The article is accepted by the journal "Linguistics of the Tibeto-Burman Area". This is the pre-typesetting version.

Please cite this article as:
Wu, M.-S, Bodt, T. A, Tresoldi, T. (2022). Bayesian phylogenetics illuminate shallower relationships Trans-Himalayan languages in the Tibet-Arunachal area. Linguistics of the Tibeto-Burman Area. [forthcoming]

Bayesian phylogenetics illuminate shallower relationships among Trans-Himalayan languages in the Tibet-Arunachal area

Mei-Shin Wu[1], Timotheus A. Bodt[2], and Tiago Tresoldi[3]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Department of East Asian Languages and Cultures, SOAS University of London, London, United Kingdom

[3]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

January 20, 2022

Abstract

Kho-Bwa, Hrusish, Mishmic, Tani, and Tshangla are language clusters that have been recurrently proposed as subgroups of the Trans-Himalayan (also known as Tibeto-Burman and Sino-Tibetan) language family. Nonetheless, their internal classification, as well as the relation with each other and with other linguistic groups in the family, is hitherto unresolved. We use lexical data on these groups and dated phylogenies to investigate such internal classifications. We base our examination on previous research into the language family in the Tibet-Arunachal area, and follow a computer-assisted approach of language comparison to perform Bayesian phylolinguistic analysis. As earlier phylogenetic studies on this family included little data related to this geographic area, we took a subset of the best available dataset and extended it with vocabularies for the Kho-Bwa and Hrusish clusters, also including one Mishmic, two Tani, two Tshangla, and five East Bodish languages to cover the major languages and linguistic subgroups neighboring these clusters. Our results shed light on the internal and external classification of

the Kho-Bwa, Hrusish and Bodish languages, and allow us to share valuable experience on the extent to which similar approaches can be applied to the phylogenetic analysis of the Trans-Himalayan language family.

Keywords

Tibeto-Arunachal, Trans-Himalayan, Bayesian phylogenetic analysis, language classification, historical linguistics

1 Introduction

Linguists have been studying the relationships among Trans-Himalayan languages for several decades[1], pursuing questions on the processes of language change, the relations between individual languages and larger sub-groups, along with the origins, migrations, and dates of historical divergence of the speakers of these languages. Several Trans-Himalayan phylogenies have been proposed since the early 19th century (Leyden 1808; Shafer 1955; Benedict 1972; Burling and Matisoff 1980; Bradley 2002; Thurgood and LaPolla 2003; Sagart 2011; Blench and Post 2014; van Driem 2014; Matisoff 2015), but some of these classifications are based more on impressionistic grounds than on the results of traditional comparative linguistic methods. The diverse names proposed for the language family itself and their highly divergent sub-groupings show how scholars seem far from reaching a level of consensus comparable, for example, to the one found in Indo-European studies.

A solution that has been proposed frequently, although also far from unanimous, is the investigation by Bayesian phylogenetics. This method has allowed re-examination of topics that have been under discussion for considerable time. For example, Gray et al. (2009) suggested that the Austronesian languages originated in contemporary Taiwan about 5230 years ago, and that the social strategy and navigation technology played a significant role in their expansion. Recently, three different studies have applied a Bayesian phylogenetic approach to infer the internal structure of the Trans-Himalayan language family, with largely convergent results (Sagart et al. 2019; Zhang et al. 2019, 2020): all of them reported that the Sinitic subgroup was the first off-branch (sometimes along with Sal), subsequently locating the family's homeland in northern China. These results have encouraged us to turn our attention to more recent linguistic

[1] In this paper, we call the language family also known as Sino-Tibetan or Tibeto-Burman as "Trans-Himalayan" (van Driem 2007: 226 fn.7), recognizing the great diversity of languages spoken on both sides of the Himalayan range. We feel this name is more adequate in expressing such a diversity than alternative names that promulgate certain subgroups based on numerical or historical importance. Following the results of Sagart et al. (2019), on which we based our dataset, we adhere to the hypothesis that treats the languages from Arunachal Pradesh as a subgroup of the Trans-Himalayan languages, which can be analyzed at a shallower lever. Note, however, that due to the low support in some of the splits the same results could be interpreted as requiring the inclusion of Kiranti languages, whose paraphyly is still under debate, as per Gerber and Grollmann (2018) contra Opgenort (2005), and which we do not include in our data.

levels, as these macro-scale studies, focusing on the most ancient splits and their dates, don't fully analyze shallower linguistic divisions in the family. This holds especially true for areas where large numbers of highly divergent languages are spoken in confined geographical regions, such as in the Indian state of Arunachal Pradesh.

Arunachal Pradesh is located in the eastern Himalayas, bordered in the north by the Himalayan ranges and the Tibetan plateau, and in the south by the alluvial plains of Assam. The same major river links the area, flowing west to east across the Tibetan plateau and east to west across the Assam plains, and is known as the Yarlung Tsangpo in Tibet, as the Siang in Arunachal, and as the Brahmaputra in Assam. Linguistic diversity in this area might be partially explained by it having served as a mountain refuge to diverse and successive population strata that for millennia migrated from both the Tibetan plateau and the Assam plains, when other population strata moved into and settled across these more easily accessible, inhabitable, and arable stretches of land.

This narrative seems to confirm that the "Zomia" geographical area extends from Southeast Asia into the Himalayas. Zomia was first proposed by van Schendel (2002, 2007[2]) and further elaborated on by Scott (2009). However, we agree with criticism of the Zomian theory offered by authors like Michaud (2010), Lieberman (2010) and Brass (2012). According to them, the people now inhabiting Southeast Asia's mountain ranges may not always have "chosen" to migrate from their original homelands to avoid being "enslaved" by "nation states". Rather, they may have been forced out by more technologically advanced and numerous migrant populations, facing linguistic, cultural, and ethnic assimilation, or worse. Moreover, most of the modern nation states in this region did not even exist before the 17th century, and the current geopolitical boundaries only stabilized in the mid-20th century. The influence of the precursors of these modern nation states was highly area-specific, and there was not a simple relation of one-sided economic and cultural dominance over the people in the mountains, neither were these mountain communities isolated from the adjacent populations, as to a large extent they could determine the level of contact on their own terms. Their most common livelihood systems – shifting cultivation and extensive livestock herding, combined with a heavy dependence on foraging and hunting in

[2] Personal communication between Willem van Schendel and Jean Michaud in February 2008. See footnote 2 in Michaud (2018: 73 fn. 2).

the forest – was determined by the topography and climate of the mountain ranges they lived in, and was not a necessity to reduce domination and predation by the peoples and states in the plains.

Hence, although the entire area may have served as a refuge for various migrant populations, these groups may never have lived in secluded refuges which they defended against outsiders with whom they supposedly minimized every contact. These communities may have accepted later migrant populations and intermixed with them linguistically, culturally, and genetically, resulting in the high level of linguistic diversity we observe today. Though often described as inhospitable, the mountain ranges and rivers in the eastern Himalayas are not impregnable, and both mountain passes and river valleys have always served as gates and roads for human movement. Although some authors attribute the linguistic diversity in the mountainous regions of the Himalayas to geographic isolation and the prevalent socio-economic situation[3], we prefer to keep a more agnostic approach in which we also consider the possible influence of migration, language contact and other factors for which the topography of the area may not actually have served as an impediment.

The existing Bayesian phylogenetic studies were not adequate to evaluate this hypothesis, as they overlooked several of the linguistic groups in Arunachal Pradesh. Neither the Kho-Bwa nor the Hrusish clusters were included in the studies by Sagart et al. (2019) and Zhang et al. (2019), and in Zhang et al. (2020) they were only represented by two and three varieties, respectively. The latter study, being a macro-level one, does not provide much insight into the shallower phylogeny of these groups and only provides some general indications regarding their phylogenetic position, despite recent progress in unraveling the linguistic history of these clusters (e.g., Anderson (2014) and Bodt and Lieberherr (2015) for Hrusish; Lieberherr (2015), Lieberherr and Bodt (2017), Bodt (2019) and Bodt (2021) for Kho-Bwa).

To shed light on these groups, we first review the literature on the Kho-Bwa and Hrusish languages. We then describe the lexical material that we used, and the Bayesian phylogenetic methods that we employed. At last we present our findings, discussing the internal structure of the Kho-Bwa and Hrusish clusters and their affiliations within the Trans-Himalayan language family, along with interpretations for the positions of other neighboring groups like Tani and

[3] Such as Zhang et al. (2020: 5) who claim that "the Himalayan region maintained high levels of ethnolinguistic diversity" because it "limited opportunities for social contact and cultural diffusion [, leading to] rapid cultural diversification."

Mishmic. We discuss the usefulness of Bayesian phylogenetic analysis for these lower-level phylogenies, hoping to provide a simple explanation of the method while sharing insights into the opportunities and limitations of these methods to the study of the Trans-Himalayan languages.

2 Languages of Arunachal Pradesh and Tibet

Hazarika (2016, 2017) believes that Northeast India has been a corridor of population movement between South Asia and Southeast Asia since the late Pleistocene or early Holocene period (estimated between 12900 YBP and 11700 YBP). The Indian state of Arunachal Pradesh, bordered by Tibet (China) to the north, Bhutan to the west, Myanmar to the east and Assam (India) to the south, also falls in this proposed Northeast Indian corridor. In the western part of this state, the Tawang, Kameng, and Tenga river valleys are home to a surprising diversity of ethnolinguistic groups, whose languages and cultures are only now starting to be adequately described. We find several of these linguistic groups also in Bhutan and in Tibet, with relatively recent national borders separating people with a shared cultural and linguistic history.

Sun (1992: 80 and 1993: 11) was the first to suggest that the languages known to him from several descriptions from the Indian side of the border as Bugun, Sherdukpen, and Lishpa-Butpa could make up a new Tibeto-Burman group. He cautiously added Sulung to the group, based on data from the Chinese side of the border. Sun also provided the first linguistic evidence for the Trans-Himalayan affiliation of the languages of this cluster beyond lexical similarities, describing the regular correspondence between the Sulung voiced stop onset and other Trans-Himalayan nasal onsets. However, he remarked that the relationship of Sulung to the other languages "does not seem very close" (Sun 1992: 80 fn. 19), basing himself on some striking characteristics of this "obscure" language, such as "rich consonantal contrasts", "an impressive set of vocalic elements", "a rudimentary system of tones", and "a set of remarkable Austroasiatic phonological features". Sun (1993: 11) proposed the name "Bugunish" for the group constituted by Bugun, Sherdukpen, Lishpa-Butpa and, tentatively, Sulung. This group of languages gradually gained recognition among linguists, with van Driem (2001) first labeling it the "Kho-Bwa cluster" after his proposed reconstructed proto-words for "water" and "fire". Within this "enigmatic" cluster, van Driem included Bugun, Sulung, Lishpa, and Sherdukpen. On the basis of Rutgers'(1999) comparative vocabulary, van Driem (2001: 476-477) noted that "the Sulung

lexicon shares many peculiar traits with Sherdukpen and Bugun, but the sheer oddity of the Sulung lexicon has led many to entertain doubts about whether the language is Tibeto-Burman at all" and that "the Sulung are lexically the most aberrant, leading scholars either to suppose an overwhelming non-Tibeto-Burman substrate influence or to question whether Sulung belongs to the Tibeto-Burman family at all". Despite these doubts, the most commonly consulted handbooks (Burling 2003; Genetti 2016) and online language catalogues (Eberhard et al. 2019; Hammarström et al. 2021) list Kho-Bwa as a branch of the Trans-Himalayan family.

In 2005, an initially unpublished study by Abraham et al. (2018[2015]) offered many new lexical and socio-linguistic data on the various linguistic varieties of Western Arunachal. They identified "Chugpa" to be close to "Lishpa", and "Sartang" to consist of different varieties all close to "Sherdukpen". While noting the difference from other "Monpa" languages of the area, they accepted the Trans-Himalayan affiliation of all these languages. Matisoff (2009: 309), also noting the correspondence earlier identified by Sun, remarked that "in spite of Sulung's relatively poor score with respect to the 'stable' vocabulary [...], there are many clear Sulung reflexes of well-established [Tibeto-Burman]-roots, of all degrees of 'basicness'." However, Blench and Post (2014: 78,92) expressed skepticism about the affiliation of the entire Kho-Bwa clade, and indeed many of the languages of Arunachal Pradesh, to the Trans-Himalayan language family.

