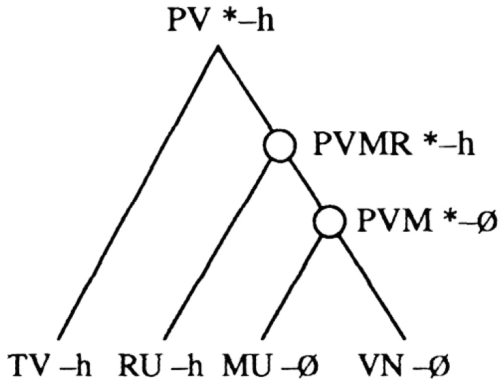

Figure 4. Hayes (1992: 216) phylogram based on development of Vietic *-h.

Also relevant is the development of Vietic *-s. As indicated in Table 3, each subgroup has reflexes of *-s which appear to be distinct, and in most cases show multiple secondary developments within each group. Areally, coda [-s] is unstable, and across multiple language families shifts to [-t] and [-h] are not uncommon.[14] The shift of [-s] to [-l] and [-n] are also unproblematic; the shift of [-s] to [-l] is assumed to be via a voiceless lateral, written *lh* here. A subsequent change of [-l] to [-n] is similarly common in Tai,[15] and it is unsurprising that it happened in Vietic considering the shared typological and historical context.

The phonological reconstruction indicates that, like *-r, Vietic *-s was unchanged in proto-Thavung-Malieng and later developed a range of mostly voiceless approximant articulations. As other subgroups branched off on the Eastern side of Vietic, the *-s underwent independent developments in Arem, Chut, and Pong-Toum. Given the similarity of the reflexes of *-s to the reflexes of *-l in Viet-Muong, and the obvious merger to *-l in Cuoi-Tho, we propose that *-s shifted to *-lh in proto-Cuoi-Tho-Viet-Muong, becoming voiced in Cuoi-Tho. The various /-t, -c/ reflexes in Pong-Toum clearly suggest a sequence *-l > *-lh > *-ç > *-c with lateral friction shifting to palatal friction before hardening to a stop coda.

[14] For example, in Austroasiatic, *rma:s ‘rhinoceros’ > Sre *rəmis*, Khmer *rəmiəh*; *ʔas ‘to swell’ > Sre *ʔas*, Khasi *ʔa:t*; *ris 'root' > Mlabri *rɛ:lh*, Chong *rè:t*, Car *ɽeh* (reconstructions from Shorto 2006).

[15] For example, Proto-Tai *bil^{A} ‘to fly’ > Siamese *bin^{A1}* (Pittayaporn 2009).

Overall, we can say that our analysis of the Vietic codas *-h, *-s, *-r, *-l is readily reconciled with our computational phylogenetic results, and we take this as pleasingly confirmatory. Below, we first assess the resulting individual lectal groups and characterize the overall phylogenetic findings.

4.4. Summary findings

Figure 5 provides a schematic representation of our overall findings. The computational and phonological results converge on the identification of five subgroups in the branching configuration indicated in the figure. We have a high level of confidence that this is a good approximation of the real family history of Vietic. We are not in a position to make strong claims about the internal configurations of each sub-group; generally, the lexical differences between individual sub-group members are moderate and may not be strongly significant in the context of the 116-item list (with the exception of Arem) although other factors may also be brought to bear in discussing the subgroup internal relations. Below we discuss each subgroup separately.

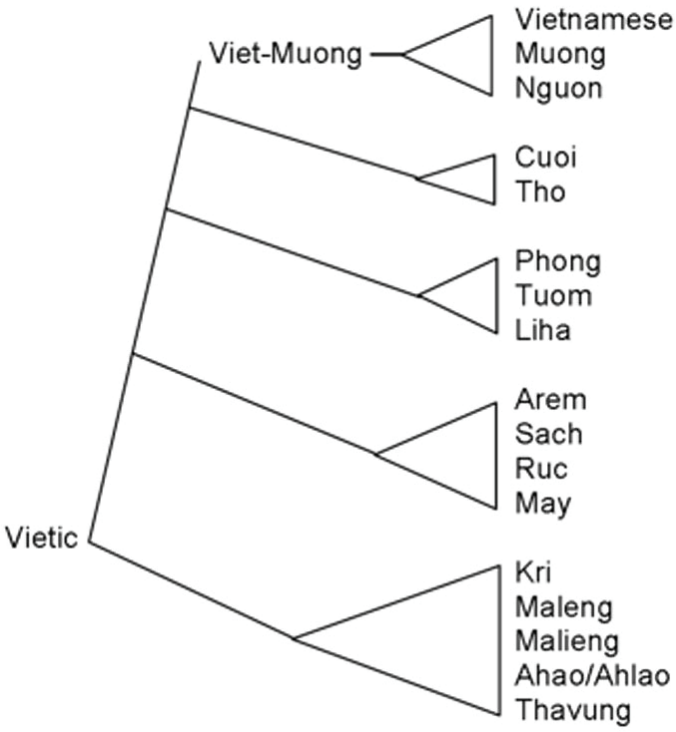

Figure 5. Vietic classification proposed in this study.

Viet-Muong (including Nguon): Our results confirm the general view that the Vietnamese, Muong, and Nguon form a coherent VM sub-branch, what can considered the Northeast clade of Vietic. While Vietnamese as a national language is spoken throughout Vietnam, Muong lects are largely concentrated in northern Vietnam. Nguon is spoken in north-central Vietnam in Quang Binh Province close to the Chut group. Interestingly, Vietnamese is grouped closer to Nguon than to Muong in the tree, with the latter so tightly clustered that the Muong lects could be regarded as one language lexically. This is contra Phan's (2012) results based on phonological changes, but unremarkable given their real lexical uniformity. The present results are based on counting cognates regardless of phonological changes, so as a lexical result this should not be controversial, and we note again that the question of VM internal

structure has no bearing on the larger question of the higher order branchings. We note here also that in addition to the grouping based on lexical data, Vietnamese, Muong, and Nguon are typologically unified and distinct from most other Vietic groups. VM languages have complex tone systems, generally 5 or 6 tones based on a 3-by-2 tone pattern, with mergers resulting in the loss of a tone. They have only monosyllabic morphemes, and syllables have at most $C_1(C_2)V(C_3)$ structure in which the C_2 can only be a liquid or glide (and in Vietnamese only /w/). They are also both recipients of much larger quantities of Chinese loanwords than any other Vietic lects.

