Journal of Language Evolution, 2024, **XX**, 1–15
https://doi.org/10.1093/jole/lzae008
Advance access publication 30 October 2024
Research Article

OXFORD

A phylogenetic study of South-Western Tibetic

Dubi Nanda Dhakal[1], Johann-Mattis List[2,3], Seán G. Roberts[4,*]

[1]Central Department of Linguistics, Tribhuvan University, Kathmandu 44618, Nepal
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany
[3]Chair for Multilingual Computational Linguistics, Faculty of Humanities and Cultural Studies, University of Passau, Dr.-Hans-Kapfingerstr. 14d, 94032 Passau, Germany
[4]School of English, Communication and Philosophy, Cardiff University, John Percival Building, Colum Drive, Cardiff, CF10 3EU, UK
[*]Corresponding author. School of English, Communication and Philosophy, Cardiff University, John Percival Building, Colum Drive, Cardiff, CF10 3EU, UK. E-mail: RobertsS55@cardiff.ac.uk

Associate editor: Prof. Michael Dunn

This study performs primary data collection, transcription, and cognate coding for eight South West Tibetic languages (Lowa, Gyalsumdo, Nubri, Tsum, Yohlmo, Kagate, Jirel, and Sherpa). This includes partial cognate coding, which analyses linguistic relations at the morpheme level. Prior resources and inferences are leveraged to conduct a Bayesian phylogenetic analysis. This helps estimate the extent to which the historical relationships between the languages represent a tree-like structure. We argue that small-scale projects like this are critical to wider attempts to reconstruct the cultural evolutionary history of Sino-Tibetan and other families.

Keywords: phylogenetics; Tibetic; Bayesian methods; historical linguistics.

Introduction

Recent computational techniques have facilitated new ways of studying historical linguistics. These include the quantitative estimation of historical relations from lists of cognate words using Bayesian phylogenetic inference to estimate the likely age of a language family and when languages diverged from each other (see, e.g. Hoffman et al. 2021). Previous studies have focussed on the Indo-European language family (e.g. Gray and Atkinson 2003; Nakhleh et al. 2005; Bouchkaert et al. 2012; Chang et al. 2015; Holm 2017; Rama 2018; Ritchie and Ho 2019; Heggarty et al. 2023), mainly because the considerable data required to perform these estimations is more readily available, and other large families (e.g. Pama-Nyungan, Bowern and Atkinson 2012), including attempts to create a global tree of languages (Jäger 2018; Bouckaert et al. 2022). More recently, these methods have been applied to a more diverse range of language families such as Dravidian (Kolipakam et al. 2018), Transeurasian (Robbeets and Bouckaert 2018), Turkic (Savelyev and Robbeets 2020), Tupi-Guarani (Ferraz Gerardi and Reichert 2021), and Mixtecan (Auderset et al. 2023), and smaller-scale projects such as studies of ten Chapacuran languages (Birchall et al. 2016) or dialects of Timor-Alor-Pantar (Kaiping and Klamer 2022). The resulting trees are important resources because they can be used in studies of the evolution of linguistic traits (e.g. the evolution of tone in Sino-Tibetan, Wu et al. 2023), but also to study the cultural evolution of non-linguistic traits (e.g. marital practices, Fortunato and Jordan 2010; folktales, Da Silva and Tehrani 2016; or subsistence, Ji et al. 2022) and as statistical controls in cross-cultural comparisons (e.g. Rácz et al. 2019; Shcherbakova et al. 2023).

However, applying these methods to smaller-scale linguistic groups is difficult because calibrating the dates in the tree relies on prior estimates of the depth of at least some parts of the tree. In this study, we study small and minority languages by leveraging existing resources and results. Our project includes primary data collection, cognate coding, and estimation of historical relations using Bayesian phylogenetics. We focus on eight Southwest Tibetic languages from the

Sino-Tibetan language family: Sherpa, Yohlmo, Jirel, Lowa, and Kagate, Nubri, Tsum, and Gyalsumdo. Previous work on the Sino-Tibetan language family includes three Bayesian phylogenetic studies that were published around the same time. Sagart et al. (2019) analyse 50 languages and estimate a tree depth for the family of around 7,200 years BP. Zhang et al. (2019) analyse 109 languages and estimate a slightly shallower tree depth of 5,900 years BP. Most recently, Zhang et al. (2020) analysed 131 languages and estimated a tree depth of about 8,000 years BP. The reasons for the conflict in the structure and depth of the trees have been debated (see, e.g. Wu et al. 2022). However, for this study, they offer a source of information to calibrate and structure the Bayesian phylogenetic processes that would be unavailable otherwise. For example, previous cognate sets can be used as a template for identifying the cognacy of forms in the target languages. This provides information well beyond the scope of our data. In addition, prior information for the dates of key branches can be extracted from the previously computed trees. These resources support us in 'filling in' the details to make better sense of the bigger picture. We hope that the methods employed here will make it possible for smaller, more specialist research groups to start contributing to the debate on computational methods for investigating historical linguistics.

Background

The Ethnologue lists 122 languages spoken and signed in Nepal from at least 5 different language families. The eight target languages included in this study are spoken in the mountainous region of Nepal Himalaya (see Table 1 and Fig. 1) and are located very close to Tibet (see Fig. 1). Several are geographically isolated, taking several days to reach from Kathmandu. All of these languages are Tibetic (or Bodic) in their genetic classification (e.g. Genetti 2016). These were chosen because some linguistic materials are available in these languages (e.g. Dhakal 2017a, b, 2018, 2019, 2020, 2023), but they have not been extensively documented. Among these eight languages, only five of them have been enumerated in the 2011 national census (Sherpa, Yohlmo, Jirel, Lowa, and Kagate), whereas speakers of Nubri and Tsum were counted together in 2021 and Gyalsumdo are yet to be recognized (see National Statistical Office 2023). Table 1 shows the location and population sizes of the languages (the latter taken from the 2021 census, or Ethnologue), showing that several languages have relatively few speakers. According to Glottoscope (Hammarström et al. 2018), the endangerment status of all of these is 'shifting' (language not being transmitted to the next generation) or 'threatened' (only some transition to the next generation). For example, based on an interview with local speakers, Hildebrandt and Perry (2011: 178) suggest that 'younger Gyalsumdo are only passive users of the language' and that 'there is an urgent need to gather this information while there is still access to regular speakers'.