After 2012, several publications (Bodt 2014a,b; Lieberherr 2015; Lieberherr and Bodt 2017; Jacquesson 2015; Lieberherr 2017; Bodt 2020) provided more data on individual Kho-Bwa languages and on their internal and external classifications. Bodt (2014a,b) identified the Kho-Bwa languages by their most common autonyms: Puroik (Sulung), Bugun (Khowa), Sherdukpen, Sartang (Butpa), Duhumbi (Chugpa), and Khispi (Lishpa), also refining the data for the latter four and grouping them as the "Western Kho-Bwa" languages. Lieberherr (2015) provided the first description of the various Puroik varieties, showing their relationship through shared sound correspondences. Along with the correspondence of Trans-Himalayan bilabial nasal onset to Puroik bilabial stop onsets identified by Sun, Lieberherr (2015: 267-268) adduced a second defining phonological innovation in Puroik, the correspondence between the sibilant onset *s-* in other Trans-Himalayan languages (*th-* in the Kuki Chin languages) to vocal onsets in Puroik. Lieberherr and Bodt (2017: 38-40) showed that both of these defining sound correspondences hold for all the Kho-Bwa languages, lending evidence to the presumption that all the languages considered as part of this group are related Trans-Himalayan languages. In addition, the latter authors concluded that, in terms of core vocabulary, Kho-Bwa is a consistent group with three

sub-groups: Western Kho-Bwa, Bugun and Puroik. More detailed grammatical descriptions of the Kho-Bwa languages Sherdukpen (Jacquesson 2015), Puroik (Lieberherr 2017), and Duhumbi (Bodt 2017, 2020) have since been published.

Nonetheless, Post and Burling (2017) would again express doubt that Puroik is a member of the Trans-Himalayan family. Neither Blench and Post (2014) nor Post and Burling (2017) presented any evidence – linguistic or otherwise – that showed that the languages of the Kho-Bwa cluster, and Puroik in particular, are indeed not Trans-Himalayan languages. We do not immediately reject the hypothesis that Puroik or even all the Kho-Bwa languages are descendants from non-Trans-Himalayan substrate languages that have been in intense contact with Trans-Himalayan languages. This may be adduced when taking only lexical data into account. However, we believe that all the other linguistic evidence that has been presented till date strongly favors both the internal coherence of the cluster and its Trans-Himalayan affiliation by descent. Phonological, lexical, and grammatical oddities of these varieties, and of the Puroik ones in particular, may stem from a variety of reasons such as linguistic substrates, a long time depth of divergence and subsequent differentiation (either in isolation or through language contact), and admixture with diverse subsequent migrant groups. Indeed, considering the linguistic evidence in combination with the until recently dominant hunter-gatherer lifestyle of the Puroik and the continued dependence of most agricultural societies in Arunachal on shifting cultivation and extensive livestock herding, including that of the mithun (*Bos frontalis*), it would be tempting to follow Blench and Post's (2014: 90-91) hypothesis on the origin and dispersal of the Trans-Himalayan languages from the eastern Himalayan regions, despite all the aforementioned macro-level phylogenetic studies supporting the more traditional views of Sinitic as the first off-branch of the family and the Yellow river basin as its homeland (Sagart et al. 2019; Zhang et al. 2019, 2020).

The Puroik generally consider themselves to be the original inhabitants of the area they inhabit, preceding Tani and Hrusish speakers that would have arrived later (Stonor 1952; von Fürer-Haimendorf 1982). The Bugun claim a close relationship to the Puroik (Stonor, 1952: 949; Soja, 2009: 17). Western Kho-Bwa speakers claim a mixed origin, initially from a migratory group from the East related to the Puroik and Bugun, mixing with a migratory group from the North, perhaps a Pre- or Proto-Bodish group, then followed by subsequent population admixtures in their respective locations (Rinchin 2011: 27-53; Bodt 2014a). Whereas we can take none of these

origin and migration stories at face value, we should keep them in mind when further analyzing the linguistic history of the communities that tell them.

Besides the Kho-Bwa languages, several other languages that are confirmed or presumed as Trans-Himalayan are spoken in western Arunachal Pradesh. Among these are varieties of the Tshangla, Tani, East Bodish, and Hrusish groups. Tshangla has its heartland across the border in southeastern Bhutan, whereas the East Bodish languages have their center of gravity in northeastern Bhutan. The Tani and Hrusish languages are spoken to the east of the Kho-Bwa speech area, as per Bodt (2014a). In order to increase the possibility of highlighting relationships between the Kho-Bwa varieties and these groups, we increased the representative sample of the Mishmi, Tshangla, and Tani subgroups, also adding representative samples of the East Bodish and Hrusish languages. Although the internal phylogenies of these two groups were not an initial goal of our research, we seized the opportunity presented to us to provide a more detailed overview of the linguistic phylogeny of the Bodish and the Hrusish groups, as well as of the larger Tibet-Arunachal area.

The Hrusish languages, including the Miji varieties of East and West Kameng and Hruso (Aka), were first identified as a subgroup by Shafer (1947, 1955) based on Hruso and West Kameng Miji data. To these, Sun (1993: 348) added Bangru. Similar to their doubts about the internal coherence and the Trans-Himalayan affiliation of the Kho-Bwa languages, Blench and Post (2014: 78,92) also expressed reservations about the Hrusish languages, and, in particular, about the position of Hruso itself. In their description of Bangru, Bodt and Lieberherr (2015) presented initial evidence that Bangru, the Miji varieties, and Hruso Aka could indeed belong to a single linguistic sub-group within the Trans-Himalayan family. The study by Zhang et al. (2020) placed the two Hrusish varieties of Aka (Hruso) and Miji together with Kho-Bwa and subsequently placed these in a larger clade together with the Sal languages of Northeast India, which includes the Bodo-Garo and Northern Naga languages. Local origin and migration histories (e.g., Grewal 1992 and Dusu 2013) present a very diverse and mixed picture of the ethnic and linguistic origins of the individual Hrusish varieties, including elements from the East, from the Brahmaputran plains in the South and from the North.

Although decidedly considered a member of the Trans-Himalayan phylum, the exact phylogenetic position of the large Tshangla group is still unresolved. Most linguists (Shafer 1955: 100-101; van Driem 2001: 991) accord Tshangla a relatively independent position close to or together with the Bodish languages, with van Driem coining the term “para-Bodish” to refer to

the language group; other authors, such as Thurgood (2003: 9-10), even place Tshangla firmly among the Bodic languages. A genetic relation between Tshangla and the Lolo-Burmese languages (Bodt 2012: 211) has not been further substantiated. Among the existing macro-level phylogenetic studies, the position of Tshangla does not reach an agreement, either. Sagart et al. (2019) placed Tshangla with the Tani and Mishmi languages of Arunachal. Zhang et al. (2019) and Zhang et al. (2020) placed Tshangla as an early offshoot of the Bodish branch.

More certainty exists about the East Bodish languages. Shafer (1954) made the first hypothesis about the East Bodish languages, with subsequent work by Michailovsky and Mazaudon (1994) and van Driem (2001: 380). More recent advances have been mainly to the credit of work by Hyslop (2013, 2014). The East Bodish languages are considered earlier offshoots of the Bodish branch of Trans-Himalayan. Hyslop and d'Alpoim Guedes (2020) tentatively dated this split to 2,500 years ago, locating the homeland in the southernmost parts of the Tibetan plateau and its Himalayan highland interface zone. On the basis of shared cultural and linguistic traits, Huber (2020) also hypothesized a common ancestral heritage with the earlier speakers of the Qiangic and Naic languages spoken to the East. The position of East Bodish as an earlier offshoot of the Bodish languages is supported by Zhang et al. (2020) and Zhang et al. (2019). As for Tani, since the reconstruction of Proto-Tani by Sun (1993), the Trans-Himalayan affiliation of the group has not been questioned, although its precise phylogenetic position within the family has not yet been agreed upon.

To complement the major linguistic subgroups of Arunachal Pradesh to the north of the Lohit-Brahmaputra river system, we included Kaman Mishmi (hereinafter called Gémàn[4]) in addition to the Yidu (hereinafter Yìdū) and Darang Mishmi (hereinafter Dáràng) varieties already present in the Sagart et al. (2019). The phylogenetic relationship between these languages, and their affiliation with the Trans-Himalayan language family, are as disputed as those of the Kho-Bwa and Hrusish languages (Blench 2017: 3,14). However, besides some perfunctory remarks, we will not pay more attention to these languages.

[4] We use Pinyin in the original Chinese sources to refer to the language names.

3 Material and Methods

3.1 Lexical data

The lexical data in our study comprise 86 linguistic varieties, i.e., "doculects" (Good and Cysouw 2013), for our purposes taking an agnostic position on whether these varieties are "languages" or "dialects". Besides the 49 linguistic varieties represented in the dataset from Sagart et al.'s (2019) study, we drew additional linguistic material from various primary and secondary sources, always selecting the same set of concepts used in Sagart et al. (2019). We based such a concept set on the Concepticon database (List et al. 2021). Whereas we prioritized data from published sources, we had to rely on primary data when such sources were not available or when we wanted to extend the concept coverage for specific linguistic varieties. In the following paragraphs, we explain which doculects were added and which additional sources we consulted, providing a complete description of our material in section S2 of the supplementary information.

3.1.1 Kho-Bwa

The Western Kho-Bwa lexical data include published material (Bodt and List 2019; Bodt 2020) and is supplemented by primary data from unpublished fieldwork by TAB. We collected the data on Bugun varieties from two different sources: Dikhyang Bugun data came from Bodt (2017) and primary resource with additional data from an unpublished database by TAB[5], while we incorporated forms for the other five varieties from Abraham et al. (2018[2015]). Data on Puroik languages include two Eastern Puroik variants spoken in Arunachal Pradesh (Soja 2009; Remsangpuia 2008) and one Western Puroik variant (Lieberherr 2017). A Puroik variety recorded in the early 1990s by Sūn et al. (1991), with additional forms by Lǐ (2004) was also included, even though it is believed that there are no speakers of Puroik in Tibet anymore.

3.1.2 Hrusish

No linguistic variety thought to belong to the Hrusish subgroup was sampled in Sagart et al. (2019). We based our data on Abraham et al. (2018[2005]) in a cross-linguistic data format

[5] Some of the concepts included in Sagart et al. (2019) were not included in the lexical data in Bodt (2017, 2020), but we were able to extend the coverage from unpublished fieldwork data.

(Abraham et al. 2019), the most extensive collection of Hrusish varieties. We extended the dataset with primary data on Bangru from Lieberherr and Bodt (2017) and an unpublished dataset by TAB. In addition, for concepts of Nafra Miji and Jamiri Hruso Aka not provided by Abraham et al. (2018[2005]), we used forms from Simon (1979) and Simon (1993[1970]), respectively.

3.1.3 Tshangla

The dataset by Sagart et al. (2019) included a single Tshangla variety, Mòtuō Ménbā also known as Pemakö Tshangla which is spoken in southeastern Tibet. We extended the sampling for this group, adding primary data from the two major Tshangla varieties, Bhutan Tshangla (i.e., Tshangla as spoken in Bhutan) and Dirang Tshangla (i.e., Tshangla as spoken in the Dirang area of West Kameng district in Arunachal Pradesh), using datasets assembled by TAB. We hoped to gain a preliminary insight into the phylogenetic position of Tshangla, assessing whether it associates more closely to the Bodish languages (the hypothesis that has most support in the literature) or to the languages of Arunachal Pradesh.

3.1.4 Bodish

The dataset by Sagart et al. (2019) had a good representation of Central Bodish (Bodic) linguistic varieties, with five Central Bodish varieties. The dataset also included three Western Himalayish varieties. We extended the dataset with forms from the under-researched East Bodish languages to extend this Bodish clade. Ideally, we would have taken Tawang Monpa, the primary East Bodish contact language for the Kho-Bwa varieties, to represent East Bodish. However, no reliable and complete lexical datasets of the varieties of Tawang Monpa are available. For the related varieties spoken across the border from Tawang in Tibet, we used data from Lù (1986) and Lù (2002) on Mama Cuona Menba (Mámǎ Cuònà Ménbā). We also added data from Lù (1986) and Lù (2002) on Wenlang Cuona Menba (Wénlǎng Cuònà Ménbā) which is spoken in southeastern Tibet, which is thought to be related to Dzalakha.[6] The Dzalakha data come from

[6] In hindsight, we could better have used the Lù's Bāngxīn data: TAB's sources state that the people of 文朗 Wénlǎng, Tibetan wan-lang, local name [uŋlaŋ] village came from the Dzalakha speaking areas of eastern Bhutan, whereas the people of 帮辛 Bāngxīn, Tibetan spang-zhing, came from the Tawang area, in particular the Pangchen, Tibetan spang-chen valley on the border with Tibet. Unfortunately, the Bāngxīn data in Lù (2002) lack 76 concepts from our original 250 concept list, and Lù (1986) does not have Bāngxīn data.

Dzongkha Development Commission (2017). Data from Bumthang are primarily from van Driem (2015), representing the Chos-'khor dialect, with additional data from Dzongkha Development Commission (2018) describing the Chu-smad[7] dialect, along with primary data for the Tang dialect[8]. The data on Khengkha are from Yangzom and Arkesteijn (1996), with additional unpublished primary data by TAB.