Cuoi-Tho: The results of our study confirm the grouping of the Cuoi-Tho lects, which are geographically farther north than other non-VM lects, and closer to VM. The results of the phylogenetic calculations divide the four lects in an unexpected way: one each with Ferlus' and Nguyen's fieldwork lects. In other words, those authors' two lects do not appear to group more closely to each other, although these differences are marginal for our purposes. Just as Cuoi-Tho are lexically closest to VM languages, they are also typologically closest: they are monosyllabic and have tone systems that parallel those of VM structurally without being neatly cognate. Of particular interest is that, in our data, Cuoi-Tho and Pong-Toum do not form a single node, in contrast with some previous analyses.

Pong-Toum: The available lexical data clearly shows that Pong, Toum, and Liha form a group, but also, crucially, are distinct from the Cuoi-Tho group. In contrast to the clear results of the lexical data, the lack of consistent, complete phonological descriptions presents challenges. As for syllable shapes, Nguyen T. L. (1992: 101) posits that approximately ten percent of Pong words are sequisyllabic (i.e. bisyllabic words with iambic stress, resulting in the presyllables that are unstressed and a limited number of consonants and neutralized vowels), and Ferlus (2014: 1) mirrors this point: it is a point of distinction between PT and CT. However, while Pong has been described as having a 4-way system divided by oppositions in registral height and phonation (Nguyen T. L. 1992: 99), in the current lexical data, Liha (in Peiros 2004 from Nguyen V. L.) has six tonal contours. If this is the case, developments in the suprasegmental systems must have occurred after this subgroup separated from Vietic. These lects do appear to be **t**ypologically transitional between the monosyllabic, tonal VM and CT groups and the remaining archaic varieties, the latter having higher rates of sesquisyllabic roots and simpler tone systems or register-phonation systems, as described below.

Chut-Arem: To the southernmost extent of the Eastern Vietic clade, Sach, Ruc and May are phonologically archaic lects that clearly form a compact sub-group lexically, generally called 'Chut'.[16] Arem is similarly archaic, but lexically divergent, such that it does not obviously fall into Chut based on the computational results. However, we do identify some five Chut-Arem isoglosses in our 116-item list, and these are listed in the first five rows of table 4, along with representative glosses for the other Vietic sub-groups. We propose that these are best regarded as common inheritances from proto-Chut rather than borrowings of Chut lexical innovations into Arem, as the latter seems unlikely to us.

Additionally, an anonymous reviewer suggested that possible Katuic influence on Arem and Thavung-Malieng should be investigated in this regard. In our 116-item list, we find only one possible shared Katuic loan into Arem and T-M, 'horn', although on phonological grounds it is striking that the Arem form more closely resembles Bahnaric reflexes, while only the T-M

[16] Nguyễn Phú Phong et al. (1998) note this as a Chamic loanword. Thurgood (1999: 315) reconstructs Proto-Chamic *cət 'mountain range', but suggests that the vowel is indicative of an Austroasiatic loanword into Chamic. It does not appear to be a wider Austronesian word or even Malayic. However, as we cannot locate any comparable forms in Austroasiatic, including neighboring Bahnaric and Katuic languages, it does seem to be a borrowed exonym.

forms match closely the Katuic etymon. One other Arem-T-M isogloss is evident in our 116-item list, namely 'earth/soil'. The form resembles Katu *katiek*, but if those forms are related, the vowel discrepancy suggests independent borrowing of formation of the word in Arem and T-M versus Katuic. The origin of this 'earth/soil' etymon is obscure, although it may be a phonological deformation of the Vietic etymon, whereby the coda of *tət became velar, and a presyllable was acquired. While an exhaustive etymological investigation might potentially reveal a very differently proportioned distribution of isoglosses, restricting ourselves to the present dataset, we tend to the view that Arem is an aberrant Chut lect.

Gloss	Arem	Sach	Viet.	Cuoi-TanHop	Liha	So-Thavung	Notes
'beast/udder'	nɒːi̯	nɔːj^{3}	*vú*	ʔu^{12}	now	təmʔák	Lao ʔák 'chest'
'nail/claw'	n̥tɒŋ	katoːŋ^{1}	*móng*	sɑ̆m^{22}	sam	kasâm	
'fat/grease/oil'	tlùŋ	tluŋ^{2}	*mỡ*	mɤ^{33ʔ}	mə	ʔatûː	
'navel'	uɗuh	duduːl^{3}	*rốn*	sun^{12}	suːɲ	cuʔbîː	
'night'	lɯm	lúm	*đêm*	tem^{32}	tɛːm	ʔamáh	
'earth/soil'	atăk	bɤ́nʔ	*đất*	tɤ̆t^{12}	tɔt	ʔaták	Katu katiek
'horn'	takaːi̯~təkeː	ʈəŋ^{2}	*sừng*	khʂiŋ^{54}	k^{h}lɔŋ	takɔ̂ːj	Katu takɔːj, Bahnar ʔəkɛː

Table 4. Arem isoglosses

Thavung-Malieng: These lects, some of which are generally tonal and highly registral (making lexical use of creaky and breathy phonation), appear to unambiguously stem together from the highest branching node in the tree. Phonologically it appears that the TM languages are conservative in regard to the outcomes of proto-Vietic coda *-r, with the ancestor of the Eastern clade undergoing a general merger of *-r and *-l to *-l, while reflexes of *-r diversified later within TM. Additionally, there appear to be Tai loans (e.g. 'good', 'to know', 'small' from Tai) within TM that are not shared with the rest of Vietic (and thus likely within more recent centuries), which speaks to the unique history of the subgroup after the initial split of Vietic into Western and Eastern clades.

Given the considerations discussed above, the tree structure we propose has a compelling logic to it. Taking the West-East split as primary, the TM languages have their own history of lexical, segmental and tonal developments; on the Eastern side, the highest branching nodes reflect archaic lects that retain sesquisyllables and show only limited tonogenesis. Below them in the Eastern branch are the more innovative subgroups, with VM lects nested at the extreme north with the most innovative lexical and phonological histories.

4.5 Previous studies with coinciding results

Vietic phylogenies proposed by other scholars also reproduce specific aspects of our findings, although the bases for these are not always clear. We have also presented a cursory overview of a dozen previous classifications of Vietic languages in Table 2. Below, a few of these are presented with complete trees and discussion of the overlap with our current model.

James Chamberlain has authored several studies analyzing Vietic vocabulary, primarily faunal lexicon, and this has informed several classifications. Phylograms are offered in both 1998 and 2018 papers (Figures 6 and 7), and both anticipate our primary result of a binary division between TM and a clade covering the rest of the Vietic languages, notably indicating a closer connection between Chut and VM. In Chamberlain's (1998) model, neither primary

clades are named, while the lower sub-groups have mostly neutral geographical names. The structure of each primary clade is flat, one with three groups: Viet-Muong, Pong-Toum, and Chut and the other with four TM lect groups.