Fig. 2 shows the current knowledge about the genetic classification of these languages according to Glottolog (Hammarström et al. 2023). The time depth of the splits is not known, and there are uncertainties about the structure of the tree around two multifurcations (Lowa/Jirel/Sherpa and Tusum to Gyalsumdo). Other disagreements exist about the structure of the tree. For example, Gyalsumdo has been classified within the 'southwestern section' of Tibetan (along with Humla, Mugu, Dolpo, Langtang, Kyirong, Lhomi, Walung, and Tokpe Gola; Tournadre 2014: 122). Similarly, Yohlmo is not part of Bradley's (1997) list of 'Central Bodish (Tibetan)' languages. Genetti (2016) enumerates these languages under the languages spoken in the northern

Table 1. The target languages in the current study.

Language	Glottolog code	Location	Latitude	Longitude	Speakers	Language status
Lowa	lowa1242	Ghilling village of the Mustang district	29.008472	83.858009	624 (2021)	Shifting
Gyalsumdo	gyal1236	Chame, district headquarters in Manang district	28.520045	84.235244	200 (2011)	Shifting
Nubri	nubr1243	Sama	28.636707	84.584472	4,284 (2021)	Shifting
Tsum	tsum1240	Nile	28.544032	85.104809		Shifting
Yohlmo	lamj1247	Paragang	27.96178	85.653074	9,658 (2021)	Shifting
Kagate	kaga1252	Likhutamakoshi	27.425911	86.177226	611 (2021)	Threatened
Jirel	jire1238	Jiri	27.635952	86.230586	5,167 (2021)	Shifting
Sherpa	sher1255	Dhunge Sangu	27.353164	87.570224	117,896 (2021)	Shifting

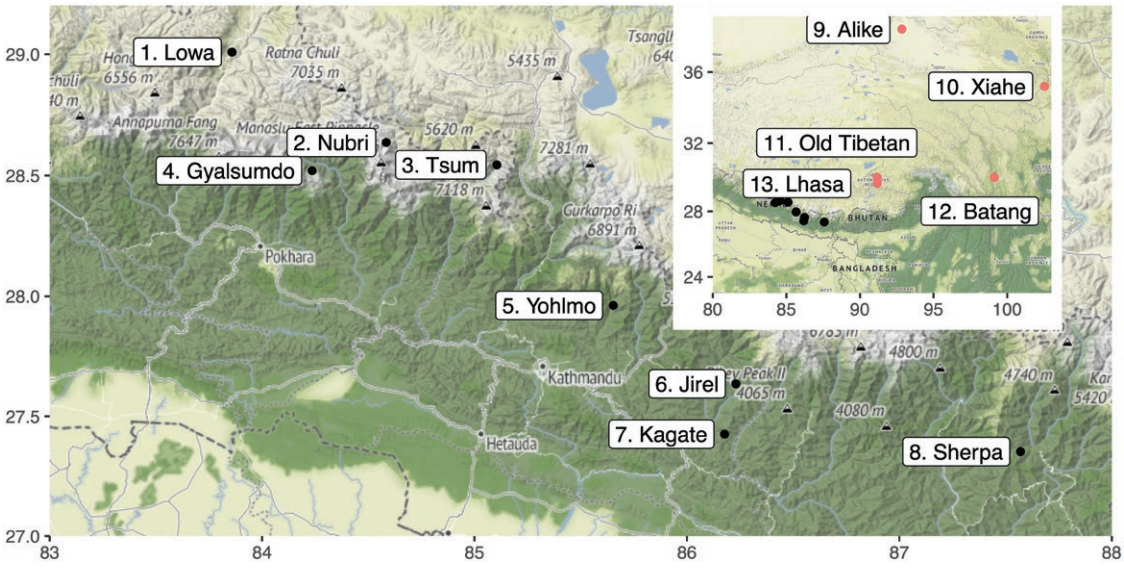

Figure 1. Map of the locations of the target languages and languages from Sagart et al. (insert). Numbers on the axes relate to latitude and longitude.