3.1.5 Tani and Mishmic

The large Tani group was only represented by Bokar Luoba (Bógāěr Luòbā) in Sagart et al. (2019), with data originally from Huáng and Dài (1992). To lessen the imbalance of data selection, we added lexical data on Galo (Post 2007) and Tangam (Post 2007) to extend the Tani subgroup. Although Western Tani (also known as Nyishi, Bengni or Bangni) is the primary Tani contact language for Puroik and other Kho-Bwa languages, we chose to extend the Tani group with Tangam and Galo: their sources (Post 2007, 2017) are by far the most complete and reliable descriptions of an eastern (Tangam) and a western (Lare Galo) Tani language, with an easily accessible and reusable lexicon. Furthermore, Post's (2017) publication has the additional benefit of providing the ProtoTani reconstructions by Sun (1993: with updates by Post): considering how Lare Galo has undergone considerable phonological change, the Proto-Tani reconstructions made it much easier to determine cognates.

The Mishmic group was represented in Sagart et al. (2019) by Yìdū and Dáràng. We added the third linguistic variety, that is sometimes classified as Mishmic: the Gémàn language from Sūn et al. (1991), obtained from the STEDT database (Matisoff 2015).

[7] This source is primarily a record of local household items, food items, plants and animals and contains few other parts of speech besides nouns.

[8] We realize that mixing dialects is methodologically problematic. However, in this context we believe it is acceptable given the limited data currently available.

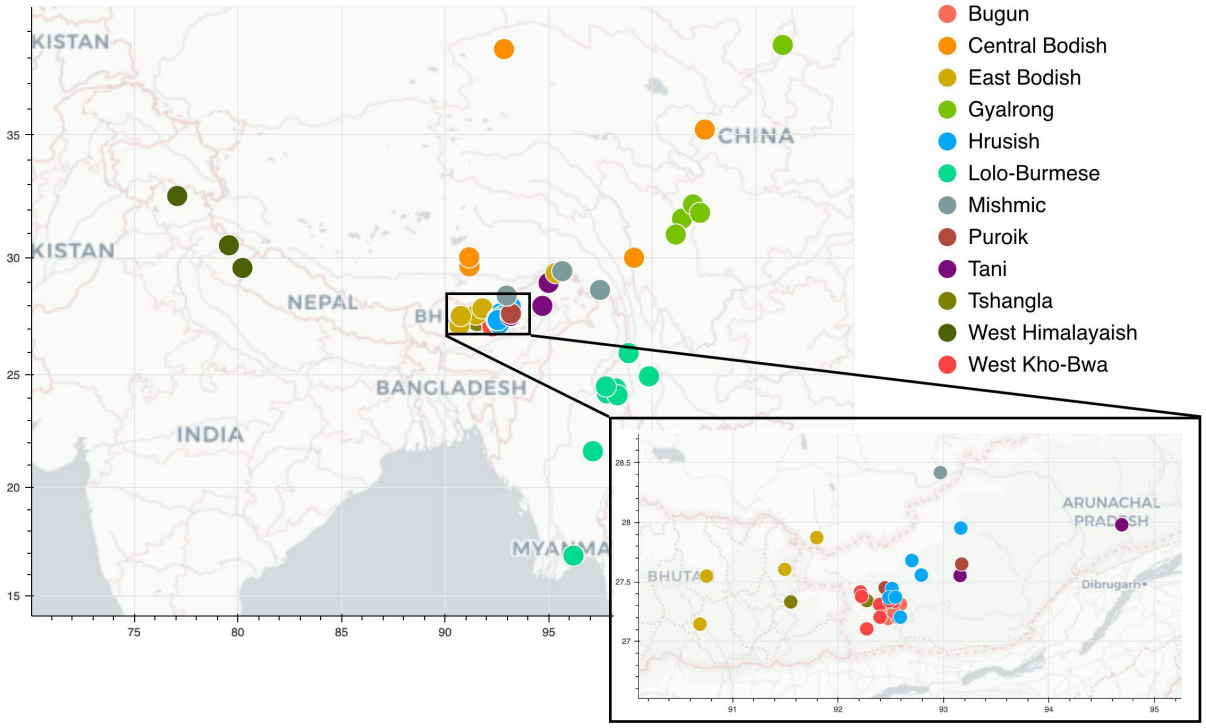

Figure 1: Languages in Arunachal and Tibet in our selection.

3.2 Concept selection

Closely following the selection and decisions of Sagart et al. (2019), our dataset has 250 concepts, but, as expected, not all concepts are represented in all doculects. While some concepts are found in the data of most or all varieties, often those expressing cross-linguistically common concepts that are easy to elicit, other more region- and culture-specific concepts are only found in a subset of our varieties. Section S3 in the Supplementary Information tabulates the coverage of our concepts, i.e., in how many of the 86 linguistic varieties in our dataset each concept is found. Compared to the original dataset by Sagart et al. (2019), the coverage of several of the concepts in the geographic area of our interest increased, showcasing the usefulness of consulting lexicons, dictionaries, and grammars of individual languages. This also attests to the substantial benefit of studies such as Sagart et al. (2019) that provide their data for replication and extension in digital formats designed for these purposes.

We computed the mutual coverage rate per concept, the total number of distinct cognate sets per concept, the number of singletons per concept (i.e., the number of cognate identifiers that are found only once) and other exploratory statistics; we provide those in the Supplementary

Information in section S3. The mutual coverage rate was the only criterion to prepare the data for the phylogenetic analyses. In all cases, no matter the subset of linguistic varieties involved, we consistently applied a concept filtering criterion of 80 per cent.

3.3 Cognate judgments

As discussed in the following section, the method we use generates linguistic trees from events of lexical substitution in base vocabulary: cases in which the most "neutral" word to express a concept is replaced in common usage. Unlike in most traditional approaches to historical linguistics, inferred sound changes are not evolutionary characters in themselves, but are pieces of evidence to detect words grouped by a common origin.

Cognate decisions, provided either by experts or by automatic methods, will to a large extent determine the outcomes. As such, the most crucial and critical step in our workflow is identifying which groups of forms can be traced to a common ancestor, comprising "cognate sets". It is the most subjective step in the entire workflow: two forms determined as cognates by one expert may be considered independent by another expert or even by the same expert under different circumstances. While methods for automatic judgment have been used to assist experts and streamline their work (List et al. 2018), the only way ahead is through achieving some level of consensus through publications of views and alternate views. Unfortunately, the study of the linguistic history of Trans-Himalayan has not reached such a level of consensus even for the concepts that have been researched the most, like numerals and body parts. Facing this inherent weakness of cognate judgment, the approach by Sagart et al. (2019), which makes all the judgments available in a format that can be easily reproduced, replicated, cross-checked and modified, is a huge and commendable step forward.

Overall, we used the cognate judgments by Sagart et al. (2019) with little change. There have been a few changes based on our insights into certain individual varieties, which can be identified by comparing the two open datasets. For the varieties that we added, we based cognate decisions on the best knowledge and insights at the moment. Essential guiding documents for these decisions have been the reconstructions of Proto-Western Kho-Bwa (Bodt 2019, 2020), Proto-Tani (Sun 1993; Post 2007), and Proto-Hrusish (Bodt and Lieberherr 2015).

3.4 Loanword handling

There is currently no consensus on how to treat loanwords in Bayesian phylogenetic analysis. In particular, there seems to be no agreement on how to treat nativized loanwords that may have taken part in the common sound changes affecting the receiving language and its descendants. While some authors remove loanwords (a practice almost impossible to assess when they do not make their models public), when considering phonological changes besides lexical replacement, loanwords can, in fact, give important insights into the relationships between different languages.

Therefore, we first marked the loanwords in our data with a "LOAN" flag whenever possible. In our initial exploratory analyses, we ran each experiment twice, once including the forms that we had marked as loans and once excluding them. We visualized both phylogenies and found that in all cases the results did not differ significantly. Considering the incomplete knowledge of the linguistic history of many of the languages in our dataset, determining whether attested forms are cognates or loans from languages in the same family involves subjective decisions of a non-trivial nature, which may frequently be biased by an implicit expectation of the results. Therefore, we decided that the most parsimonious decision would be to not add another level of subjective decisions, thus including all the attested forms irrespective of their suspected origin in our analyses.

3.5 Bayesian phylogenetic analysis

Bayesian inference in phylogenetic analysis is an analytical method for incorporating prior information and model likelihood to deduce the evolutionary relations among taxa ("language varieties"). First introduced for molecular and biological studies, it has been gaining traction in historical linguistics in recent times (Greenhill et al. 2020), building upon the foundation of the Markov Chain Monte Carlo (MCMC) algorithmic implementations. These methods have recently risen in popularity, as they allow to detect the tree-like signal of vertical transmission even in cases where the language evolution involves many horizontal events, such as lexical borrowings, population admixture, and so on, with a Bayesian approach forcing scientists to declare quantitatively their assumptions and analyses' limits as "priors". In essence, Bayesian phylogenetics performs a probabilistic inference of trees and parameters of a model: the "posterior" probability of a vast series of trees is calculated as a function of the "prior" probability of a tree and the "likelihood" of the data available within a specific evolutionary model and its

parameters. In other words, the method collects with a statistically-oriented sampling a set of most likely trees when both the linguistic data and a model of linguistic evolution are considered, giving a "score" of how likely each tree is when all elements are considered.

Among the advantages of this method, it allows to date splits in these trees, especially when historical languages can calibrate probability distributions in terms of expected changes over a certain time interval. In the end, different processes can be used to combine the best trees (i.e. those with the highest probabilities) into a summary tree ("consensus") or into a representation that highlights conflicting signals (i.e. groups of trees that illustrate different and not reconcilable evolutions, but with comparably high probabilities). As a complex topic involving expertise in quantitative methods and familiarity with alternative evolutionary models, notably when applied to historical linguistics, phylogenetics needs specific works for an exhaustive summary (e.g. Gamerman and Lopes 2006; Gilks et al. 1996; Ravenzwaaij et al. 2018; Greenhill et al. 2020).

We base our Bayesian phylogenetic analysis on discrete characters of lexical replacement. Despite some researches experimenting with different phenomena, chiefly those most frequent in "traditional" historical linguistics such as phonological innovations, the most accepted approach is this usage of lexical substitutions in the expression of "basic concepts": each time a new word (i.e., a word member of a different cognate set) replaces a previous one as the most "neutral" way of expressing a concept, we have an evolutionary event of "lexical replacement", whose effects will be transmitted to descendant languages. Cognate sets given by linguistic experts are converted into a binary matrix where 1 encodes the presence of the cognate found in a given language, 0 encodes the absence of such cognate, and question marks represent missing information (such as for non-exhaustive language data).

Once the binary matrix is ready, we express our assumptions with statistical distributions, the "priors". These priors contain factors related to the family's evolution, such as rates of lexical substitution (a probability distribution of how often the word for expressing a concept changes) and of language birth and death. At the base, we use a binary covarion model (Huelsenbeck 2002) to infer the trees. This model introduces a "fast" or "slow" state of change, which controls the transition rates between presence or absence of a cognate (Maurits et al. 2017). We set the visible frequencies as 0.99 and 0.01, so that the state of each cognate changes from absence to presence faster than the opposite. We modeled the branch lengths' development following a *"relaxed*

molecular clock" with a log-normal distribution (Drummond et al. 2006)[9], following an underlying belief that lexical changes do not follow a fixed rate through time and that the rate of evolution of a branch is autonomous from the rates of its mother, sister, and daughter branches. This assumption seems closer to the real-world scenario and has been frequently adopted to produce language phylogenies.

In addition, we calibrated the taxa and set a time frame on splits and the root so that the branch lengths are calculated in proportion to time. Calibrations tend to rely on written records and archaeological excavations, but archaeological research on the Tibetan plateau, and especially in Arunachal Pradesh, is still in its infancy. Most of the languages of Arunachal Pradesh were, even until recently, spoken by hunter-gatherer or early agriculturalist societies, a fact which, combined with the hot and humid climatic conditions, left us without written records and with limited archaeological data, save for unstratified, scattered and undated stone tools (Hazarika 2017; Ashraf 1990; Tada et al. 2012). Because of this limitation, it is hard to provide calibration dates to the Kho-Bwa and Hrusish languages.