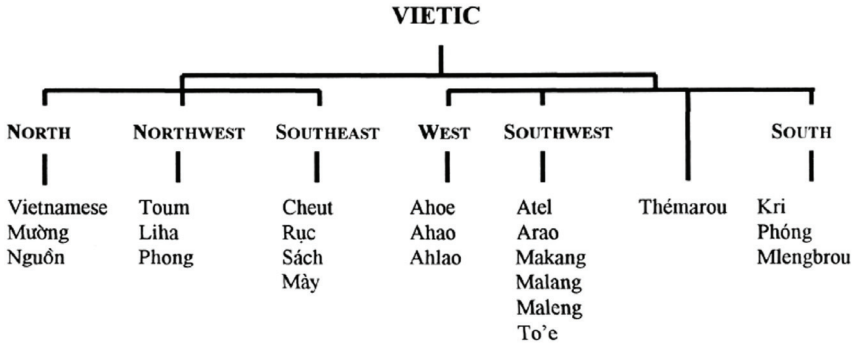

Figure 6. Chamberlain (1998: 106) "Suggested modifications to Vietic subgrouping".

Chamberlain (2018) offers a somewhat refined classification (Figure 7) that indicates some nested branching. Additionally, he invents unique names for language sub-groupings and nodes in the tree which are unrelated to equivalent commonly used terms. Specifically, Pong-Toum and Chut are grouped in opposition to Viet-Muong, and multiple levels of nested branching are indicated within the TM clade, which is called "Nrong-Theum". While showing closer affiliation of Chut with VM, as opposed to the other bisyllabic Vietic lects, as our model does highlight the closer connection of Pong-Toum and Cuoi-Tho lectal groups with VM, all of which are monosyllabic (with a small number of sesquisyllables noted in Pong-Toum) and highly tonal.

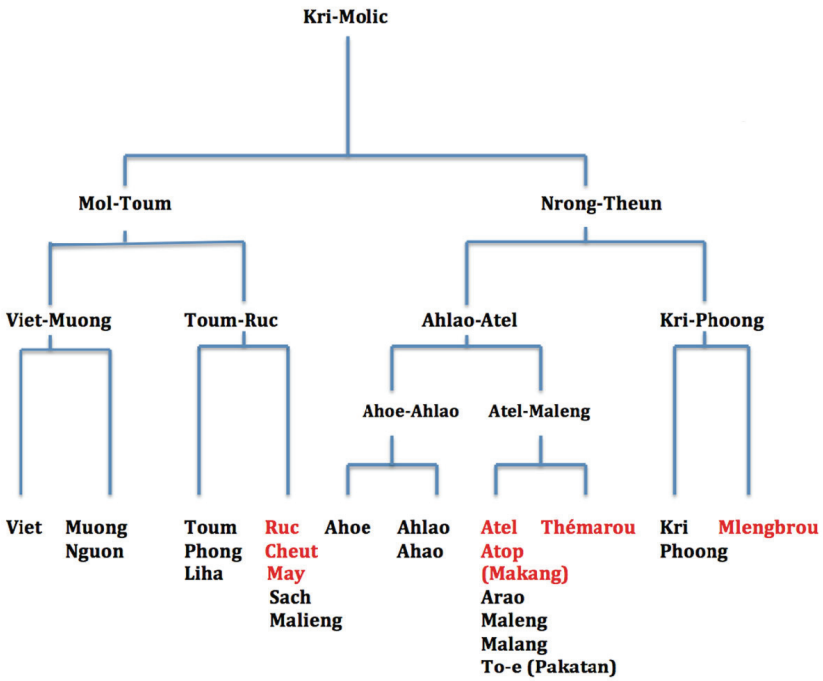

Figure 7. Chamberlain (2018: 12) Kri-Molic (Vietic) phylogram based on faunal lexicon.

The other proposal of note is the lexicostatistical analysis by Peiros (2004), which is based on the Swadesh 100 items, and processed with the StarLing software package (see Figure 7). Peiros' results closely parallel ours in respect of the lower branchings, notably offering similar

configurations of Viet-Muong and the Pong-Toum and Cuoi-Tho clades (although Liha is displaced by one node). At the higher levels, Peiros does place Thavung on a primary branch opposed to the rest of Vietic, as we do, but he does not group Thavung with Malieng, and this is difficult to account for given the rather strong sub-grouping of these evident in our dataset, which is very similar to that of Peiros. We assume that the main reasons for the differences between Peiros' results and ours relate to the shorter wordlists (100 versus 116 items) and his particular assessments of cognates and loans. Overall, we take this comparable precedent as reinforcing the overall results we have obtained from both computational data as well as the comparative method in terms of segmental and prosodic issues.

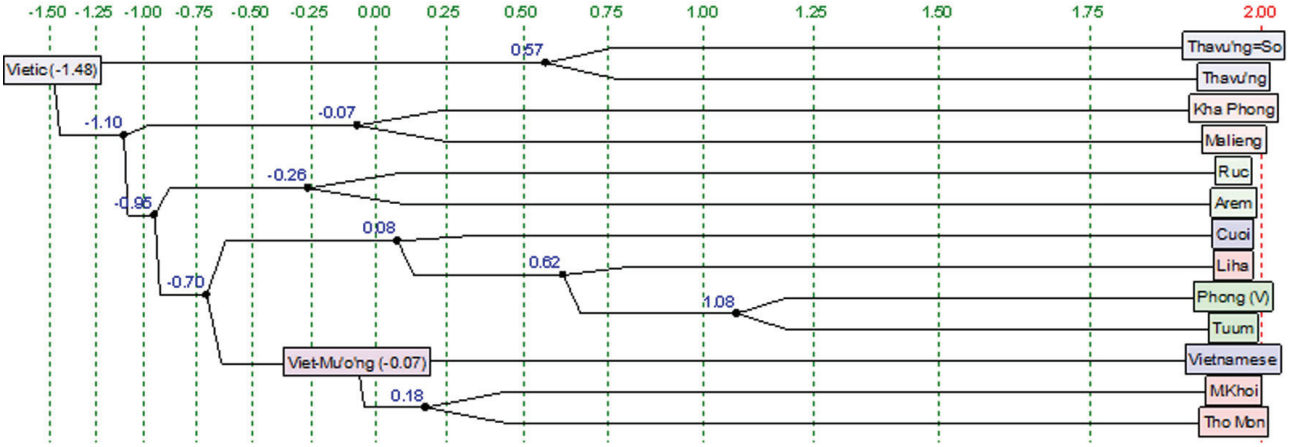

Figure 8. Peiros (2004) Vietic phylogram based lexicostatistics (100 word list).