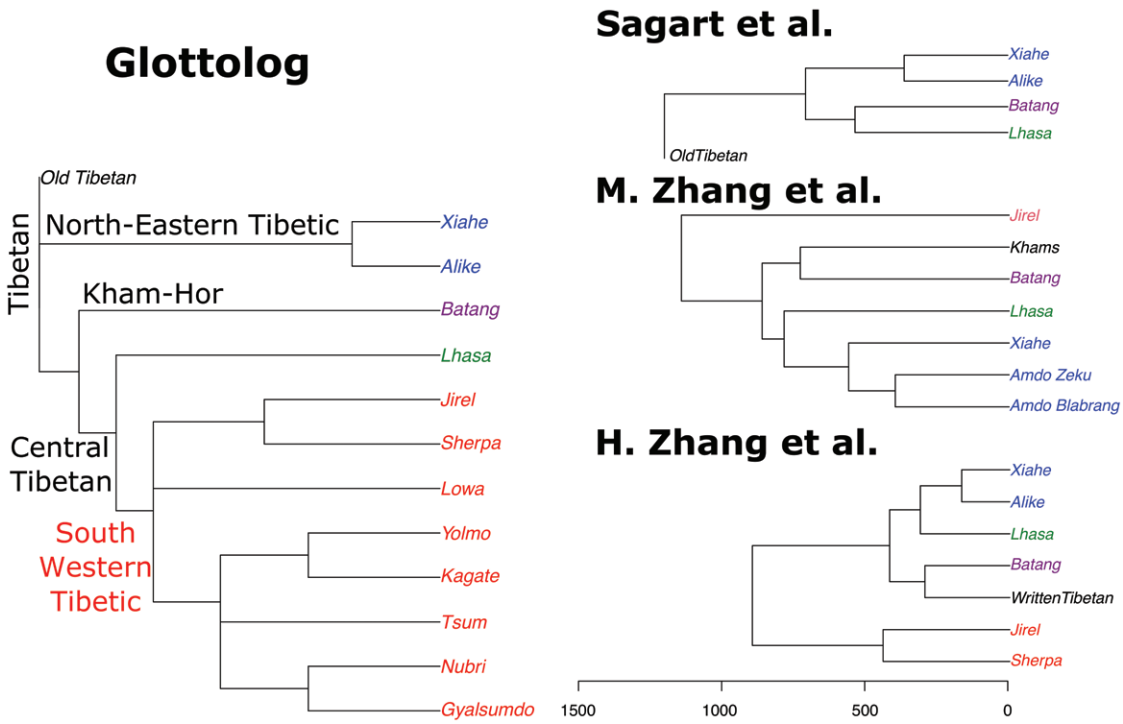

Figure 2. Genetic classification according to Glottolog (left, branch lengths are not meaningful) and Bayesian analyses.

Nepalese border area (Nepal) belonging to Central Tibetan languages, but Gyalsumdo is missing in this list.

In Sagart et al'.s maximum clade credibility tree (MCCT, the tree that has the most support in the posterior distribution of trees), the four existing Tibetan

languages follow the same structure as the Glottolog tree. H. Zhang's tree is also similar to the Glottolog tree structure, though it places Lhasa with Xiahe and Alike and has an overall shorter time depth than the other two Bayesian trees. However, H. Zhang et al. (2020) and M. Zhang et al. (2019) suggest that the Southwestern Tibetic clade diverged before the others. In fact, in M. Zhang's et al'.s (2019) tree the order of clade divergence from oldest to youngest is Southwestern Tibetic, Kam-Hor, Lhasa Tibetan, and North-Eastern Tibetic, effectively the reverse of the Glottolog order.

To leverage this existing work for analysing the target languages, data for an additional six languages were taken from Sagart et al. These included five languages in the Tibetan family (Batang, Old Tibetan, Lhasa, Xiahe, and Alike) and Old Chinese, which was used as an outgroup. Zhang et al.'s data include more Tibetan languages, but the available data for the Sagart paper was more extensive, including explicit phoneme alignments and formats compatible with various software tools for processing our data (LingPy and EDICTOR, see below). This made it much easier to integrate our data with the existing data. In addition, Sagart et al. proposed prior dating for Old Tibetan (1,200 before the present) and Old Chinese (2,500), based on archaeological evidence. Our suggestion is to calibrate our tree of Southwest Tibetic using these older languages and the estimation of the dating of the most recent common ancestor (in this case, the root of Sagart's tree, mean estimate of 7,184 before the present, with 95% highest posterior density interval [5,093–9,568]). In order to do this, new word list data need to be collected for the target languages and coded for cognacy alongside the previous data. Part of this step involves standardizing the two datasets, including extending the cognate judgements. The final cognate judgements are analysed manually, so while the two projects use very slightly different transcription conventions, this does not affect the cognate analysis.

Methods

Word list data

Word form data collection was led by D.N.D. A basic concept list was compiled in order to elicit the forms from the eight target languages. The concept list was built on the list of 210 concepts used in the Linguistic Survey of Nepal (see Gautam 2020), with 33 additional concepts focussing on words used in everyday conversation for the target languages.

A native speaker of each of the target languages was recruited. A list of words in Nepali for these concepts was shown to each native speaker and they were recorded producing equivalent words in their target language. The recordings were made between 4 August and 19 September 2018, in various places in the Kathmandu Valley (Lowa in Kirtipur; Syuba, Sherpa, and Yohlmo in Kapan; Nubri, Tsum, and Gyalsumdo in Swayambhu). These recordings were transcribed by Nepalese linguists using standard International Phonetic Alphabet (IPA) conventions. Since tone systems have not been formally described for all target languages, only a basic tone transcription was carried out.

Standardization

In order to facilitate both the data curation and the data reuse, the data were standardized, following the recommendations of the Cross-Linguistic Data Formats (CLDF) initiative (https://cldf.clld.org, Forkel et al. 2018). CLDF offers standard format representations of data in the form of comma-separated values extended by metadata and a specific ontology that allows us to check consistently to which degree a given dataset corresponds to the standards laid out by the CLDF initiative. As an example, when language varieties are complemented with Glottocodes (Forkel and Hammarström 2022), CLDF offers tools to check automatically whether the Glottocodes in question exist.

One of the core aspects of CLDF is to recommend integrating linguistic data with existing reference catalogues. Reference catalogues are metadata collections that assemble information on linguistic objects such as language varieties (Glottolog, https://glottolog.org, Hammarström et al. 2023, Version 4.8), concepts frequently used in wordlist elicitation (Concepticon, https://concepticon.clld.org, List et al. 2023, Version 3.1), or speech sounds (Cross-Linguistic Transcription Systems (CLTS), https://clts.clld.org, List et al. 2021, Version 2.1). The advantages of providing Glottocodes for language varieties, Concepticon concept set identifiers for the concepts for which data were collected, and making sure that phonetic transcriptions cohere to the standard laid out by CLTS(see Anderson et al. 2019) are 2-fold. On the one hand, existing data can be easily enriched with the additional information provided by the respective reference catalogues; on the other hand, data can be easily reused at later stages, and problems resulting from misinterpretations can be avoided.

In order to convert a given dataset into CLDF, CLDFBench, a Python package (https://pypi.org/project/cldfbench, Forkel, List & Rzymski, 2023 , Version 1.14) can be used (see Forkel and List 2022). The PyLexibank package (https://pypi.org/project/pylexibank, Forkel et al. 2022, Version 3.4.0) offers an extension to CLDFBench that provides additional integration of standards important for wordlists, including consistent handling of concepts through the

Concepticon and speech sounds through the CLTS reference catalogues (see List et al. 2022 for additional details on the standardization invoked by PyLexibank).