To overcome this obstacle, we used the calibration dates provided by Sagart et al. (2019) for the Old Tibetan, Old Burmese and Tangut languages. Figure 1 lists the language subgroups which comprise our analysis and their sampled locations. To set up the model for Tibet-Arunachal phylogeny, we used aforementioned priors and the calibration dates on Old Burmese, Old Tibetan and Tangut. To set the root date, also known as tree height, we consulted the phylogeny that was offered by Sagart et al. (2019) and Blench and Post (2014: 18), and set a uniform distribution between 5000 to 6800 YBP. The phylogeny in Sagart et al. (2019) shows the origin of languages in Tibet and Arunachal is dated around 5000 - 5500 YBP. Furthermore, Hazarika (2016) indicated that yak domestication on the Tibetan Plateau took place around 6700 YBP. A uniform distribution shows that our prior treats all the time points between 5000 to 6800 YBP as equal. The common ancestor of the selected languages in the Tibet-Arunachal phylogeny could, in theory, appear at anytime between 5000 to 6800. This is the optimal solution when there is no other study allowing us to favor any particular hypothesis at the time when we conducted the

[9] A strict clock assumes a constant mean rate of change across all branches, being somewhat similar to glottochronology, while a relaxed clock allows different rates of change for each concept in each branch.

experiment[10]. We set a normal distribution for Proto-East-Bodish with the mean at 2500 YBP as Hyslop (2013) stated that East Bodish originated at this date[11]. We assigned a normal distribution with the mean at 1350 YBP for the proto-Tani language, as the time frame (5th century AD - 7th century AD) is indicated in Krithika and Vasulu (2018).

The Bayesian phylogenetic analysis mimics how the numbers of lineages can change in a time frame with a "Birth–Death Skyline Serial model" (Stadler et al. 2013; Gavryushkina et al. 2014), so each taxon in the model can lead to a specification event, or the taxon can become extinct. But we specifically requested the model not to consider the extinction rate.

We performed the phylogenetic inference, running every model for 10^8 iterations, and we sampled trees every 5000 iterations. After running the analysis, we discarded the first 10 percent of the iterations ("burn-in"). The reason for setting a burn-in is that the initial likelihood is low because Bayesian analysis starts from a random tree. The algorithm will enter a high-probability zone of the posterior after a certain amount of iterations, spending the rest of the study in such a high-probability zone. We only select the sampled trees generated from the high-probability zone to avoid the random trees which were generated by the initial states (Nascimento et al. 2017).

From the major overall Trans-Himalayan (Sino-Tibetan) phylogeny (see supplementary S4), we observed that the deepest level (not the root) forms a binary structure. Therefore, we selected the big clade that contains Kho-Bwa, Tshangla, Mishmic and Tani languages. In addition to the selected languages of Tibet and Arunachal, the Lolo-Burmese and rGyalrong languages also form part of the same larger clade. For that reason, in our subsequent lower-level tree, we included these languages as well. Section 4 shows the consensus tree of Tibet-Arunachal phylogeny. We display all the sampled Tibet-Arunachal phylogenies in a *DensiTree* visualization in the Figure

[10] A more comprehensive summary about the yak domestication on the Tibetan Plateau can be found in Jacques et al. (2021).

[11] Here, we rely exclusively on Hyslop's assumption that East Bodish is, indeed, a valid taxon with the ancestral language having an age of approximately 2500 YBP. Our earlier modeling without monophyletic constraints actually showed that East Bodish is a polyphyletic group (see supplementary SS4); however, as this may be due to sampling bias (three of the four Dakpa-Dzala varieties are from Chinese sources in Tibet) or intense language contact, we did not take those results into consideration during further modeling.

2 in supplementary section S5 (Bouckaert 2010). In addition, the complete Trans-Himalayan phylogeny is also shown in the section S4 in supplementary.

3.6 The workflow

The workflow comprises several software packages for data management and curation, Bayesian analysis, and visualization. Our raw data is stored in the Cross-Linguistic Data Format (CLDF, Forkel et al. 2018). The merged data is a *LingPy* wordlist format (List et al. 2019), which was generated from our CLDF dataset via the *CLDFBench* toolkit (Forkel and List 2020) with the *pylexibank* plugin (Forkel et al. 2021). TAB made the cognate judgments and lexical data annotations for the additional languages with the help of the *EDICTOR* web application (List 2021).

After the cognate judgments, we coded Python scripts to convert the data into a distance matrix and analyze the resulting neighbor-net (Bryant and Moulton 2004) with *SplitsTree 4* (Huson and Bryant 2005). We coded additional Python scripts to draw a language subset, filter concepts with 80 percent or above mutual coverage, and build data files used for generating the Bayesian phylogenetic models via *BEAUti* (Drummond et al. 2012) and a customized tree prior template. We computed all the Bayesian phylogenies via *Beast2* version 2.6.5 (Bouckaert et al. 2019), also writing Python scripts for the post-analysis.

We used *DensiTree* version 2.2.3 (Bouckaert 2010) to visualize the remaining sampled trees and to inspect the well-supported clades and conflicting signals. All images exported from *DensiTree* are provided in the supplementary. We used *TreeAnnotator* (Drummond and Rambaut 2007) to compute the maximum clade credibility trees. The algorithm calculated the node heights, which have either the maximum sum of posterior clade probabilities as the consensus tree or rescale the phylogeny to reflect the posterior mean. We then used *ggtree* (Yu 2020), a R library, to visualize the consensus tree. For the post-analysis, we calculated the amounts of shared cognates between two varieties and repeated this calculation through all the language pairs in the Tibet-Arunachal phylogeny. We used *seaborn* (Waskom 2021), a Python library, to visualize the shared cognate counts with a heatmap, and followed experts' grouping to arrange the languages in the heatmap. In addition, we also visualized the shared cognate counts for the entire Sino-Tibetan language data. The two heatmaps are presented in the supplementary.

We designed our workflow on the basis of FAIR principle guidelines (Wilkinson et al. 2016; List et al. 2021). Therefore, the data, the entire workflow, and the experiments are provided in the supplementary under an open license. Our supplementary is archived on Open Science Framework (OSF)[12]. In addition, we archive our raw data in .tsv, .xslx, and .ods formats on Zenodo so that users can inspect the data with Excel or LibreOffice.[13]

4 Results

Figure 2 shows the phylogeny of Trans-Himalayan languages of the Tibet-Arunachal area, with internal nodes ("splits") numbered from 0 (the root) to 60. Table 1 lists the time estimations in years before present (YBP) and the posterior support of the corresponding internal nodes in Figure 2; the posterior indicates the percentage of trees in the sampled set, after filtering and burn-in are performed, in which the split is observed. Most of the posteriors of the phylogeny are very high with the exception of two branching events, (a) Tshangla as the ancestral split among the 6 Bodish subgroups and (b) the splitting between Tangut, Japhug and Maerkang rGyalrong languages. Visual analysis of Figure 2 in the supplementary suggests that the low posterior among rGyalrong languages is due to the difficulty of internally resolving the clade, even though it is well-supported as a clade without major conflicting signals from other groups. Since the rGyalrong languages are well grouped together and it is not the focus of our current experiment, we defer this issue to future studies.

The root of the Tibet-Arunachal phylogeny is estimated at 6149 YBP (node 0, 95% highest posterior density (HPD): 5256 - 6800 YBP). In the Arunachal clade, the most recent common ancestor (MRCA) of the Hrusish and Kho-Bwa languages formed at 4092 YBP (node 32, 95% HPD: 2967 - 5255 YBP). The diversification of Kho-Bwa languages started at 2843 YBP (node 39, 95% HPD: 1996 - 3747 YBP) and the Hrusish languages started branching later, at 1846 YBP (node 33, 95% HPD: 1121 - 2674 YBP). The internal structure of Kho-Bwa agrees with previous findings that reported that Western Kho-Bwa is the ancestral split followed by a separation between Bugun and Puroik. The branching events at the shallowest layers of Kho-Bwa languages are all placed within the recent 1000 years, reflecting the dialect continuum within the three main Kho-Bwa language groups Bugun, Puroik, and Western Kho-Bwa. The internal structure of

[12] The project repository is archived on Open Science Framework (see supplementary S1).

[13] DOI: 10.5281/zenodo.5554780

Hrusish is also in agreement with earlier linguists' findings that the diversification started with the split of the ancestor of Hruso Aka. The later branching events of Bangru followed by the Dammai and Namrei languages similarly matches the current classifications. Tani and the Mishmic languages are language subgroups that are genealogically the closest relatives of Kho-Bwa and Hrusish. This entire clade is well correlated with the geographic location of the respective speakers within Arunachal Pradesh.

In the Tibet clade, the ancestor of the Tshangla group split from the other languages about 5700 YBP (node 1, 95% HPD: 4413 - 6530 YBP), but the internal diversification of the Tshangla varieties started much more recently at 824 YBP (node 20, 95% HPD: 475 - 1192 YBP). As mentioned, Bradley (1997) and van Driem (2014) considered Tshangla a subgroup of the Bodic group, while Hammarström et al. (2021) places Tshangla in the same clade with East Bodish. Our phylogeny does not support either of these classifications; furthermore, by inspecting the conflicting signals via the *DensiTree* software (Bouckaert 2010), we observed conflicting signals that link Tshangla with other languages spoken in Arunachal Pradesh (see figure in section S5). We elaborate on this finding in the discussion.

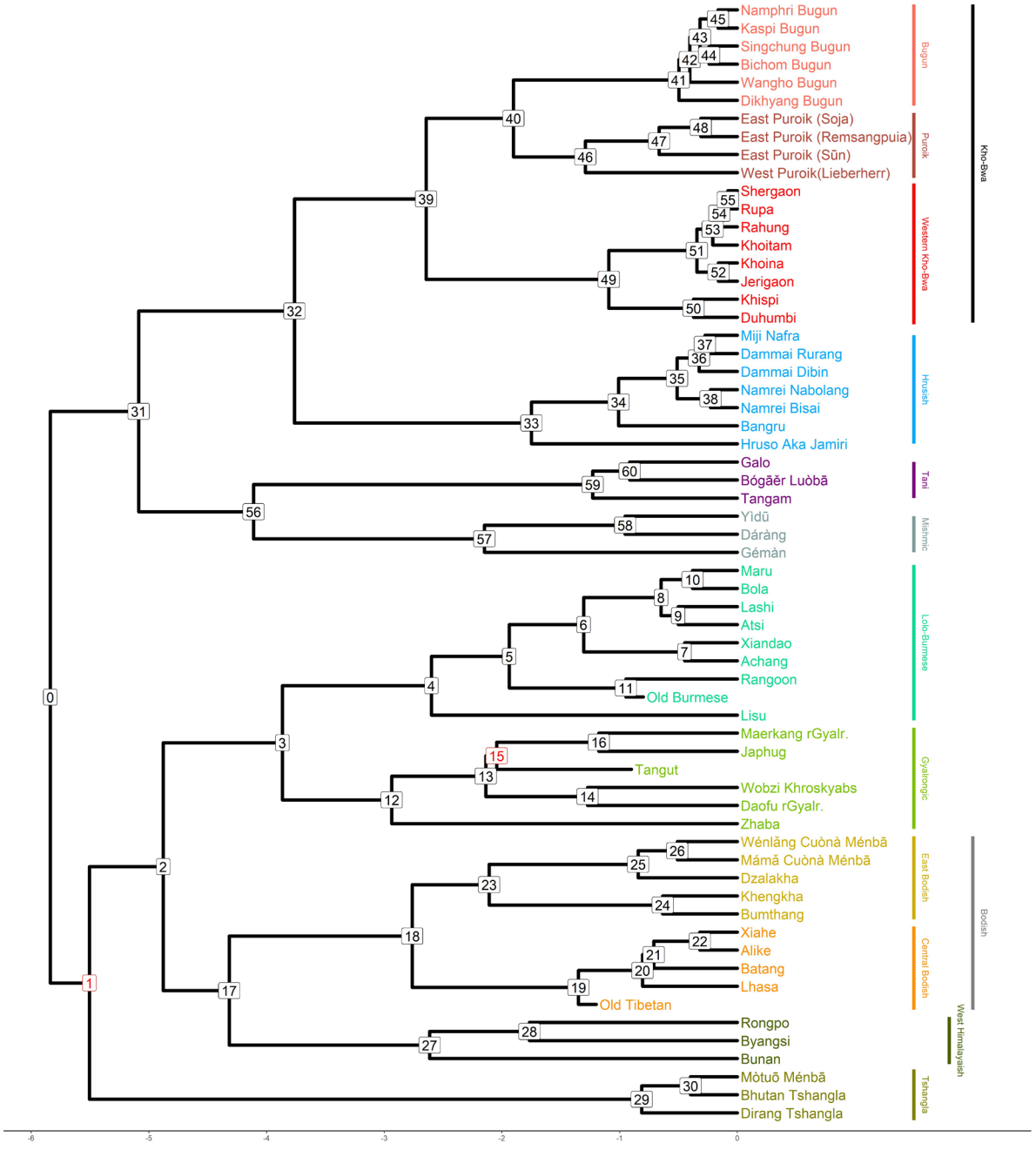

Figure 2: Phylogeny of Trans-Himalayan languages of Tibet-Arunachal

ID	Node age (95% HPD) unit:YBP	Posterior	ID	Node age (95% HPD) unit:YBP	Posterior