5. Historical Geographic Considerations

The modern position of Vietic groups is telling of the geographic history of the greater Vietic speech community. Until the late 1440s, the southernmost extent of Vietic speakers was in north-central Vietnam, and the great expansion southward only began in the late 1400s in the period known as the Nam Tiến, or the "Southward Advance," after the fall of the Cham polities following centuries of Cham-Vietnamese military conflicts.[17] This expansion led to the movement primarily of those speaking Vietnamese, which was already a fully distinct language group from the rest of Vietic. The southernmost reaches of Vietnam (e.g. the Mekong Delta) were populated by Vietnamese speakers only in the past few hundred years. Thus, the modern map of Vietic languages from north-central to northern Vietnam (see Figure 9) may be a reasonable representation of the geographic distribution of Vietic groups in the pre-expansion period.

Archaeological, genetic, historical, and linguistic data altogether show that the heartland of Vietnamese is the Red River delta, going back to the Han period of Chinese colonization (cf. Alves 2021), and consistent with this, the Muong lects are spoken inland not far to the west of the delta. The Nguon speakers are located much further south, but this may reflect the relocation of a garrison recruited in the Viet-Muong area under Chinese command (Ferlus pers

[17] It is striking that this follows the fall of Angkor, traditionally dated to 1431 after being sacked by Ayutthaya, and these days considered to have been aggravated by climate change that made the Mekong Delta a more attractive locus for agriculture (Penny et al. 2018). Regardless of the causes, this period is one of considerable political change and migration in the region, and thus one in which social conditions contributed to a dynamic linguistic environment.

com.). Even if this specific historical scenario is not the case, the close affiliation of Nguon with the rest of VM suggests that this outlying geographic location is the result of migration of a VM language, and so this does not affect the identification of the Viet-Muong historical locus in the north.

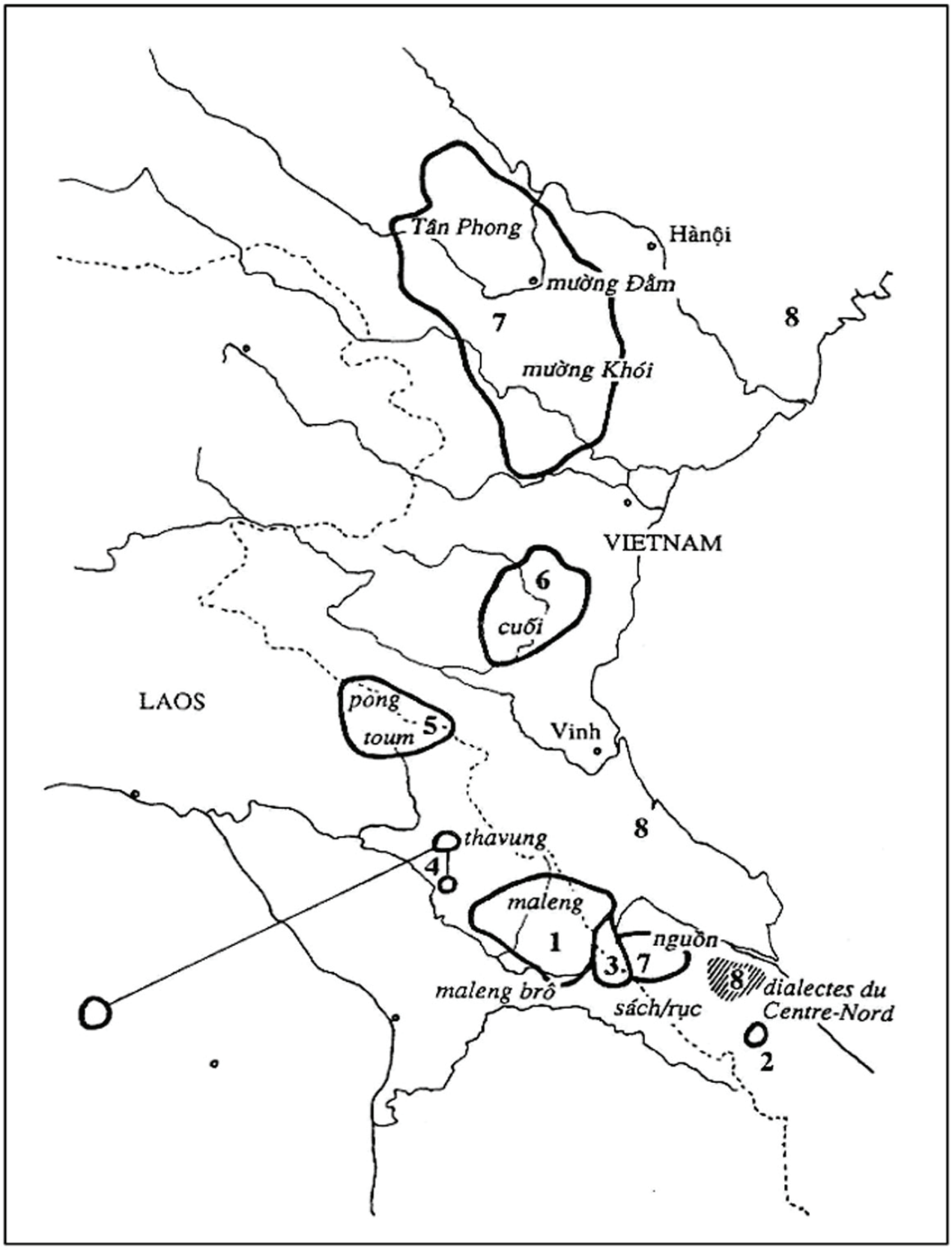

Figure 9. Map of Vietic languages approximate locations (Ferlus 1998: 27)

Other Vietic subgroups are all located south of the delta, mainly in the hills of the Annamite Range, plus some migrations further afield, such as the Thavung in Thailand documented by Premsrirat (2000). The geographical distribution broadly follows our proposed phylogeny; the Pong-Toum and Cuoi-Tho subgroups are located closer to Viet-Muong, being roughly in the centre of the Vietic range. The more archaic subgroups Chut, Arem, Thavung-Malieng are further south, with Thavung-Malieng to the west of Chut and Arem, mostly on the Lao side of the border. Assuming a least-moves analysis, this places the locus of genetic diversity in Vietic approximately in the north-west of Quang-Binh Province, implying that Viet-Muong moved northwards into its present range. However, ancient loanword evidence suggests otherwise.