In order to standardize our data following the CLDF recommendations, we linked the language varieties to Glottolog, mapped the concepts to concept sets in Concepticon, and refined the IPA transcriptions to adhere to CLTS. Concept linking was facilitated since the base list of 210 items by Backstrom et al. (1992) was already linked to the Concepticon project (https://concepticon.clld.org/contributions/Backstrom-1992-210a).

Our workflow of data curation and data annotation now consists of two steps. Data are annotated in external tools, such as the EDICTOR tool discussed below. In regular intervals, data are converted to CLDF. The conversion allows to include automatic consistency checks of the data and allows for the conversion of the base data into additional formats needed to carry out specific analyses (such as phylogenetic reconstruction).

Since the data by Sagart et al. (2019) are also available in CLDF (curated on GitHub at https://github.com/lexibank/sagartst), and the concept list has been linked to the Concepticon as well (https://concepticon.clld.org/contributions/Sagart-2019-250), we could directly compute the overlap with our dataset, finding that there are 132 concepts both datasets have in common. This illustrates the reuse potential of standardized data, which also helps minimize the risk of individual errors.

Cognate coding

Cognate coding was carried out with the help of the EDICTOR, a web-based tool that facilitates various tasks in computer-assisted language comparison and offers specific modules for the coding of cognates (https://edictor.org, Version 2.1, List 2023; List and van Dam 2024). EDICTOR takes TSV files as input that can be easily generated from data stored in CLDF with the help of the PyEdictor tool (https://pypi.org/project/pyedictor/, Version 0.4, List 2021). Cognate codings can be carried out on the level of entire words and on the level of individual morphemes. The annotation is additionally supported by the possibility to carry out phonetic alignments, which in turn allow to inspect sound correspondence patterns in an interactive manner (List 2019), and by glossing individual morphemes in complex words (Hill and List 2017).

Initially, we attempted to merge our cognate coding with Sagart et al'.s using a novel automatic method. LingPy (List and Forkel 2023, Version 2.6.13, https://pypi.org/project/lingpy) was used to automatically generate starting suggestions for cognate sets. Within each concept, each form from the target languages was assigned to the best matching cognate class from the Sagart et al. data. Matching was done by weighted sequence alignment using the Needleman–Wunch algorithm (Needleman and Wunsch 2005, implemented in LingPy), and a form was assigned to the cognate set with the best average alignment score. However, although this might be a useful method for bootstrapping cognate coding for many language families, it became clear that an analysis at the morpheme level was necessary for the current data.

Like many Sino-Tibetan and South-East Asian languages, South-West Tibetic languages are also rich in complex words consisting of more than one base morpheme. Wu and List (2023) show that, for Southeast Asian languages, the approach to cognate coding of words with multiple morphemes can change cognate judgements and affect the estimation of a phylogeny. They suggest that partial cognate coding is required in order to make sure cognate judgements are based on the most relevant morphemes and to make the evidence for cognacy transparent. In order to provide the necessary cognate coding on entire word forms (rather than individual morphemes), they suggest converting partial cognates manually into 'full cognates' by paying specific attention to those morphemes in complex words whose individual meanings are judged to be 'salient' with respect to the meaning of the entire word.

In order to code cognates in the data, we proceeded in two steps. First, we coded partial and full cognates for the South-West Tibetic data alone, without taking additional data from Sino-Tibetan languages into account. In this stage, morphemes in complex words were glossed consistently, using morpheme glosses, phonetic alignments were carried out to facilitate the recognition of regular sound correspondence patterns, and sounds were grouped into 'evolving units' typically observed for South-East Asian languages (List et al. under review). Cognate judgments on entire words were based on partial cognate judgments, following the procedure outlined by Wu and List (2023). Coding the individual data collected as the basis for the study for cognates in this form has the advantage that it provides us with very specific information on regular sound correspondence patterns and language-internal cognates resulting from morphemes reused across several words in the same language. An example of such patterns of cognates spanning several concepts within the same language (but also across related languages) is given in Fig. 3.

In a second step, we created a combined dataset, consisting of the Tibetic languages in the dataset of Sagart et al. (2019) and Old Chinese as an outgroup for the purpose of phylogenetic reconstruction. Sagart et al. conducted cognate coding of their data on entire words, using phonetic alignments to make sure that their codings were consistent. When combining

ID	DOCULECT	CONCEPT	FORM	TOKENS	MORPHEMES	COGIDS	COGID
48	Gyalsumdo	Moon	ʈakaɹ	ʈ a + k aɹ	moon/da + star/white	15[16] 2565[22]	501[5]
51	Gyalsumdo	Star	kaɹma	k aɹ + m a	star/white + ma-suffix/ma	2565[22] 11[51]	759[8]
162	Gyalsumdo	White	kaɹpu	k aɹ + p u	star/white + :bu-suffix/pu	2565[22] 1581[54]	1106[6]

Figure 3. Morpheme *kaɹ* in Gyalsumdo, recurring in the words for 'moon', 'star', and 'white', pointing to a common motivation underlying the terms. Each row shows an individual lexical item. Forms and tokens are in IPA and are coloured by sound class according to the EDICTOR scheme (e.g. vowels are red). The morphemes column shows a gloss for each morpheme. The COGIDS column shows an ID for the cognate set at the morpheme level and the COGID shows the ID for the cognate set at the lexical level. Superscript numbers show the frequency of those cognate sets.