0	6194 (5256 – 6800)	1	31	5525 (4204 – 6533)	0.67
1	5700 (4413 – 6530)	0.62	32	4092 (2967 – 5255)	1
2	5013 (3808 – 6077)	0.95	33	1846 (1121 – 2674)	1
3	4015 (2976 – 5031)	1	34	1056 (579 – 1602)	1
4	2696 (1860 – 3551)	1	35	532 (286 – 802)	1
5	2013 (1398 – 2681)	1	36	340 (160 – 546)	1
6	1351 (849 – 1881)	1	37	286 (105 – 443)	0.64
7	468 (179 – 809)	1	38	236 (68 – 445)	1
8	660 (370 – 970)	1	39	2843 (1996 – 3747)	1
9	512 (209 – 764)	0.84	40	2036 (1356 – 2749)	1
10	388 (154 – 652)	1	41	507 (274 – 770)	1
11	948 (815 – 1136)	1	42	412 (222 – 607)	0.80
12	3071 (2194 – 3960)	1	43	326 (166 – 479)	0.79
13	2234 (1597 – 2909)	1	44	248 (73 – 336)	0.68
14	1349 (699 – 2024)	1	45	169 (64 – 293)	1
15	2136 (1459 – 2680)	0.45	46	1371 (830 – 1956)	1
16	1234 (564 – 1940)	1	47	699 (355 – 1073)	1
17	4486 (3250 – 5591)	0.89	48	326 (125 – 538)	1
18	2855 (2036 – 3703)	1	49	1143 (592 – 1800)	1
19	1355 (1205 – 1564)	1	50	378 (119 – 695)	1
20	824 (475 – 1192)	1	51	354 (193 – 541)	1
21	733 (359 – 1020)	0.57	52	167 (53 – 301)	1
22	333 (117 – 587)	1	53	217 (108 – 337)	1
23	2158 (1494 – 2831)	1	54	159 (76 – 255)	0.99
24	644 (250 – 1138)	1	55	84 (33 – 147)	1
25	869 (424 – 1137)	1	56	4806 (3405 – 6029)	0.91
26	526 (203 – 913)	0.99	57	3107 (1944 – 4311)	1
27	2733 (1783 – 3732)	1	58	1225 (581 – 1977)	1
28	1880 (1096 – 2712)	1	59	1282 (715 – 1862)	1
29	838 (361 – 1415)	1	60	955 (359 – 1365)	0.69
30	411 (141 – 751)	1			

Table 1: The posterior and the node age (95% height) in our phylogeny. The internal nodes received posteriors under 0.5 are marked in red.

5 Discussion

5.1 Findings and interpretations

The emergence of the Trans-Himalayan languages in Tibet and Arunachal Pradesh is estimated at 6149 YBP (95% HPD: 5256 - 6800 YBP) according to our Tibet-Arunachal phylogeny. Our Tibet-Arunachal phylogeny has a bipartite structure with one clade comprising languages spoken in present-day Arunachal Pradesh and another clade comprising Bodish, Tshangla, West Himalayish and the Lolo-Burmese languages. We hesitate to give a definite answer about the origin of the common ancestor of the selected groups in our study due to two reasons. First, in addition to the written records about the history of the Tibetan plateau in written Tibetan and Chinese sources and the archaeological evidence that has been unearthed from the Tibetan plateau and its eastern fringes (see, for example, Aldenderfer 2011), there have been only few stratified excavations from Arunachal (see, for example, Tada et al. 2012; Ashraf 1990). Second, language diversification, dispersal or expansion can be associated with human migration, with cultural contact, or with both. The actual scenario of Trans-Himalayan language differentiation in Tibet-Arunachal area is more complex than a mere phylogeny can explain.

According to our phylogeny, we observed a lot of language differentiation events occurring between 1000 - 3000 YBP. Jeong et al. (2016) investigated the genetic structures of human remains in three archaeological sites in Northern Nepal spanning the period between 1250 - 3150 YBP, identifying a strong affinity between contemporary East Asian populations and the ancient DNA, which suggested a Southwestward expansion from the Tibetan plateau into present day Nepal. However, Jeong et al.'s (2016) study does not provide any explanation about language differentiation in Arunachal Pradesh. Perhaps, the differentiation events that our study highlights in the same period as the study by Jeong et al. (2016) indicates that expansion of East Asian DNA material and the Trans-Himalayan languages was not limited to Nepal, but occurred across the southern Himalayan region.

However, due to the absence of evidence from paleolinguistic studies about the prehistory of the people of Arunachal Pradesh, we cannot give any hypotheses regarding the origin of the Kho-Bwa, Hrusish, Tani and Mishmic clades. If there had been reliable paleolinguistic studies reconstructing vocabulary in the proto-languages relating to, for example, agricultural crops, flora and fauna and the climate and weather, we could further extend our inference. At this moment, we can only give a rough estimation of the root at 6149 YBP (95% HPD: 5256 - 6800 YBP) and the emergence of Trans-Himalayan languages spoken in Arunachal Pradesh occurring at 5625 YBP (95% HPD: 4204 - 6533), but we can not infer much with regards to where these clades originate from - whether they were native to the area itself, came from the Tibetan plateau, or originate elsewhere. However, we are able to make some statements about the internal structure of the lower level clades. In the following sections, we detail the internal structure of Kho-Bwa and Hrusish, and then we attempt to interpret our phylogeny according to the evidence available to us.

5.2 Interpretation about the internal structure of Kho-Bwa and Hrusish

Our phylogeny supports the hypothesis that Kho-Bwa and Hrusish are members of the Trans-Himalayan language family, placing proto-Kho-Bwa and proto-Hrusish at around 2843 YBP and 1846 YBP, respectively. Kho-Bwa and Hrusish are shown to be genealogically closer to the Tani and Mishmic language subgroups than to the Bodish, rGyalrong and Lolo-Burmese languages.

The internal structure of the Kho-Bwa languages agrees with current linguistic studies that the Western Kho-Bwa group branched off first, followed by the separation between Bugun and Puroik. Furthermore, our phylogenies agree with earlier linguistic studies by showing that the ancestor of Khispi-Duhumbi was the first to split among the Western Kho-Bwa languages, that the Sherdukpen varieties (Rupa and Shergaon) are closely related, and that Khoina and Jerigaon are likewise closely related. Our tree also shows that the positions of Rahung and Khoitam are not settled, with the Sartang variety Rahung occupying an intermediate position between the Sherdukpen varieties (Rupa and Shergaon) and the other Sartang varieties (Khoitam, Jerigaon, and Khoina).

In addition, we observe in the neighbor-net network (Supplement S6), that all the Kho-Bwa subgroups are very well supported as clades. Even the single Western Puroik variety (KB West Puroik Lieberherr), which in the phylogenetic tree of Figure 2 is the first split, has a clear signal

in common with all other Puroik varieties. While Bugun shows a comparatively recent network signal with Western Kho-Bwa, Puroik shows similar signals with both Bugun and Hrusish; both signals are compatible with a hypothesis of more recent language contact, in addition to the older genetic relation between Bugun and Western Kho-Bwa and between Puroik and Bugun, as these varieties are located geographically close to each other and their populations are known to have had socioeconomic and cultural contact. This latter observation – a comparatively recent network signal between Puroik and Hrusish – is worth mentioning, because, assuming a strict clock, this conflicting signal likely arose around the same time when the ancestor of the Western Kho-Bwa varieties split from the ancestor of Puroik and Bugun and hence lends evidence to the local origin and migration stories that relate how the advent of the Hrusish speakers led to migration and differentiation of the Kho-Bwa speakers.

The internal structure of the Hrusish languages has as noticeable feature that, although our phylogeny shows that Hruso Aka is a member of the Hrusish group, we observe that the split of Hruso Aka is much more ancient than the subsequent splits of the other Hrusish languages. This leads us to suspect that either there may be another Trans-Himalayan language subgroup that was not in our selection of languages that is related to Hruso Aka; that Hruso Aka has a non-Trans-Himalayan substrate; or that other linguistic varieties closer related to Hruso Aka than to the other Hrusish varieties went extinct in the past. This observation is consistent with earlier allusions by Blench and Post (2014) regarding the possibly distinct position and linguistic history of Hruso Aka, although at this point we have more possible explanations than simply that this is due to the non-Trans-Himalayan nature of the language. Similar situations, where a single contemporary language appears to have split from its closest genetic relatives at a considerable time depth, without having any other more recent linguistic relatives, can be observed for Gémàn, Zhaba, Lisu and Bunan, but the neighbor-net in Supplement S6 also shows that Zhaba and Lisu are very weakly connected to their respective larger subgroups. We cannot make informed comments about the latter three varieties, and it is likely that other linguistic varieties exist that split more recently from, and are hence more closely related to, these varieties, but which not included in our sample. However, like with Hrusish, we know that the three Mishmic varieties in our sample cover all the known varieties. Hence, the position of Gémàn may be attributed to the same three possible interpretations offered for Hruso Aka. This is consistent with Blench (2017), who observed the close linguistic affiliation of Yìdū and Dáràng but the clearly distinct linguistic nature of the Gémàn language.

Among the Mijic varieties, the ancestor of Bangru is the first to split, as was also indicated in Bodt and Lieberherr (2015). The overall pattern also matches the description in Abraham et al. (2018[2005]) and Bodt and Lieberherr (2015) that the Miji varieties can be divided in Western and Eastern Miji, with the two Namrei varieties making up an Eastern Miji clade, and the two Dammai varieties plus Nafra Miji making up a Western Miji clade. However, our results show that Dammai Rurang is more closely related to Nafra Miji than to Dammai Dibin. We speculate that these three varieties are mutually intelligible, which is also aligned well with Abraham et al. (2018[2005]) survey. Abraham et al.'s (2018[2005]) stated that the distinction between "Miji" and "Dammai" is a difference in nomenclature.

The split of Khispi and Duhumbi from the Sartang and Sherdukpen varieties in the Western Kho-Bwa clade, and the split of Bangru from the other Mijic varieties in the Miji clade is very close to each other in time, which, as we explained above, may lend evidence to local stories that the diversification of the Western Kho-Bwa varieties was initiated by the arrival of the Mijic speakers (e.g., Bodt and Lieberherr 2015, Lieberherr and Bodt 2017).

The neighbor-net we present in Supplement S6 indicates that Tani and Mishmi derive from a common and relatively old common ancestor, with relatively rapid diversion from the other Arunachal languages in our sample. Our phylogeny furthermore shows that after the common ancestor of Hrusish, Kho-Bwa, Tani and Mishimic emerged, the Proto-Hrusish, Proto-Kho-Bwa, Proto-Tani and Proto-Mishimic languages all existed for a long time without diversification, and that the subsequent differentiation of languages in Arunachal Pradesh is relatively recent. At first, we suspected that this may have been because there were languages in the same subgroups that were not sampled. However, in the case of Tani, we know that most other Tani varieties except perhaps Apatani and Milang (Modi and Post 2009; Macario 2015) are remarkably similar to the Tani varieties we already included in our dataset. In the case of the Hrusish, the Kho-Bwa and the Mishmic group, our sample covers basically all or the vast majority of known varieties.

We have four different interpretations, namely that (a) the languages related to the selected subgroups at a higher level were not included or insufficiently represented in our complete sample, and hence did not show up in our Trans-Himalayan phylogeny as being related to these language subgroups; (b) some older languages in this clade went extinct without being documented, like we observe for Tangut or Old Tibetan in other clades; (c) the languages in Arunachal Pradesh were isolated for a long time, until, in more recent times, multiple waves of migration triggered language differentiation; and (d) the cognate decisions obscured extant

relations between the Tani or the Hrusish languages and other languages in our entire sample. One possibility for discovering more about the value of the first interpretation would be to progressively expand our experiment by including more and more languages and linguistic subgroups of the Trans-Himalayan language family, with experts on these additional languages making the cognate decisions. About the second interpretation, the written records of languages that were once spoken in the Tibet-Arunachal area are basically absent except for Tibetan, therefore, we can not add other old languages that would attest to historical splits. As for the third interpretation, the mountains or high altitude areas are often seen as natural barriers that prevent people from moving between different locations freely. However, Huber and Blackburn (2012: 102) give evidence that although there was indeed migration from the southern fringes of the Tibetan Plateau to the neighboring highland regions of Arunachal Pradesh, such moves could have been part of longer cycles of shifting back and forth between higher and lower sites in response to a range of changing economic, political and ecological conditions. Due to a lack of historical and other evidence, we can not at this moment be sure whether this third interpretation applies to the speakers of Kho-Bwa, Tani, Mishmic and Hrusish languages in Arunachal Pradesh. Hence, at this moment, we prefer to take a cautious approach. More light may be shed through historical-comparative linguistic and phylolinguistic studies on other Trans-Himalayan subgroups, on clearly distinct languages such as Milang and Koro Aka in Arunachal and Gongduk, Ole Monkha and Lhokpu in Bhutan, but also on larger languages that have hitherto evaded classification such as Tshangla, Lepcha, Chepang and Karbi. Inclusion of such languages and linguistic subgroups and conscientious cognate decisions may likely reveal additional links to the languages in our sample.