It is apparent that there was a much older phase of Tai-Vietic contact before proto-Vietic began to disperse, potentially dating to the BCE-era Dongson period, evidenced by a number of proto-level items, e.g. 'duck', 'winnow basket', 'drum', 'water pipe of bamboo', 'water spinach', 'bush-knife', etc. (Alves 2021: 25). This is also consistent with genetic material in human remains from the Dongson-era Nui Nap archaeological site in the Red River Delta, showing shared alleles of both Tai and Kinh (i.e. ethnic Vietnamese) populations (Lipson et al. 2018: 2).[18] This was also the period prior to the dispersal of Southwestern Tai, so we expect that Tai was more geographically constrained to the northern parts of Vietnam and bordering areas of southeastern China.

If our interpretation of these old loans is correct, a Vietic homeland in the Red River Delta or proximal region of Northern Vietnam is supported, the issue being to explain the distribution to the south of the other Vietic subgroups. Indeed, we find it provocative that the Man Bac site (in the Red River Delta) of the Phung Nguyen culture (c. 4000–3500 BP) is among the earliest sites with evidence of incoming agriculturalists from the north (e.g. Matsumura et al. 2008), and that there is a demonstrated sequence of archaeological cultures from the Phung Nguyen culture to the Dong Son culture (e.g. Kim 2015: 106). Thus, in the larger picture, the spread of Vietic is tied to and may follow from the original Austroasiatic dispersal. Moreover, it is surely relevant that the range of Vietic languages approximates the reach of Chinese administrative control in the Han period and some Vietic groups may have moved south and upland to avoid Chinese control, or may have been moved southward by the Chinese to help provide a buffer to the Chams. Another possibility is that, when Sinitic speakers arrived, Vietic speakers were already spread from the Red River Delta south into Thanh Hoa and areas of north-central Vietnam and bordering parts of modern-day Laos. In this situation, the extent of the more typologically innovative Viet-Muong shows the approximate range of a substantial Sinitic presence and language contact from the early first millennium CE, but whether there was Sinitic language contact with other Vietic languages in that early period is as yet unclear. However, to determine which scenario is most viable, we need clearer indications of the disposition of Vietic groups before Han period, and in this respect, perhaps all we have to go on are the archaeological continuities in the Red River delta before and after Chinese domination, hinting that the Viet-Muong have a deeper history of settlement in the north.

There remains the question of to what extent language contact is relevant to the higher branching of Vietic, which may be a potentially rich area for subsequent study. There are clear Tai loanwords among TM languages, and relevant items from our 116-word list are aggre-

[18] Lipson et al. (2018) also note the shared alleles encompass Austronesian speakers. However, this is harder to resolve as (a) Austronesian is a huge language family with many branches, making it difficult to interpret in a meaningful way, (b) there are no archaeological indications of an Austronesian community in northern Vietnam in the Dongson period, and (c) the connection between Kradai and Austronesian is increasingly accepted, making a claim of both Austronesian and Tai from the mid-1st millennium possibly redundant.

gated in table 5. Comparisons are made with Lao, Thai and Shan to establish the Tai origins, although local Phu Thai[19] may be an important source of influence. Additional Tai loanwords can be found among these lects (cf. Hayes 1982 on Tai loanwords in Thavung), but in terms of our 116-item list, they are not relevant. The Thavung spoken in Northeast Thailand has in particular a large number of probable recent Lao (Isaan Thai) loanwords, including grammatical ones, which is apparent from examining the dictionary of Premsrirat (2000) and the Thavung So grammar of Srisakorn (2008).

Gloss	Lect	Item	Compare
breast/udder	Thavung (S)	təmʔə́k~təpʔə́k	Lao ʔə́k 'chest, breast'
chest	Thavung (F)	ʔək[1]	
cloud	Thavung (F), Malang	me:k	Thai ˈmêːk 'cloud'
good	Thavung (S), Thavung (PS), Thavung (F), Pakatan, Malang	di:	Thai, Lao di: 'good'
hard/solid	Maleng Bro	lɛ̰:ŋˈ	Thai rɛ:ŋ 'forcefully, hard'
to know (facts)	Thavung (S); Thavung (F); Kri	húʔ; huʔ; ɽuˈ / ʐuˈ	Lao hû:, Thai rú: 'to know'
mountain	Thavung (S)	phû:	Lao pʰú: 'hill, mountain'
person	Malang	khon	Lao kʰón 'person'
small	Thavung (S)	ʔi:t	Shan ʔit[4] 'small'
to speak/say	Thavung (F) Kri vs̆u̯ˈ	vaw[4]	Lao wâu 'to speak/say'
walk	Thavung (S)	ʔaɲa:ŋ	Lao ɲā:ŋ 'to walk, stride'

Table 5. Tai loans in Thavung-Malieng basic vocabulary

Another indicator of language contact is seen in noun phrase structure: while Chut languages show noun phrase structure that parallels that of Vietnamese, So Thavung mirrors the Lao/Thai type, while Kri has flexible word order, allowing for either type, all of which suggests substantial language contact with Vietnamese and Laos respectively (Alves 2020). In contrast to the TM lects, the Chut lects in our small-scale study show no Tai loanwords in their basic vocabulary. Not surprisingly, in our larger data sets, likely Vietnamese loanwords have been discovered, including some grammatical lexemes, likely to have been borrowed in recent decades.

Broadly, we can say that Tai influence is strongly evident in the TM lects, while Vietnamese influence (especially recent influence) is apparent in Chut lects. This speaks to the geographical history, indicating that TM has been isolated from the rest of Vietic on the western slopes of the Annamite Range for a significant period of time, but the relevant chronology is unclear without a more detailed study.

6. Concluding observations

In this short paper, we report on our phylogenetic analysis of the Vietic languages, summarising our results in the family tree given in Figure 5. The computational phylogenetics based on a 116-item basic word list, and the historical phonology of Vietic codas *-h, *-s, *-r, *-l, converge on the identification of five subgroups, Thavung-Malieng, Chut-Arem, Pong-Toum, Cuoi-Tho, and Viet-Muong. The tree has a simple descending binary branching structure, with

[19] Phu Thai is a broad term used to refer to diverse upland Tai, and available lexicons do not readily permit accurate localization of specific lexical forms.

Thavung-Malieng and Eastern Vietic being the first split, and Viet-Muong the last. The configuration of language relations within the subgroups is largely untouched by our study as the limited data set permits only rather low-resolution results, and we would welcome more richly detailed analyses. While the geography of the languages is suggestive of a locus of dispersal in the southern end of the Vietic range, indications from old loanwords, archaeological studies, and consideration of the history of Chinese occupation equally suggest a northern locus of dispersal, although further research is needed to clarify this important question. The data on which our conclusions have been reached is available for download, we offer this as an example of transparency and will gladly cooperate with any researchers wishing to reproduce or augment this study.