DOCULECT	CONCEPT	ALIGNMENTS	OLD	NEW
Gyalsumdo	TREE	ʈ o ŋ + b u		1039
Jirel	TREE	ⁿd o ŋ + b o		1039
Kagate	TREE	ʈ o ŋ + b o		1039
Lowa	TREE	ʈ o ŋ + b o		1039
Nubri	TREE	ʈ o ŋ + p o		1039
Sherpa	TREE	ɖ o ŋ + b u		1039
Tsum	TREE	ʈ o ŋ + b o		1039
Yolmo	TREE	ʈ u ŋ + b u		1039
Batang	TREE	x ĩ - 55 + pʰ ũ - 55		1448
Lhasa	TREE	ɕ i ŋ 55 + t u ŋ 55		1448
OldTibetan	TREE	ɕ i ŋ - - - - - -		1448
Alike	TREE	- - - - ɣ d o o ŋ + ŋ o	1039→	1449
OldTibetan	TREE	ɕ i ŋ + s d o o ŋ - - -	1448→	1449
Xiahe	TREE	- - - - - d o o ŋ + w o	1039→	1449
OldChinese	TREE	ᶜk + mᵉ o o k		1450

Figure 4. Changes to Sagart et al.'s cognate sets for the concept 'TREE', in the same format as Fig. 3. The column labelled 'OLD' shows Sagart et al.'s cognate set ID where it differs from the current data. The current cognate set IDs are shown in the column labelled 'NEW'.

the two datasets, we followed their base cognate judgments and adjusted the internal cognate judgments for South-West Tibetic where needed. However, as it turned out, almost no cognate judgements for the combined dataset split or lumped any cognate sets from Sagart et al.'s data. The exception is for the words for 'tree' (Fig. 4). The form in Old Tibetan is a compound, with the first part being /siŋ/ (= wood), and the second /sdooŋ/. Forms in Xiahe and Alike are clearly cognate with the second morpheme, so we corrected for this case.

This coding procedure results in 2 separate datasets, 1 consisting of 243 concepts translated into 8 South-West-Tibetan varieties, and one consisting of 127 concepts translated into 14 Sino-Tibetan varieties, including Old Tibetan and Old Chinese. All of our cognacy judgements are available to view online (https://edictor.org/edictor.html?file=tibetic-combined.tsv) or in the Supplementary data.

Filtering

Obvious borrowings from languages outside the sample and minor phonetic variants were removed (e.g. the word for NAME was borrowed from Chinese in several languages). Sagart et al. removed concepts that are likely to contain compounds (NOON and

FIREWOOD, see Sagart et al. SI p. 15), but this is not necessary in our coding because morphemes were coded separately.

Finally, 15 concepts were excluded because they did not have any variation in cognacy. At least 96% of concepts were represented in all languages (mean coverage = 99%). Sagart et al. excluded concepts if there were variants for less than 85% of languages in the sample. All concepts in the sample met this criteria except for 2 which had variants for only 11 out of 14 languages. However, these were kept in the sample since this is just one language under the threshold, and in one case the missing language was the outgroup. The final sample used in the Bayesian estimation consisted of 458 cognate sets in 117 concepts with (between 2 and 13 cognate sets per concept, mean of 3.9 sets per concept).

A neighbour net representing the distances between languages was calculated using the average normalized hamming distance between cognate codes for each language.

Phylogenetics

To analyse the historical relationships between languages, we used Bayesian phylogenetics. This is a process that aims to estimate a set of binary branching trees that explain the patterning of cognates within the existing languages. This is described in more detail elsewhere (e.g. Hoffman et al. 2021). Phylogenies were estimated using BEAST 2.7.5 (Bouckaert et al. 2019). The cognate data were converted to binary sites (each row representing a language and each column representing the presence or absence of a particular cognate set) in order to be used in the phylogenetic inference. Concepts were partitioned individually and ascertainment correction was added to each partition. To aid computation, it is typical to assume that all partitions follow the same evolutionary dynamics. However, Greenhill and Hoffman (2019, see Barido-Sottani and et al. 2018) note that it is reasonable to assume that concepts with larger variations in cognates may evolve at faster rates. Accordingly, we tested three site model configurations: (1) a single site model for all concepts, (2) two models (2–3 cognates—around 60% of the data—and 5 + cognates), and (3) three models (2–3 cognates, 4–6 cognates, 7+ cognates).

This analysis used a binary covarion model, which is typically used to model cognate date (Hoffman et al. 2021). We tested a strict clock and optimized uncorrelated relaxed clock with a log-normal distribution (Drummond et al. 2006), using a Fossilized Birth–Death model for tree generation (Heath et al. 2014). This model requires assumptions about the proportion of languages in the family that we have sampled. Glottolog lists Central Tibetan as having sixteen main varieties, but we are sampling some doculects classed as dialects. So we estimate there are between twenty and thirty varieties at a similar level to our doculects that we could have collected data for. We are sampling twelve of these, which suggests between eight and eighteen unsampled (40%–60%). So we used a beta distribution with $a = 8$, $b = 12$. The clock rate and bcov_alpha priors were set as uniform between 0 and 100.

The fossilization configuration allows us to distinguish present-day languages from extinct languages This prevents the model from assuming that the branch lengths for extinct languages extend to the present day and therefore skews the estimate of rates of change. Accordingly, the tip dates for Old Chinese and Old Tibetan were fossilized at 2,500 years before the present and 1,200 years before the present, respectively (see Sagart et al.).

Priors for the date of the root of our tree were taken from Sagart et al'.s final phylogenetic tree for the most recent common ancestor of the languages in our analysis (normal distribution with mean = 7,184 years BP, sigma = 500 years, 95% confidence intervals = [6200,8160]). A monophyletic prior was placed on all Tibetan languages to ensure that Old Chinese would remain an outgroup, with a relatively uninformative uniform prior between 0 and 13,000 years.

The phylogenetic process was run for 20,000,000 iterations. Tracer (Rambaut et al. 2018) was used to investigate the convergence and set a burn-in rate. The remaining iterations were resampled to provide a final distribution of 18,000 trees. TreeAnnotator (Helfrich et al. 2018) was used to identify the MCCT using median heights (common ancestor heights are not possible due to the inclusion of fossilized taxa).

Results

Fig. 5 shows a neighbour net for the average of the hamming distances for all concepts, excluding Old Tibetan and Old Chinese. This mirrors some structure of the Glottolog tree. For example, the main split is between the four languages from Sagart et al. and the eight languages based in Nepal. The latter shows three main groups: (1) Sherpa and Jirel, (2) Kagate and Yolmo, and (3) Nubri and Gyalsumdo. The relations between Tsum and Lowa are less clear. There is also considerable conflicting signal, suggesting that different concepts have different histories of inheritance between the languages.