5.3 The undetermined position of Tshangla

Our classification does not support any of the earlier hypotheses that consider the Tshangla varieties to be members of the Bodish language clade. While the consensus tree points to, as already mentioned, a scenario with the ancestor of Tshangla as the first split in a group also comprising the ancestors of Bodish, Lolo-Burmese, and rGyalrong languages, it is only the most likely hypothesis among others that must also be investigated. In a second scenario, Tshangla could be grouped with the languages spoken in Arunachal Pradesh. Indeed, the density tree presented in supplement S5 indicates a possibly closer connection between the Tani languages

and all the Tshangla varieties, and not just Mòtuō Tshangla which has Tani languages as known contact languages. In a third scenario, Tshangla could be grouped with Lolo-Burmese languages. In a fourth scenario, derived from the density tree presented in supplement S5, the possibility of a non-Trans-Himalayan substrate is indicated by a line that exceeds the root of our tree. And last but not least, the Neighbor-net in Supplement S6 shows that Tshangla is almost a paraphyletic group, and that the low posterior support that we observe in the phylogenetic tree is due to conflicting signals with Rongpo, Byangsi and Bunan, i.e., the 'West Himalayish' group. This last scenario could be compatible with the idea that at least part of the Tshangla lexicon is derived from the ancient but extinct language of Zhangzhung, which is also thought to be related to the West Himalayish languages (cf., e.g. Matisoff 2001 and Widmer 2014: 53-56).

5.4 Limitations

5.4.1 Issues related to word compounding

Many Trans-Himalayan languages are marked by polymorphemic forms expressing a single concept. Often, these are lexical compounds. Usually, there are no clear monomorphemic "roots" expressing a concept in all varieties that would be straightforward to compare. This is not a feature unique to Trans-Himalayan languages, as compounding is a prevailing word formation mechanism found in several language families, such as among the Hmong-Mien (Ratliff 2010) and Bantu (Currie et al. 2013) languages. Although linguists are well aware of this phenomenon of "partial cognacy" (List 2016), the customary approach to cognate judgments is still to judge the cognacy on the lexical level (i.e. the entire word). To compromise the shortcoming of this classical methodology, it is common to consider only one morpheme per gloss in cognate judgments, regardless of whether compounding took place or not (Ratliff 2010). This approach has some caveats: the word forms are not well preserved, potentially causing confusion; linguists may not agree with the morphemes which are selected to represent the words (the "salient" ones); and, the most serious issue of all, a dataset tends to lose comparability with other sources after such "data compression".

To avoid the aforementioned issues, Sagart et al. (2019) used the "common morpheme" approach and the advantage of *LingPy* wordlist format (List et al. 2019), which is to collect words, including synonyms, to make sure that at least one common morpheme is shared among languages. The common morpheme approach solved the issue that may be caused by the "one

morpheme per gloss" approach, and the *LingPy* wordlist format provides columns to preserve the full word forms. Although commendable, this solution does not totally facilitate the work of extending a database or combining multiple sources. Our attempt to use the cognate judgments made by Sagart et al. (2019) as a baseline was quite challenging. In many of their earlier cognate judgments, it was unclear which morpheme in a particular variety was compared and judged cognate with which other morpheme in the other varieties. This also implies that, in several cases, forms that were at least partially cognate with forms in other varieties were not marked as such.

In order to make our cognate decisions transparent and reproducible, we used *EDICTOR* version 2.6.6 to annotate our cognate judgments. Its functions allow for the cognate terms to be displayed together, showing all forms that belong to a given cognate ID to a much closer scrutiny of which morpheme is judged as cognate. Later, *EDICTOR* implemented the option to annotate morphemes with semantic and grammatical features (List 2021). Unfortunately, an equivalent functionality was probably not available to Sagart et al. (2019), which would have made their cognate judgments, and in turn our own, much more insightful. However, even if these options were to be fully explored, this would not enable judging morphemes as cognates beyond the concepts that are actually the object of this study: a given word or morpheme may have cognate forms that, through semantic change, have shifted to a different meaning, and hence are not reflected in the dataset unless the new meaning happens to be included in the concept list. We discuss this issue in the following subsection.

5.4.2 Issues related to cross-semantic cognates

As in almost all phylolinguistic studies, Sagart et al. (2019) only performed cognate judgment among words for the same cross-linguistic concept. Although this practice is justifiable in several aspects, from scope restriction to adequacy with what the quantitative models expect, it overlooks the phenomenon of semantic change, where a form in one variety is cognate with a form in another variety but with a different, albeit usually related, meaning. A common theoretical reading in phylogenetic contexts holds that semantic change is itself an event of lexical substitution, so that this decision has little influence on the results and might even be desirable since the semantic change is transmitted to descendants as well as other lexical substitutions. However, semantic changes tend to be gradual and rarely "shift" in meaning, with a progressive extension or reduction of the semantic field involved being more common.

Considering our dataset, Dzala *'me.loŋ* means "eye", whereas the Wénlăng Cūonà Ménbā form for "eye" is *mek*^{55}, while the cognate is actually the Wénlăng Cūonà Ménbā form *me*^{55}*.loŋ*^{55} "eyebrow". There was a semantic change between "eyebrow" in Wénlăng Cūonà Ménbā which became "eye" in Dzala. Unless both the concepts "eyebrow" and "eye" are present in the concept list used for the comparison, we may not denote these forms as cognate. Similarly, Dirang Tshangla *a.ta* means "grandfather", whereas Bhutan Tshangla *a.ta* means "elder brother": these two forms are cognate, but subject to semantic changes. However, in a dataset for a phylogenetic study, even when both the concepts "elder brother" and "grandfather" are present, we would not denote these two forms as cognate unless we specifically annotate cross-semantic cognates.

Semantic change can occur at both the entire word level and the morphemic level, which, as seen, is particularly relevant for the family under study. More than that, semantic change is not just prevalent in, but even inherent to situations where we compare languages. Even within relatively recent and low-level sub-groups, such as the Western Kho-Bwa languages, we can find plenty of examples of semantic change. The failure to recognize such semantic change will only become more relevant the higher we ascend in the phylogenetic tree. Not only are we comparing across a wide range of time periods, going back some millennia, we are also comparing across a wide range of highly divergent cultural complexes, from what are basically hunter-gatherers like the Puroik to complex, highly evolved and stratified societies like the Old Chinese and modern Sinitic cultures, and we are comparing across a wide range of highly diverse habitats, from tropical jungles in river valleys to the highest plateau on Earth, to deserts, and to coastal cities and towns. The linguistic evidence indicates that some Trans-Himalayan languages display various degrees of creolization due to language contact (DeLancey 2013). DeLancey states that Tshangla, for example, is an extreme case of a creoloid language, and that the Bodish languages also show a significant degree of creolization. These creoloid traits may skew the relatedness among language groups if they are not considered carefully (van Driem 2021: 108). This combination of time depth, varying developmental patterns, livelihood systems, and environmental habitats means that the same inherited word form may have obtained significantly different meanings in the related descendant languages.

Koptjevskaja-Tamm (2008) introduced the idea of colexifications which addresses the semantic shift phenomenon in the process of synchronic and diachronic language change. A large-scale colexification database was developed by Rzymski et al. (2020) to improve the customary practice in quantitative historical linguistic studies. A function to automatically detect

the colexification among words, which is also known as cross-semantic cognates, was implemented in the computer-assisted workflow described by Wu et al. (2020). Since 2020, *EDICTOR* has also been provided with the interface to inspect the cross-semantic cognates. Unfortunately, once more the Sagart et al. (2019) study did not have such an option available at the time.

5.4.3 Issues about sampling bias

The Trans-Himalayan language family consists of more than 400 highly diversified languages. We added a substantial number of languages that are spoken in Arunachal Pradesh as well as in the adjacent area of the Tibetan plateau and Bhutan, but we recognize the language subgroups in the dataset that are not fully or equally sampled. The phylogenetic models we use are, at least in theory, partially resistant to this problem, which we also took into account when stipulating the parameters of our evolutionary model.

5.4.4 Insufficient archaeological and population genetic evidence

Meyer et al. (2009) stated that the semi-nomadic populations started yak herding on the Tibetan plateau around 6700 YBP, however, we cannot confidently link their research result to ours without having more evidence on the ancient human genome as well as modern population genetics studies on this matter[14].

Deriving a solid time frame about the peopling of Arunachal Pradesh from the existing studies to integrate with our model was not feasible. Although there have been sporadic archaeological findings, such as stone tools, we cannot establish an immediate connection between the material cultures that produced them and the modern populations in this area (Ashraf 1990). Likewise, the modern Trans-Himalayan speakers in Arunachal Pradesh show genetic admixture with populations from Southern China, Southeast Asia, and India. Therefore, we could not assign

[14] Jacques et al. (2021) summarized the archaeological and linguistic evidence and estimated a much later dates than 6700 YBP of the yak domestication. According to their linguistic evidence, yak domestication happened two times on the Tibetan Plateau. The first time occurred among the speakers of the linguistic ancestor of Tibetan and the second time occurred among speakers of Proto-rGyalrong. The dates that are given in Jacques et al. (2021) could benefit future Bayesian phylolinguistic study related to Bodish languages.

calibration dates to the internal nodes that are related to all the selected languages spoken in Arunachal Pradesh nowadays.

Although we cannot provide much information to the algorithm via the existing archaeological or population genetic studies, the evidence provided by these disciplines demonstrates that complex admixtures occurred in the past, shaping today's languages and population in the area. Our study highlights the challenges and hopes that these obstacles raise the attention from the other disciplines.

6 Conclusion

Admittedly, any attempt by Bayesian phylolinguistics to describe language evolution with statistical models greatly simplifies reality. Nonetheless, it enables us to examine hypotheses and provide statistical evidence, particularly when the knowledge about the history of the languages involved is still limited, such as in the case of Arunachal Pradesh. By being aware of these limits, which were set out above, we were able to use this method fruitfully, especially in evidencing the internal structure and time periods of diversification of the two comparatively understudied Kho-Bwa and Hrusish groups.

Our phylogeny reported a date that the common ancestor of the selected language subgroups emerged around 6149 YBP and the language diversification in Arunachal Pradesh started around 5624 YBP. At a linguistic sub-grouping level, our results agree with the previous linguistic studies that the Kho-Bwa, Hrusish, and Tshangla language groups are clades of the Trans-Himalayan language family, also resolving their internal structures. We found support for the hypothesis that the individual, contemporary Mijic, Puroik, Western Kho-Bwa and Bugun varieties emerged only in the last couple of hundred years, although the higher Hrusish and Kho-Bwa clades emerged much earlier with a common ancestor around 4092 YBP and internal differentiation starting around 1846 YBP for the former and 2843 YBP for the latter. For Kho-Bwa, we found support for a Western and a "core" group, with the former starting to divide more recently, around 1100 years ago into a group composed of Khispi and Duhumbi and one involving the other Western Kho-Bwa languages. The "core" Kho-Bwa group shows a more complex structure, with the Puroik and the Bugun languages starting to differentiate around 2036 years ago. We inferred that either the continuous internal and external language contacts slowed down the language differentiation, potentially leading the algorithm to report younger split dates,

or that some longer branches might be explained by the survival or dominance of a single variety of ancient clades, with a comparatively much more recent and at times on-going diversification.

Our findings for Central Bodish and Lolo-Burmese mirror, as expected, the results in Sagart et al. (2019), and East Bodish neatly divides around 2100 YPB into two related subgroups composed by Khengkha and Bumthang on one side and Dzalakha along with the Cuona Memba varieties in the other. We found that Tani and Mishmic are distinct but related groups, sharing a common ancestor around 4806 YBP.

Unfortunately, we cannot provide reliable answers to some of the questions regarding the linguistic history of this area. For example, the relationship between the Tshangla varieties and the Bodish subgroups, or other language subgroups in Arunachal Pradesh for that matter, is not entirely clear (although an alignment with Kho-Bwa and Mishmic is less supported). Our models show that the Tshangla varieties have complex admixture from other language subgroups on the individual language level. This observation shows that using only one consensus tree to represent the diversification process may overly simplify the complexity of language evolution in an area which has long-term language contacts. Therefore, we encourage linguists who seek to use Bayesian phylogenetic methods in groups with equivalent contact histories to investigate the entire set of trees in the sample, and not just a consensus tree that might obscure support for seemingly less likely hypotheses.

Although we have expanded the sampling of languages and groups in this area compared to other studies, there are still language subgroups in Northeast India that are understudied. We believe that the way forward to further explaining the phylogenetics of the Trans-Himalayan language family and its linguistic evolution is to follow a bottom-up approach. Here, we envisage experts on linguistic subgroups to use a base dataset to add new linguistic varieties of their expertise, select the concepts, make the cognate judgments, and then probe and run models that will give initial ideas and clues about the position of the linguistic varieties and the linguistic subgroups they added within the language family. These results can then be compared to the results of the traditional method of comparative linguistics, including sound correspondences and consideration for other forms of linguistic evolution (such as reticular relationships) and contact between different families, and in this way, insights into the phylogenetics of the language family can advance.