Appendix 1: Excluded words and rationale for exclusion

Item	Remarks
'all'	The 'all' gloss is frequently associated with the meaning 'all gone/ nothing left' rather than the concept of 'every item in a set'.
'cold'	The languages of the area often distinguish between internal sensation of 'cold' versus objects being 'cold', and the sense of 'cool' in contrast with 'cold' adds further challenges.
'to come'	The meaning is not clearly distinguished between 'arrive' and being on the way to a destination.
'man/husband'	Languages vary as to distinguishing 'man', 'husband', or 'male'.
'many/much'	The meaning was not well represented in the sources, and items present many are suspected of being Vietnamese loans.
'name'	The meaning was not well represented in the sources, and where items were present many were suspected of being Vietnamese and Tai loans.
'not'	Multiple negators were typically listed and no clear basis for identifying comparable items was apparent.
'seed'	Lists did not clearly distinguish 'seed for planting' versus 'seed found within edible fruit'.
'sun'	'Sun' is commonly encoded by a compound 'eye (of) day'.
'that'	Languages vary in their demonstrative systems, with cases of multiple distal forms rendering comparison difficult.
'we (incl.)'	There were too many gaps and inconsistencies in the lists of relevant pronominal forms.
'woman/wife'	Languages vary as to distinguishing 'woman', 'wife' and 'female'.
'yellow'	Most lists simply had reflexes of Chinese 黃 *huáng*.

References

Alves, Mark. 2003. Ruc and Other Minor Vietic Languages: Linguistic Strands Between Vietnamese and the Rest of the Mon-Khmer Language Family. In: Karen L. Adams et al. (eds.). *Papers from the Seventh Annual Meeting of the Southeast Asian Linguistics Society*: 3–19. Arizona State University, Program for Southeast Asian Studies.

Alves, Mark J. 2006. Linguistic research on the origins of the Vietnamese language: An overview. *Journal of Vietnamese Studies* 1.1–2: 104–130.

Alves, Mark J. 2017. Etymological research on Vietnamese with databases and other resources. In: *Ngôn Ngữ Học Việt Nam, 30 Năm Đổi Mới và Phát Triển (Kỷ Yếu Hội Thảo Khoa Học Quốc Tế)*: 183–211. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

Alves, Mark. 2020. Initial steps in reconstructing Proto-Vietic syntax. In: Mathias Jenny, Paul Sidwell, Mark Alves (eds.). *Austroasiatic Syntax in Areal and Diachronic Perspective*: 46–81. Boston: Brill.

Alves, Mark J. 2021. The Đông Sơn Speech Community: Evidence for Vietic. *Crossroads 19 (2021)*: 1–41.

Babaev, Kirill V., Irina V. Samarina. 2018. *Materials of the Russian-Vietnamese linguistic expedition: May language* (in Russian). Moscow: Yask Publishing House.

Barker, Milton E. 1963. Proto-Vietnamủởng Initial Labial Consonants. *Văn-hoa Nguyêt-san* 12.3: 491–500.

Barker, Milton E., Muriel A. Barker. 1970. Proto-Vietnamuong (Annamuong) final consonants and vowels. *Lingua* 24.3: 268–285.

Blagden, Otto. 1913. The classification of the Annamese language. *Journal of the Royal Asiatic Society*: 427–432.

Cadière, Léopold. 1905. Les hautes vallées du sông Gianh. *Bulletin de l'École Française d'Extrême Orient* 5: 349–367.

Chamberlain, James R. 1998. The origin of Sek: implications for Tai and Vietnamese history. In: S. Burusphat (ed.). *The International Conference on Tai Studies*: 97–128. Institute of Language and Culture for Rural Development, Mahidol University, Bangkok, Thailand.

Chamberlain, James R. 2003. Eco-Spatial History: a nomad myth from the Annamites and its relevance for biodiversity conservation. In: X. Jianchu, S. Mikesell (eds.). *Landscapes of Diversity: Proceedings of the III MMSEA Conference, 25–28 August 2002*: 421–436. Lijiang: Center for Biodiversity and Indigenous Knowledge.

Chamberlain, James R. 2018. *A Kri-Mol (Vietic) Bestiary: Prolegomena to the Study of Ethnozoology in the Northern Annamites. Kyoto Working Papers on Area Studies No. 133*. Kyoto University.

Chazée, Laurent. 1999. *The Peoples of Laos: Rural and Ethnic Diversities.* Bangkok: White Lotus.

Chéon, A. 1907. Note sur les Dialectes Nguon, Sac Et Muong. *Bulletin de l'Ecole Francaise d'Extreme-Orient* 7.1–2: 87–99.

Enfield, Nicholas J., Gérard Diffloth. 2009. Phonology and sketch grammar of Kri, a Vietic language of Laos. *Cahiers de linguistique Asie orientale* 38.1: v–69.

Ferlus, Michel. 1974. Le groupe viet-muong (Recherches dans le cadre de l'Atlas Ethnolinguistique). *Asie du Sud-Est et Monde Insulindien* 5.1: 69–77.

Ferlus, Michel. 1975. Vietnamien Et Proto-Viet-Muong. *Asie du Sud-Est et Monde Insulindien* 6.4: 21–55.

Ferlus, Michael. 1979. Lexique Thavủng-Français. *Cahiers de Linguistique, Asie Orientale* 5: 71–94.

Ferlus, Michel. 1991. *Vocalisme du Proto-Viet-Muong*. Paper circulated at the Twenty-fourth ICS-TL&L. Chiang Mai University, Oct. 10–11, 1991.

Ferlus, Michel. 1996. Langues et peoples Viet-Muong. *Mon-Khmer Studies* 26: 7–28.

Ferlus, Michel. 1997. Problèmes de la formation du systèm vocalique du vietnamien. *Cahiers de Linguistique, Asie Orientale* 26.1: 37–51.

Ferlus, Michel. 1998. Les systèmes de tons dans les langues viet-muong. *Diachronica* 15.1: 1–27.

Ferlus, Michel. 2007. *Lexique de racines Proto Viet-Muong (Proto Vietic Lexicon)*. Unpublished Ms. Available online at: http://sealang. net/monkhmer/database.

Ferlus, Michel. 2014. *Proto Viet-Muong (Proto-Vietic)*. Unpublished Ms. Available online at: https://halshs.archives-ouvertes.fr/halshs-02490370

Ferlus, Michael. 2017. *Phong Toum comparatif*. Unpublished Ms.