A 10% burnin was applied to all traces. The model with the best posterior likelihood was the three-site relaxed clock model (see Appendix A1), and this was chosen as the main model. For this final model, all

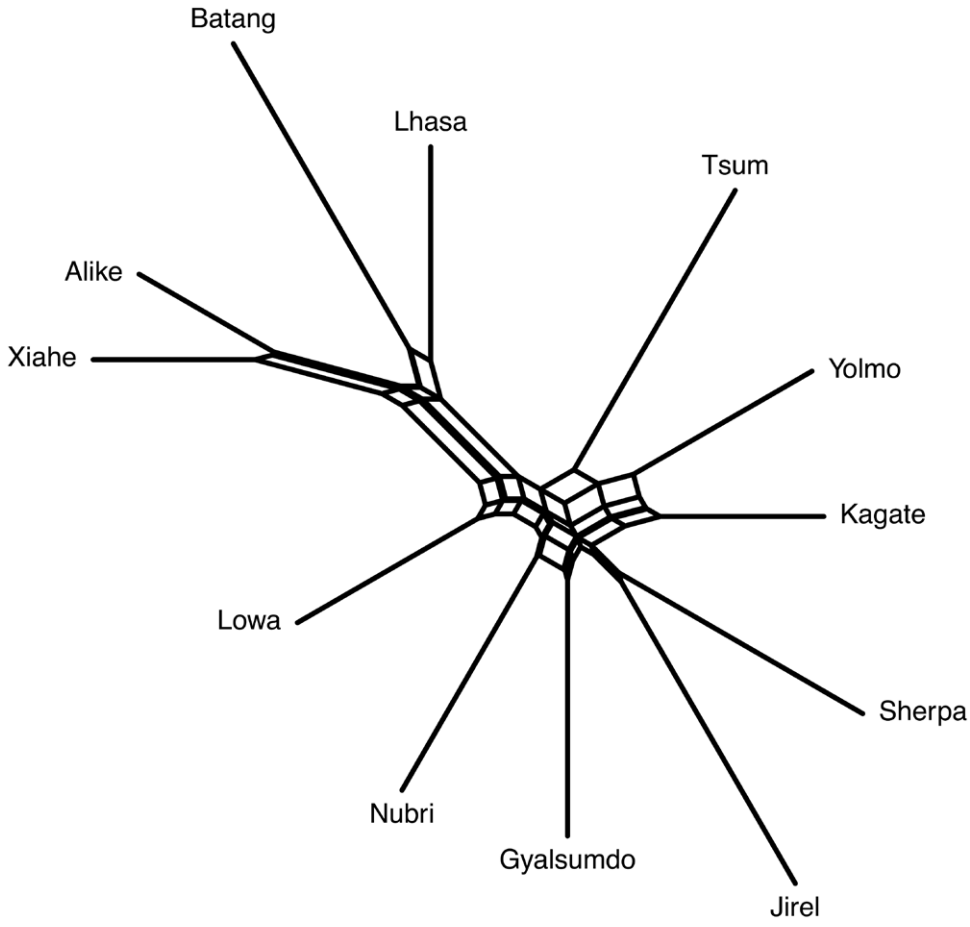

Figure 5. Neighbour net for the languages based on cognate differences.

clock and tree model ESS values were above 1,000 and the bcov and frequency values were above 300.

Fig. 6 shows the MCCT for the posterior sample of trees. This is the most representative single tree from the distribution. Initially, for simplicity, we describe the interpretation of this tree as if the languages diverged in a strictly binary-branching tree. Later, however, we analyse the full posterior sample of trees in order to identify the extent to which the data represents inheritance compared to a mix of inheritance and borrowing and other areal effects.

Regarding the MCCT: the root is estimated at around 6,000 years, which is lower than Sagart et al.'s estimate, but this is probably only because there are fewer languages in the current analysis. The final branch length of Old Tibetan is very short, but it is unlikely that this can be interpreted literally (that Old Tibetan went extinct immediately after its sister languages diverged from it). Rather, it indicates that the model is placing this split as recently as it can given that Old Tibetan is fossilized at 1,200 years. This might suggest that this fossilized date should be more recent (extending the Old Tibetan branch length), or that more data on missing languages is required to correctly estimate the amount of change in the Tibetic clade (pushing back the node that splits Old Tibetan from the rest of the Tibetic languages).

However, the relationships between existing languages are the focus of our investigation, not the root or fossilized languages. As expected, Xiahe, Alike, Batang, and Lhasa diverge from the rest of the existing languages first. According to this, Xiahe and Alike diverged from the rest of the languages around 900 years ago, and the eight target languages diverged within the last 600 years.

There are clear relationships between the phylogenetic structure and geographic distances between languages. Using the great circle distances between the main populations of these languages (based on the hometown of the informants) and the cophenetic distances in the phylogenetic trees, the correlations between the two are positive and highly significant, based on the Mantel test (for MCCT: Mantel $r = 0.85$,