We hope that our approach and our workflow can give an impetus to other linguists to apply the methodology to find out more about the internal structure and external relationships of under-

studied subgroups. And finally, we hope to draw the attention of archaeologists and population geneticists to the Tibetan plateau and in particular to Arunachal Pradesh, promoting a bottom-up and cross-disciplinary approach to reconstruct the topology of the Trans-Himalayan language family and improving the estimation of dates.

7 Acknowledgment

This research work was supported by ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (abbrv. CALC, https://calc.digling.org, MSW and TT), British Academy Postdoctoral Fellowship PF20_100076 "Substrate language influence in the southern Himalayas" (TAB) hosted by SOAS University of London, United Kingdom, and Riksbankens Jubileumsfond MXM19-1087:1 "Cultural evolution of texts" (TT). We thank Dr. Denise Kühnert and Mr. Konstantin Hoffmann who provided comments and expertise that assisted the Bayesian phylogenetic analysis. We thank our reviewers and Dr. Johann-Mattis List for comments that largely improved the manuscript.

References

Abraham, Binny, Kara Sako, Elina Kinny & Isapdaile Zeliang. 2018[2005]. *Sociolinguistic research among selected groups in Western Arunachal Pradesh: Highlighting Monpa*. Dallas: SIL International.

Abraham, Binny, Kara Sako, Elina Kinny & Isapdaile Zeliang. 2019. CLDF dataset derived from Abraham et al.'s "Sociolinguistic research on Monpa" from 2018[2005]. doi: 10.5281/ zenodo.3537601. 10.5281

Aldenderfer, Mark. 2011. Peopling the Tibetan Plateau: Insights from archaeology. *High Altitude Medicine & Biology 12*(2), 141–147. doi: 10.1089/ham.2010.1094.

Anderson, Gregory. D.S. 2014. On the classification of the Hruso (Aka) language. Paper presented at the 20th Himalayan Languages Symposium.

Ashraf, A. A. 1990. *Prehistoric Arunachal: A report on archaeological exploration and excavation at Kamla Valley with reference to Parsi Parlo of Lower Subansiri District, Arunachal Pradesh*. Itanagar: Directorate of Research, Govt. of Arunachal Pradesh.

Benedict, Paul K. 1972. *Sino-Tibetan: A conspectus*. Cambridge: Cambridge University Press.

Blench, Roger. 2017. The 'Mishmi' languages, Idu, Tawra and Kman: a mismatch between cultural and linguistic relations. Draft circulated for International Consortium for Eastern Himalayan Ethnolinguistic Prehistory 2017. http://www.rogerblench.info/Language/NEI/Mishmi/MisOP/Blench%20ICEHEP%20Melbourne%202017%20Text.pdf

Blench, Roger & Mark W. Post. 2014. Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages. In Thomas Owen-Smith and Nathan Hill (Eds.), *Trans-Himalayan Linguistics*, pp. 71–104. Berlin, Boston: De Gruyter. doi: 10.1515/9783110310832.71.

Bodt, Timotheus A. 2012. *The new lamp clarifying the history, people, languages and traditions of Eastern Bhutan and Eastern Mon* (2 ed.). Wageningen: Monpasang Publications.

Bodt, Timotheus A. 2014a. Ethnolinguistic survey of westernmost Arunachal Pradesh: A fieldworker's impressions. *Linguistics of the Tibeto-Burman Area 37*(2), 198–239. doi: 10.1075/ltba.37.2.03bod.

Bodt, Timotheus A. 2014b. Notes on the settlement of the Gongri river valley of Western Arunachal Pradesh. In Anna Balikci Denjongpa and Jenny Bentley (Eds.), *The dragon and the hidden land: social and historical studies on Sikkim and Bhutan. Proceedings of the Bhutan-Sikkim panel at the 13th Seminar of the International Association for Tibetan Studies, Ulaanbaatar, Mongolia, July 21-27, 2013.*, pp. 153–190. Gangtok, Sikkim: Namgyal Institute of Tibetology.

Bodt, Timotheus A. 2017. Dikhyang Bugun language data: Overview file. doi: 10.5281/zenodo.1116313.

Bodt, Timotheus A. 2019. The Duhumbi perspective on Proto-Western Kho-Bwa rhymes. *Die Sprache 52*(2), 141–176.

Bodt, Timotheus A. 2020. *Grammar of Duhumbi (Chugpa)*. Leiden, Boston:Brill.

Bodt, Timotheus A. 2021. The Duhumbi perspective on Proto-Western Kho-Bwa onsets. *Journal of Historical Linguistics 11*(1), 1–59. doi: 10.1075/jhl.19021.bod.

Bodt, Timotheus A. & Ismael Lieberherr. 2015. First notes on the phonology and classification of the Bangru language of India. *Linguistics of the Tibeto-Burman Area 38*(1), 66–123.

Bodt, Timotheus A. & Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology 4*(1), 22–44.

Bouckaert, Remco R., Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler & Alexei J. Drummond. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology 15*(4), e1006650. doi: 10.1371/journal.pcbi.1006650.

Bouckaert, Remco. R. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics 26*(10), 1372–1373. doi: 10.1093/bioinformatics/btq110.

Bradley, David. 1997. Tibeto-Burman languages and classification. In David Bradley (Ed.), *Papers in Southeast Asian Linguistics 14: Tibeto-Burman Languages of the Himalayas*, Volume 86 of *Pacific Linguistics, Series A*, pp. 1–72. Canberra: Australian National University.

Bradley, David. 2002. The subgrouping of Tibeto-Burman. In Christopher Beckwith (Ed.), *Medieval Tibeto-Burman languages*, International Association for Tibetan Studies Proceedings 9 and Brill Tibetan Studies Library 2, pp. 73–112. Leiden: Brill.

Brass, Tom. 2012. Scott's "Zomia", or a populist post-modern history of nowhere. *Journal of Contemporary Asia 42*(1), 123–133. doi: 10.1080/00472336. 2012.634646.

Bryant, David & Vincent Moulton. 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution 21*(2), 255–265. doi: 10.1093/molbev/msh018.

Burling, Robbins. 2003. The Tibeto-Burman languages of Northeastern India. In Graham Thurgood and Randy J. LaPolla (Eds.), *The Sino-Tibetan languages* (1 ed.)., Routledge Language Family Series, pp. 169–192. London, New York: Routledge.

Burling, Robbins & James A. Matisoff. 1980. Variational semantics in Tibeto-Burman: The 'organic' approach to linguistic comparison. *Language 56*(4), 888. doi: 10.2307/413505.

Currie, Thomas E., Andrew Meade, Myrtlle Guillon & Ruth Mace. 2013. Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. *Proceedings of the Royal Society B: Biological Sciences 280*(1762), 20130695. doi: 10.1098/rspb.2013.0695.

DeLancey, Scott. 2013. Creolization in the divergence of the Tibeto-Burman languages. In Thomas Owen-Smith and Nathan Hill (Eds.), *Trans-Himalayan Linguistics: Historical and Descriptive Linguistics of the Himalayan Area*, pp. 41–70. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110310832.41.

Driem, George van. 2001. *Languages of the Himalayas: An ethnolinguistic handbook of the greater Himalayan Region*. Handbook of oriental studies. Section two, India, Handbuch der Orientalistik. Indien. Leiden: Brill.

Driem, George van. 2007. The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese. *Bulletin of Chinese Linguistics 1*(2), 211–270.

Driem, George van. 2014. Trans-Himalayan. In Thomas Owen-Smith and Nathan Hill (Eds.), *Trans-Himalayan linguistics*, pp. 11–40. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110310832.11.

Driem, George van. 2015. Synoptic grammar of the Bumthang language. *Himalayan Linguistics*.

Driem, George van. (2021). *Ethnolinguistic prehistory: The peopling of the world from the perspective of language, genes and material culture*. Number volume 26 in Brill's Tibetan studies library. Languages of the Greater Himalayan region. Leiden: Brill.

Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips & Andrew Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLOS Biology 4*(5), e88. doi: 10.1371/journal.pbio.0040088.

Drummond, Alexei J. & Andrew Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology 7*(1), 214. doi: 10.1186/1471-2148-7-214.

Drummond, Alexei J., Marc A. Suchard, Dong Xie & Andrew Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution 29*(8), 1969–1973. doi: 10.1093/molbev/mss075.

Dusu, Sambyo. 2013. *Akas of Arunachal Pradesh A historical study till 1947 AD*. PhD thesis, Department of History, Rajiv Gandhi University, Itanagar, India, Doimukh.

Dzongkha Development Commission. 2017. *Dzongkha-English-dZalakha lexicon*. Thimphu: Dzongkha Development Commission.

Dzongkha Development Commission. 2018. *Bumthangkha-Dzongkha-English lexicon*. Thimphu: Dzongkha Development Commission.

Eberhard, David, Gary Simons & Chuck Fennig. 2019. *Ethnologue: Languages of the world*. Dallas: SIL International.

Forkel, Robert, Simon Greenhill & Hans-Jörg Bibiko. 2021. Pylexibank. the python curation library for lexibank [software library, version 2.8.2]. https://github.com/lexibank/pylexibank.

Forkel, Robert & Johann-Mattis List. 2020. CLDFBench: Give your cross-linguistic data a lift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6995– 7002. Marseille: European Language Resources Association.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russel D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data 5*(1), 180205. doi: 10.1038/sdata.2018.205.

Fürer-Haimendorf, Christophn von. 1982. Tribes of India: the struggle for survival. Berkeley: University of California Press.

Gamerman, Dani & Hedibert Freitas Lopes. 2006. *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed.). Number 68 in Texts in statistical science series. Boca Raton: Taylor & Francis.

Gavryushkina, Alexandra, David Welch, Tanja Stadler & Alexei J. Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLOS Computational Biology 10*(12), e1003919. doi: 10.1371/journal.pcbi.1003919.

Genetti, Carol. 2016. *The Tibeto-Burman languages of South Asia: The languages, histories, and genetic classification*, pp. 130–155. Berlin, Boston: De Gruyter Mouton.

Gerber, Pascal & Selin Grollmann. 2018. What is Kiranti? A critical account. *Bulletin of Chinese Linguistics 11*(1-2), 99–152. doi: 10.1163/2405478X-01101010.

Gilks, W. R., S. Richardson & D. J. Spiegelhalter (Eds.). 1996. *Markov chain Monte Carlo in practice*. Boca Raton: Chapman & Hall.

Good, Jeff & Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language documentation & conservation 7*.

Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science 323*(5913), 479–483. doi: 10.1126/science.1166858.

Greenhill, Simon J., Paul Heggarty & Russell D. Gray. 2020. *Bayesian phylolinguistics*, Chapter 11, pp. 226–253. New Jersey: John Wiley Sons, Ltd. doi: 10.1002/9781118732168.ch11.

Grewal, Dalvindar Singh. 1992. *The Aka Miji and their kindred in Arunachal Pradesh: An enquiry into determinants of their identity*. PhD thesis, University of North Bengal, Siliguri, India.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2021. *Glottolog 4.4*. Leipzig. doi: 10.5281/zenodo.4761960.

Hazarika, Manjil. 2016. Tracing post-Pleistocene human movements and cultural connections of the eastern Himalayan region with the Tibetan plateau. *Archaeological Research in Asia 5*, 44–53. doi: 10.1016/j.ara.2016.03.003.

Hazarika, Manjil. 2017. *Prehistory and archaeology of Northeast India: multidisciplinary investigation in an archaeological Terra incognita*. New Delhi: Oxford University Press.

Huber, Toni. 2020. *Source of life: Revitalisation rites and bon shamans in Bhutan and the Eastern Himalayas*. Vienna: Austrian Academy of Sciences Press.

Huber, Toni & Stuart Blackburn. 2012. *Origins and Migrations in the Extended Eastern Himalayas*. Leiden: Brill. doi: 10.1163/9789004228368.

Huelsenbeck, John P. 2002. Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution 19*(5), 698–707. doi: 10.1093/oxfordjournals.molbev.a004128.

Huson, Daniel H. & David Bryant. 2005. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution 23*(2), 254–267.

Huáng, Bùfán and Qìngxià Dài (Eds.). 1992. *Zàngmiǎn yǔzú yǔyán cíhuì*《藏缅语族语言词汇》*[A Tibeto-Burman Lexicon]*. Běijīng: Zhōngyāng Mínzú Dàxué [中央民族大学].

Hyslop, Gwendolyn. 2013. On the internal phylogeny of East Bodish. In Gwendolyn Hyslop, Mark W. Post, and Stephen Morey (Eds.), *North East Indian Linguistics*, Volume 5, pp. 91–110. India: Foundation Books. doi: 10.1017/9789382993285.005.