Gage, William W. 1985. Vietnamese in Mon-Khmer Perspective. In: Suriya Ratanakul et al. (eds.). *Southeast Asian Linguistic Studies Presented to André-G. Haudricourt*: 493–524. Institute of Language and Culture for Rural Development, Mahidol University.

Greenhill Simon J., Paul Heggarty, Russell D. Gray. 2020. Bayesian Phylolinguistics. In: R. D. Janda, B. D. Joseph, B. S. Vance (eds.). *The Handbook of Historical Linguistics, Volume II*: 226–253. Wiley-Blackwell: New Jersey.

Guignard, Théodore. 1911. Note sur une peuplade des montagnes du Quáng-Bīnh: les Tầc-cúi. *Bulletin de l'Ecole française d'Extrême-Orient* 11: 201–205.

Haudricourt, André-Georges. 1953. La Place Du Viêtnamien Dans Les Langues Austroasiatiques. *Bulletin de la Société de Linguistique de Paris* 49.1: 122–8.

Haudricourt, André-Georges. 1954a. De L'Origine Des Tones En Viêtnamien. *Journal Asiatique* 242: 69–82.

Haudricourt, André-Georges. 1954b. Comment Reconstruire Le Chinois Archaique. *Word: Journal of the International Linguistic Association* 10: 351–364.

Hayes, La Vaughn H. 1982. The Mutation of *R in Pre-Thavung. *Mon-Khmer Studies* 11: 83–100.

Hayes, La Vaughn H. 1992. Vietic and Việt-Mường: a new subgrouping in Mon-Khmer. *Mon-Khmer Studies* 21: 211–228.

Huffman, Franklin E. 1978. *On the centrality of Katuic-Bahnaric to Austroasiatic*. Paper presented at the 2nd International Conference on Austroasiatic, Mysore (India), December 18–21, 1978. Available online at: https://sites.google.com/view/paulsidwell/the-sical-papers.

Jenny, Mathias, Paul Sidwell (eds.). 2014. *The handbook of Austroasiatic languages (2 vols.)*. Leiden / Boston: Brill.

Kasuga, Atsushi. 2008. *Arem Vocabulary*. Ms., organized in order Vietnamese-Arem-English.

Kim, Nam. 2015. *The Origins of Ancient Vietnam*. (*Oxford Studies in the Archaeology of Ancient States*.) Oxford University Press.

Lipson, Mark et al. 2018. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361.6397 (2018): 92–95.

Logan, James Richardson. 1856. Ethnology of the Indo-Pacific Islands. Part II, Chap. VI, App. A: Comparative Vocabulary of the Numerals of the Mon-Anam Formation. App. B: Comparative Vocabulary of Miscellaneous Words of the Mon-Anam Formation. *Journal of the Indian Archipelago* 1.

Mạc, Đường. 1964. *Các dân tộc miền núi Bắc Trung Bộ*. Hà Nội: Khoa Học Xã Hội.

Malglaive, Joseph de. 1902. Vocabulaire Hang-Tchek, Khas Xos, Harème (by M. Riviére). In: J. de Malglaive. *Voyages au centre de l'Annam et du Laos et dans les regions sauvages de l'est de l'Indo-Chine*: 285–290. Paris: E. Leroux.

Maspéro, Henri. 1912. Etude sur la phonétique historique de la langue annamite: les initiales. *Bulletin de l'Ecole Française d'Extrême Orient* 12: 1–127.

Matsumura Hirofumi et al. 2008. Morphometric affinity of the late Neolithic human remains from Man Bac, Ninh Binh Province, Vietnam: Key skeletons with which to debate the 'two layer' hypothesis". *Anthropological Science* 116.2 (2008): 135–148.

Nguyễn, Hữu Hoành. 1999. Về sự phân định các ngôn ngữ của nhóm Việt-Mường [On the distribution of the Viet-Muong languages]. *Ngổn Ngữ* 5 (1999): 35–42.

Nguyễn, Hữu Hoành. n.d. *Wordlists of about 4,000 words in two Cuoi lects (Thái Hòa & Tân Hợp)*. Unpublished Ms.

Nguyễn, Hữu Hoành, Văn Lợi Nguyễn. 2019. Tones in the Cuoi Language of Tan Ki District in Nghe An Province, Vietnam. *Journal of the Southeast Asian Linguistics Society* 12.1: lvii- lxvi.

Nguyễn, Phú Phong. 1997. *Le Parler Nguồn: Langue d'une Minorité ethnique des hautes vallées du Sông Gianh Quảng Bình, Việt Nam. Cahiers d'Etudes Vietnamiennes*. Paris: Université Paris 7 – Denis Diderot.

Nguyễn, Tài Cẩn. 1995. *Giáo trình lịch sử ngữ âm tiếng Việt (sơ thảo) [Textbook of Vietnamese historical phonology]*. Hà Nội: Nhà Xuất Bản Gíao Dục.

Nguyen, Tuong Lai. 1992. Poong language – The first contact of languages between Viet and Thai. In: *Pan-Asiatic linguistics: proceedings of the Third International Symposium on Language and Linguistics, Bangkok, Thailand*: 98–107. Bangkok: Chulalongkorn University.

Nguyễn, Văn Lợi. 1993. *Tiếng Rục [The Rục language]*. Hà Nội: Nhà Xuất Bản Khoa Học Xã Hội.

Nguyễn, Văn Lợi, Hữu Hoành Nguyễn. 2001. *Field data of wordlists for Malang Pakatan*. Unpublished Ms.

Nguyễn, Văn Lợi, Hữu Hoành Nguyễn. 2019. *Field data of wordlists for Kri Phoong*. Unpublished Ms.

Nguyễn, Văn Khang, Bùi Chi, 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 [The phonetics of the Mường language across its various dialects]*. Hanoi: Nhà Xuất bản Từ điển Bách khoa.

Peiros, Ilia. 2004. *Genetičeskaja klassifikacija avstroaziatskix jazykov. Doctoral dissertation*. Moskva: RSUH.

Penny, Dan, Cameron Zachreson, Roland Fletcher, David Lau, Joseph T. Lizier, Nicholas Fischer, Damian Evans, Christophe Pottier, Mikhail Prokopenko. 2018. The demise of Angkor: Systemic vulnerability of urban infrastructure to climatic variation. *Science Advances* 4(10).