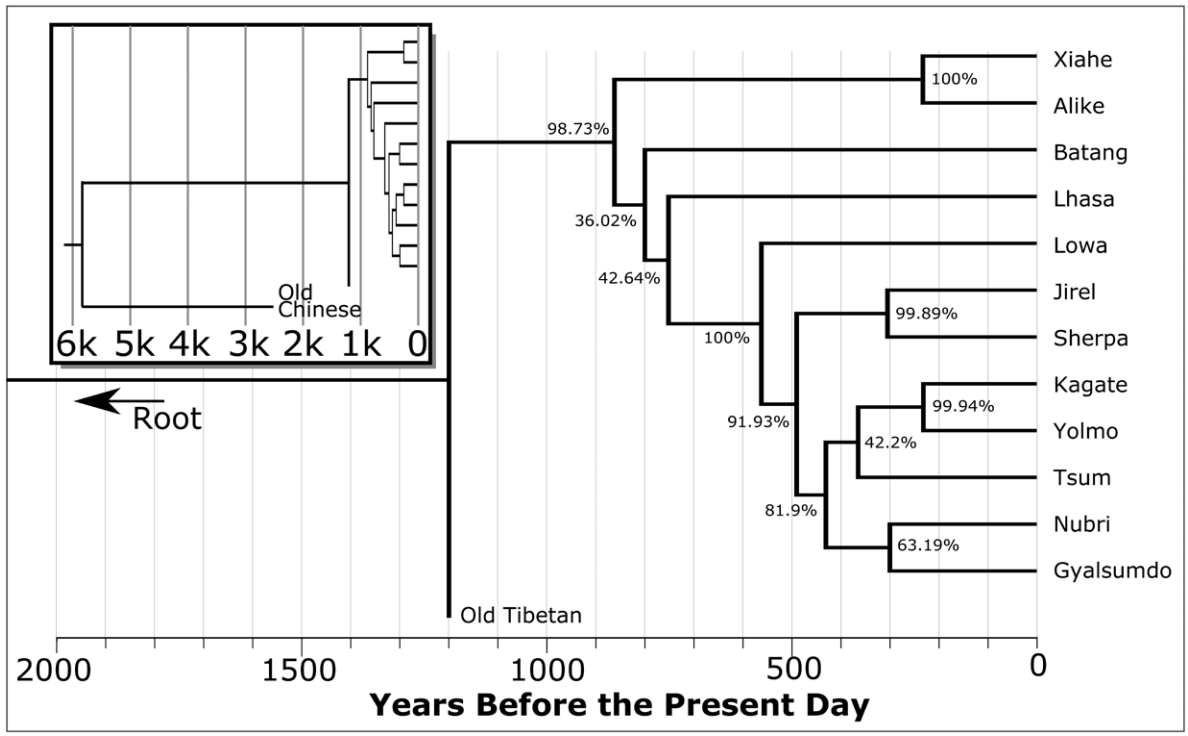

Figure 6. The Maximum Clade Credibility Tree. Branch lengths are scaled to represent years since the present day, shown on the bottom axis. Numbers at nodes represent the percentage of trees in the posterior sample that include the given split. The insert shows the full tree with the most recent common ancestor with Old Chinese.

$P < .001$) and congruence among distance matrices (for MCCT: Kendall's $W = 0.93$, $P < .001$; for posterior tree distribution: mean $W = 0.896$, min $W = 0.77$, max $P = .022$, for details see Koile et al. 2022). This correlation is on a similar scale to previous studies of the relationship between geography and linguistic divergence (e.g. Kolie et al. 2022). However, future analyses could take into account travel distance rather than geographic distance, especially since the topography of Nepal is an important factor in contact.

Some of the splits in the MCCT tree have good support in the posterior sample. For example, all trees in the posterior sample split the eight target languages from the four languages from Sagart et al. However, there are some less certain structures, as evident in some nodes with low representation in the posterior sample, which point to possible conflicting signals. Fig. 7 shows a 'densitree' that plots every tree in the posterior sample on top of each other. There are two major points of uncertainty. The first point is the structure that joins the northern languages with the Tibetan languages. In about half of all trees, Batang splits first, and then Lhasa splits from the Tibetan languages (Fig. 7 top left, for similar analyses, see King et al. 2024). In contrast, in about 43% of trees, Batang and Lhasa are on their own branch, and in a small proportion of trees, it is Lhasa that splits first.

The second major point of uncertainty is the position of Tsum in the Tibetan languages (Fig. 7 bottom left). Three topologies represent 58% of the tree distribution (a considerable proportion considering there are over 130,000 possible binary trees). The topology is identical to the consensus tree in each of these three topologies except for the position of Tsum. Tsum is placed either with Yolmo and Kagate (relatively distant languages in the Southeast), with Nubri and Gyalsumdo (relatively near to the West), or with Nubri (its closest neighbour). These uncertainties suggest multiple sources of inheritance for different concepts, which might have occurred through borrowing or contact, and are consistent with the conflicting signal shown in the Neighbour Net. To dig a little deeper into this uncertainty, Table 2 shows how cognate sets in Tsum are shared with the three other languages it is typically placed with. These reveal some interesting patterns. In general, there are many different patterns of sharing cognates, suggesting a complex history of relationships between the languages. Some of these have some obvious geographical explanations. For example, Tsum shares a cognate for WHEAT with Kagate, which is located 160 km to the Southeast in the