Hyslop, Gwendolyn. 2014. A preliminary reconstruction of East Bodish. In Nathan Hill and Thomas Owen-Smith (Eds.), *Trans-Himalayan Linguistics*, pp. 155–179. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110310832.155.

Hyslop, Gwendolyn & Jade d'Alpoim Guedes. 2020. Linguistic evidence supports a long antiquity of cultivation of barley and buckwheat over that of millet and rice in Eastern Bhutan. *Vegetation History and Archaeobotany 30*(4), 571–579. doi: 10.1007/s00334-020-00809-8.

Jacques, Guillaume, Jade d'Alpoim Guedes & Shuya Zhang. 2021. Yak domestication: A review of linguistic, archaeological, and genetic evidence. *Ethnobiology Letters 12*(1), 103—114. doi: 10.14237/ebl.12.1.2021.1755.

Jacquesson, François. 2015. *An introduction to Sherdukpen language*, Volume 39 of *Diversitas linguarum*. Bochum: Universitätsverlag Dr. N. Brockmeyer.

Jeong, Choongwon, Andrew T. Ozga, David B. Witonsky, Helena Malmström, Hanna Edlund, Courtney A. Hofman, Richard W. Hagan, Mattias Jakobsson, Cecil M. Lewis, Mark S. Aldenderfer, Anna Di Rienzo & Christina Warinner. 2016. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proceedings of the National Academy of Sciences of the United States of America 113*(27), 7485–7490. doi: 10.1073/pnas.1520844113.

Koptjevskaja-Tamm, Maria. 2008. Approaching lexical typology. In Martine Vanhove (Ed.), *From Polysemy to Semantic Change*, Number 106 in Studies in Language Companion Series, pp. 3–52. Amsterdam: John Benjamins Publishing Company.

Krithika, S. & T. S. Vasulu 2018. *Folklore versus genetics: A mitochondrial DNA investigation about the origin and antiquity of the Adi sub-tribes of Arunachal Pradesh, India*, pp. 161–185. Singapore: Springer. doi: 10.1007/978-981-13-1843-6_11.

Leyden, John. 1808. On the languages and literature of the Indo-Chinese nations. *London: Asiatic Researches 10*, 158–289.

Lieberherr, Ismael. 2015. A progress report on the historical phonology and affiliation of Puroik. In Linda Konnerth, Stephen Morey, Prizankoo Sarmah, and Amos Teo (Eds.), *North East Indian Linguistics (NEIL) 7*, pp. 235–286. Canberra: Asia-Pacific Linguistics Open Access.

Lieberherr, Ismael. 2017. *A grammar of Bulu Puroik*. PhD thesis, Universität Bern, Bern, Switzerland.

Lieberherr, Ismael & Timotheus A. Bodt. 2017. Sub-grouping Kho-Bwa based on shared core vocabulary. *Himalayan Linguistics 16*(2). doi: 10.5070/H916232254.

Lieberman, Victor. 2010. A zone of refuge in Southeast Asia? Reconceptualizing interior spaces. *Journal of Global History 5*(2), 333–346. doi: 10.1017/S1740022810000112.

List, Johann-Mattis. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution 1*(2), 119– 136. doi: 10.1093/jole/lzw006.

List, Johann-Mattis. 2021. Edictor. a web-based interactive tool for creating and editing etymological datasets. https://digling.org/edictor/.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2021. Lexibank: A public repository of standardized wordlists with computed phonological and lexical features. doi: 10.21203/rs.3.rs-870835/v1.

List, Johann-Mattis, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2019. LingPy. A Python library for quantitative tasks in historical linguistics. doi: 10.5281/zenodo.3554103.

List, Johann-Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt & Robert Forkel (Eds.). 2021. *Concepticon 2.5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution 3*(2), 130–144. doi: 10.1093/jole/lzy006.

Lù, Shàozhūn. 1986. *cuò nà mén bā yǔ jiǎn zhì*《错那门巴语简志》*[A sketch grammar of Cuona Menba]*. Běijīng: Mínzú chūbǎn shè [民族出版社].

Lù, Shàozhūn. 2002. *Ménbāyǔ fāngyán yánjìu*《门巴语方言研究》*[A study of Menba]*. Běijīng: Mínzú chūbǎn shè [民族出版社].

Lǐ, Dàqín. 2004. *Sūlóngyǔ yánjīu*《苏龙语研究》*[A study of Sulong]* (1 ed.). Běijīng: Mínzú chūbǎn shè [民族出版社].

Macario, Florens Jean-Jacques. 2015. The genetic position of Apatani within Tibeto-Burman. In *North East Indian Linguistics 7*, Volume 7, pp. 213. Canberra: Asia-Pacific Linguistics.

Matisoff, James A. 2001. The interest of Zhangzhung for comparative Tibeto-Burman. *New Research on Zhangzhung and Related Himalayan Languages (Bon Studies 3). Senri Ethnological Studies* (19), 155–180.

Matisoff, James A. 2009. Stable roots in Sino-Tibetan/Tibeto-Burman. In Y. Nagano and K. M. Hakubutsukan (Eds.), *Issues in Tibeto-Burman historical linguistics*, Number 75 in Senri Ethnological Studies, pp. 291–318. Osaka: National Museum of Ethnology.

Matisoff, James A. 2015. *The Sino-Tibetan Etymological Dictionary and Thesaurus project (STEDT)*. California: University of California.

Maurits, Luke, Robert Forkel, Gereon A. Kaiping & Quentin D. Atkinson. 2017. BEASTling: A software tool for linguistic phylogenetics using BEAST 2. *PLOS ONE 12*(8), e0180908.

Meyer, M., Ch.-Ch. Hofmann, A.M.D. Gemmell, E. Haslinger, H. Häusler, and D. Wangda. 2009. Holocene glacier fluctuations and migration of Neolithic yak pastoralists into the high valleys

of Northwest Bhutan. *Quaternary Science Reviews 28*(13-14), 1217–1237. doi: 10.1016/j.quascirev.2008.12.025.

Michailovsky, Boyd & Martine Mazaudon. 1994. Preliminary notes on the languages of the Bumthang group. In Per Kvaerne (Ed.), *Tibetan Studies, Proceedings of the 6th seminar of the International Association for Tibetan Studies, Fagernes 1992*, pp. 545–557. Oslo: The institute for comparative research in human culture.

Michaud, Jean. 2010. Editorial –Zomia and beyond. *Journal of Global History 5*(2), 187–214. doi: 10.1017/S1740022810000057.

Michaud, Jean. 2018. Zomia and beyond. In A. Horstmann, M. Saxer, and A. Rippa (Eds.), *Routledge Handbook of Asian Borderlands*, pp. 73–88. London: Routledge.

Modi, Yankee & Mark W. Post. 2009. The sociolinguistic context and genetic position of Holon (Milang) in Tibeto-Burman. International Conference on Sino-Tibetan Languages and Linguistics 42, Chiang Mai.

Nascimento, Fabrícia F., Mario dos Reis & Ziheng Yang. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution 1*(10), 1446–1454. doi: 10.1038/s41559-017-0280-x.

Opgenort, Jean Robert. 2005. *A grammar of Jero: With a historical comparative study of the Kiranti languages*. Brill's Tibetan studies library. Leiden: Brill.

Post, Mark W. 2007. *A grammar of Galo*. PhD thesis, La Trobe University, Melbourne, Australia.

Post, Mark W. 2017. *The Tangam language: Grammar, lexicon and texts*. Leiden: Brill.

Post, Mark W. & Robbins Burling. 2017. The Tibeto-Burman languages of Northeast India. *The Sino-Tibetan Languages*, 213–242.

Ratliff, Martha S. 2010. *Hmong-Mien language history*. Studies in language change. Canberra: Pacific Linguistics.

Ravenzwaaij, Don van, Pete Cassey & Scott D. Brown. 2018. A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review 25*(1), 143–154. doi: 10.3758/s13423-016-1015-8.

Remsangpuia. 2008. *Puroik phonology*. Shillong: Don Bosco Centre for Indigenous Cultures.

Rinchin, Megejee. 2011. *Stratification and change among the Sherdukpens An anthropological study on a Buddhist Tribe of Arunachal Pradesh*. PhD thesis, Rajiv Gandhi University, Itanagar, India.

Rutgers, Leopold Roland. 1999. Puroik or Sulung of Arunachal Pradesh. Paper presented at the 5th Himalayan Languages Symposium, Kathmandu.

Rzymski, Christoph, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E.Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Ar-java, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps,Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel & Johann-Mattis List. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross- linguistic polysemies. *Scientific Data 7*(13), 1–12. doi: 10.1038/s41597-019-0341-x.

Sagart, Laurent. 2011. The homeland of Sino-Tibetan-Austronesian: where and when? *Communication on Contemporary Anthropology 5*(1). doi: 10.4236/coca.2011. 51021.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill & Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences 116*(21), 10317–10322. doi: 10.1073/ pnas.1817972116.

Schendel, Willem van. 2002. Geographies of knowing, geographies of ignorance: Jumping scale in Southeast Asia. *Environment and Planning D: Society and Space 20*(6), 647–668. doi: 10.1068/d16s.

Scott, James C. 2009. *The art of not being governed: An anarchist history of upland Southeast Asia*. New Haven: Yale agrarian studies series. Yale University Press.

Shafer, Robert. 1947. Hruso. *Bulletin of the School of Oriental and African Studies 12*, 184– 196.

Shafer, Robert. 1954. The linguistic position of Dwags. *Oriens 7*(2), 348–356. doi: 10.1163/1877837254X00071.

Shafer, Robert. 1955. Classification of the Sino-Tibetan languages. *Word 11*(1), 94–111.

doi: 10.1080/00437956.1955.11659552.

Simon, Ivan M. 1979. *Miji language guide*. Shillong: Philological Section, Directorate of Research, Govt. of Arunachal Pradesh.

Simon, Ivan M. 1993 [1970]. *Aka language guide*. Shillong: Research Department, North Eastern Frontier Agency.

Soja, Rai. 2009. *English-Puroik dictionary*. Shillong: Living Word Communicators.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer & Alexei J. Drummond. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences 110*(1), 228–233. doi: 10.1073/pnas.1207965110.

Stonor, C. R. 1952. The Sulung Tribe of the Assam Himalayas. *Anthropos 47*(5/6), 947–962.

Sūn, Hongkai, Panghsin Ting, and Di Jiang. 1991. *Záng Miǎn yǔ yǔ yīn hé cí huì*《藏缅语语音和词汇》*[Tibeto-Burman phonology and lexicon]*. Běijīng: Zhōngguó Shèhuì Kēxué Chūbǎnshè [中国社会科学出版社].

Sun, Jackson T.-S. 1992. Review of Zangmicmyu Yuyin He Cihui "Tibeto-Burman Phonology and Lexicon". *Linguistics of the Tibeto-Burman Area 15*(2), 73–113.

Sun, Jackson T.-S. 1993. *A historical-comparative study of the Tani (Mirish) branch in Tibeto-Burman*. PhD thesis, Univerisity of California at Berkeley, Berkeley, USA.

Tada, Tage, J. C. Dutta & Nabajit Deori. 2012. *Archaeological heritage of Arunachal Pradesh: A book exclusively based on the findings of archaeological investigations of two decades (1991-2011)*. Itanagar: Government of Arunachal Pradesh, Department of Cultural Affairs, Directorate of Research.

Thurgood, Graham. 2003. A subgrouping of the Sino-Tibetan languages: the interaction between language contact, change, and inheritance. In Graham Thurgood and Randy J. LaPolla (Eds.), *The Sino-Tibetan languages*, pp. 3–21. London: Routledge.

Thurgood, Graham and Randy J. LaPolla (Eds.). 2003. *The Sino-Tibetan languages*. Number 3 in Routledge language family series. London: Routledge.

Waskom, Michael L. 2021. Seaborn: statistical data visualization. *Journal of Open Source Software 6*(60), 3021. doi: 10.21105/joss.03021.

Widmer, Manuel. 2014. A tentative classification of West Himalayish. In *A descriptive grammar of Bunan*, pp. 33–56. Bern: University of Bern.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data 3*(1), 1–9. doi: 10.1038/sdata.2016.18

Wu, Mei-Shin, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill & Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data 6*(1), 2. doi: 10.5334/johd.12.

Yangzom, Deki & Marten Arkesteijn. 1996. *Khengkha lessonbook*. Thimphu: SNV Bhutan.

Yu, Guangchuang. 2020. Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics 69*(1), e96. doi: 10.1002/cpbi.96.

Zhang, Hanzhi, Ting Ji, Mark Pagel & Ruth Mace 2020. Dated phylogeny suggests early Neolithic origin of Sino-Tibetan languages. *Scientific Reports* 10(1): 20792. doi:10.1038/s41598-020-77404-4.

Zhang, Menghan, Shi Yan, Wuyun Pan & Li Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in Northern China in the Late Neolithic. *Nature 569*(7754), 112–115. doi: 10.1038/s41586-019-1153-z.