Phạm, Đức Dương. 1975. Về mối quan hệ thân thuộc giữa các ngôn ngữ thuộc nhóm Việt-Mường miền tây tỉnh Quảng Bình [On the close relationship between languages in the Viet-Muong group in Western Quang Binh province]. In: Bê Viêt Dăng (ed.). *Về Vấn Đề Xác Định Thành Phần Các Dân Tộc Thiểu Số ở Miền Bắc Việt Nam [On the problem of defining the social position of the minority groups in northern Vietnam]*: 500–517. Hanoi: Nhà xuất bản Khoa học Xã hội.

Phạm, Đức Dương. 1979. Về mối quan hệ nguồn gốc của các ngôn ngữ nhóm Việt Mường [On the original relationship among the Viet-Muong languages]. *Ngôn Ngữ* 1979.1: 46–58.

Phan, John D. 2012. Mường is not a subgroup: Phonological evidence for a paraphyletic taxon in the Viet-Muong sub-family. *Mon-Khmer Studies* 40: 1–18.

Pittayaporn, Pittayawat. 2009. *The Phonology of Proto-Tai*. PhD thesis. Cornell University.

Premsrirat, Suwilai. 2000. *So (Thavung) Preliminary Dictionary*. Salaya / Melbourne: Institute of Language and Culture for Rural Development, Mahidol University – University of Melbourne.

Przyluski, Jean. 1924. Les langues Austroasiatiques. In: Antoine Meillet, Marcel Cohen (eds.). *Les Langues du Monde (Collection linguistique publiée par la société de linguistique de Paris)*, 16: 335–403. Paris: Librarie Ancienne Edouard Champion.

Samarina, Irina. n.d. *Spreadsheet of comparative Vietic data*. Unpublished Ms. (early version of the lexicon reproduced on pp. 259–263 in Babaev & Samarina 2018).

Schmidt, Wilhelm. 1903. The Sakai and Semang languages in the Malay Peninsula and their relation to the Mon-Khmer languages. *Journal of the Straits Branch of the Royal Asiatic Society* 39: 38–45.

Schmidt, Wilhelm. 1904. Grundzüge einer Lautlehre der Khasi-Sprache in ihren Beziehungen zu derjenigen der Mon-Khmer-Sprachen. Mit einem Anhang: die Palaung-Wa-, und Riang-Sprachen des mittleren Salwin. *Abh. Bayrischen Akademie der Wissenschaft* 1.22.3: 677–810.

Schmidt, Wilhelm. 1905. Grundzüge einer Lautlehre der Mon-Khmer-Sprachen. *Denkschrift der Akademie der Wissenschaften, Wien, Philologisch-Historische Klasse* 51: 1–233.

Schmidt, Wilhelm. 1906. Die Mon-Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens. *Archiv für Anthropologie, Braunschweig*, 5: 59–109.

Shorto, Harry L. 2006. *A Mon-Khmer Comparative Dictionary.* Canberra: Pacific Linguistics.

Sidwell, Paul. 2009. *Classifying the Austroasiatic languages: history and state of the art.* Munich: Lincom Europa.

Sidwell, Paul. 2014. Austroasiatic classification. In: Matthias Jenny, Paul Sidwell (eds.). *The Handbook of Austroasiatic Languages. Vol. 1*: 144–220. Leiden / Boston: Brill.

Sokolovskaja, N. K. 1978. Materialy k sravnitel'no-ètimologicheskomu slovar'u vjetmyongskix jazykov. In: V. M. Alpatov (ed.). *Issledovanija po fonologii i grammatike vostochnyx jazykov*: 126–180. Moskva: Nauka.

Srisakorn, Preedaporn. 2008. *So (Thavung) grammar. Doctoral dissertation*. Thailand: Mahidol University.

Thomas, David D., Robert K. Headley. 1970. More on Mon-Khmer subgroupings. *Lingua* 25: 398–418.

Thompson, Laurence C. 1976. Proto-Viet-Muong Phonology. In: Philip N. Jenner, Laurence C. Thompson, Stanley Starosta (eds.). *Austroasiatic Studies, Vol. 2*: 1113–1204. Honolulu: University of Hawaii Press.

Trần, Trí Dõi. 2011. *Một vài vấn đề nghiên cứu so sánh - lịch sử nhóm ngôn ngữ Việt - Mường [A historical-comparative study of Viet-Muong group]*. Hà Nội: Nhà xuất bản Đại Học Quốc Gia Hà nội.

Trần, Trí Dõi. 2018. Một tham chiếu về nguồn gốc của tiếng Việt [A reference regarding the origin of Vietnamese]. In: Đinh Văn Đức (ed.). *Tiếng Việt Lịch Sử: Một Tham Chiếu Hồi Quan [Vietnamese Language History: A Reference]*: 10–86. Hanoi: Nhà Xuất Bản Văn Học.

Vương, Hoàng Tuyên. 1963. *Các dân tộc nguồn gốc Nam Á ở miền Bắc Việt Nam [Ethnic groups of Austroasiatic origin in northern Vietnam]*. Hanoi: Nhà Xuất Bản Giáo Dục.

Пол Сидвелл, Марк Алвес. Вьетские языки: филогенетический анализ

В работе представлена новая внутренняя классификация вьетских языков, включающая все общепризнанные подгруппы и языки, для которых доступны языковые данные, необходимые для сравнения. В ходе анализа согласованы результаты двух различных методик: (1) вычислительная филогенетика, основывающаяся на 116-словном сравнительном списке и (2) сравнительный анализ диахронических изменений в финалях слоговых морфем. Анализ позволяет выделить пять основных подгрупп: тхавынг-малиенг (ТМ), тыт-арем, поонг-тум, куой-тхо и вьет-мыонгскую (ВМ). Хотя идентификация этих подгрупп сама по себе не оригинальна, ряд особенностей ветвления дерева выявлен впервые. Так, установлено, что вьетское дерево имеет бинарную структуру, разделяясь на ТМ и все остальные языки, которые можно в совокупности назвать *восточновьетской* кладой. Внутри этой восточной клады языки подразделяются на северные и южные, из которых северная группа (ВМ) оказывается более инновативной, а южная – более консервативной. В прошлом исследователи склонялись к тому, чтобы объединять ТМ и тыт языки на основании общей для них архаичности структуры словоформы, не предлагая при этом каких-либо общих фонетических инноваций. Наши результаты показывают, что перестройка структуры слога и тоногенез (отличительные черты вьетских языков) в значительной степени развивались в разных подгруппах независимо друг от друга, несмотря на многочисленные сходства в фонологических изменениях, которые способствовали появлению ареальной конвергенции. В работе также вкратце обсуждается вопрос о возможной прародине вьетской группы; в пользу ее локализации на севере свидетельствуют как старые лексические заимствования, так и археологические данные.

Ключевые слова: вьетские языки; австроазиатские языки; классификация языков; вычислительная филогенетика.