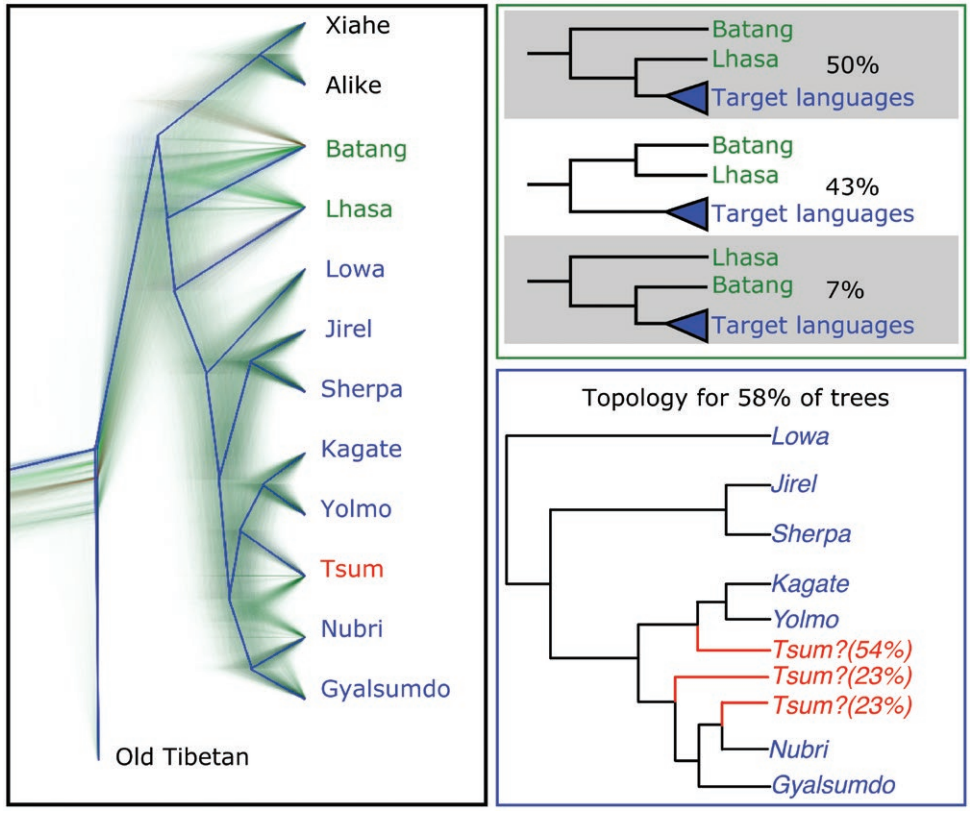

Figure 7. Left: Densitree of the posterior sample of trees, with the consensus tree overlaid as a solid line. The apparent vertical line for Old Tibetan is an artefact of this language being fossilized at 1,200 years. Top right: the distribution of topologies for the northern part of the tree. Bottom right: the topology of the target languages that is represented in 58% of the tree distribution, with the uncertainty of the location of T_{sum} represented.

Table 2. A table showing how cognates present in T_{sum} are present (green) or absent (red) in three other languages. For example, T_{sum}, Gyalsumdo, and Kagate have shared cognates for RIVER, but do not share cognates for RIVER in Nubri. Patterns where cognates are not present in T_{sum} are not shown.

T_{sum}	Gyalsumdo	Nubri	Kagate	CONCEPT
TRUE	TRUE	FALSE	TRUE	RIVER, MUD, MOSQUITO, FATHER, GOOD, WHERE, THAT, HAIL
TRUE	TRUE	TRUE	FALSE	EAR, FRUIT, HOT, RIGHT, WE (INCLUSIVE), MIDDAY, HEAR, SMOKE
TRUE	TRUE	FALSE	FALSE	BARLEY, COME
TRUE	FALSE	TRUE	TRUE	HOUSE, WOMAN, THIS, GIVE, STICK
TRUE	FALSE	FALSE	TRUE	SKIN, WHEAT, WIFE, WALK, RUN, THROW
TRUE	FALSE	TRUE	FALSE	WET, ALL, FAECES

Tibetan district and grows the most wheat (USDA 2024). On the other hand, Tsum shares a cognate for BARLEY with Gyalsumdo, 85 km to the West, which is closer to the districts that grow the most barley (USDA 2024). These seem like clear candidates for areal effects.

The MCCT shares a lot of structure with the Glottolog tree (Fig. 8). For example, Glottolog lists Jirel and Sherpa as a monophyletic pair and 99% of trees in the Bayesian estimate were consistent with this, showing good agreement. There is similar support for Xiahe and Alike (100% of trees) and Kagate and Yolmo (99% of trees). Other common pairs are less frequently represented such as Nubri and Gyalsumdo (67%). Other parts of the Glottolog tree are less

Figure 8. Comparison between the structures of the Glottolog tree (left) and Bayesian tree (right). Branch lengths on the left are not meaningful. Colours mark matching sub-clades for ease of comparison.

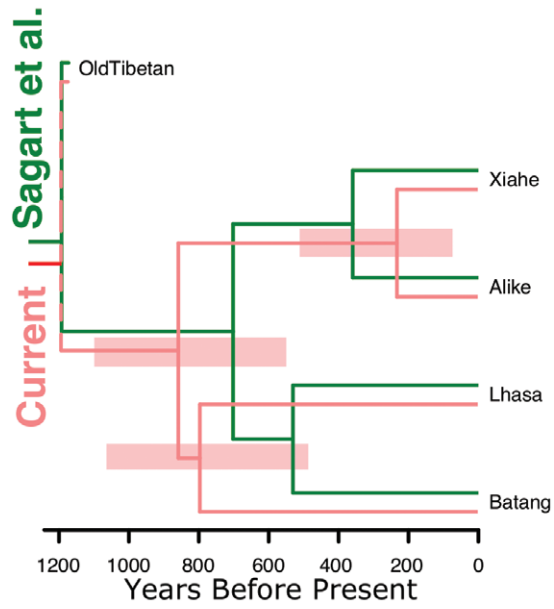

Figure 9. Comparison between the overlapping parts of the tree by Sagart et al. (dark green) and the current tree (light red). The transparent red bars show the 95% Height Posterior Density for each node of the current tree. Both have Old Tibetan fossilized at around 1,200 years.

supported by the data. For example, only 3% of the Bayesian trees included Kagate, Yolmo, Tsum, Nubri, and Gyalsumdo as an exclusive clade. The major difference is that the Bayesian tree suggests some resolutions to the polytomies in the Glottolog tree. The Bayesian tree has a high confidence that Lowa splits before the rest of the target languages, but as discussed above there is less certainty about the position of Tsum.

The structure of this study's tree is compatible with Sagart et al.'s tree, which is not surprising given this was the source of some of the data. Fig. 9 compares the branch lengths of the two trees, showing that the current tree has a more recent time for the split between Xiahe and Alike and an older time for the other two splits. The differences may be attributed to the slightly different subset of concepts analysed, but these are the median node ages, and the 95% HPDs for each node overlap between the two trees, suggesting that the differences may not be significant.

Fig. 10 compares the common parts of M. Zhang et al.'s tree, Z. Zhang et al.'s tree, and the MCCT tree from the current study. The two previous trees suggest that the Southwest Tibetic languages diverged before the northern languages, whereas the current tree suggests the opposite. In terms of dates, the current tree's common ancestor of Southwestern Tibetic and North-Eastern Tibetic is similar to H. Zhang et al.'s tree, but is shallower than M. Zhang et al.'s study. While estimates of time depths can be affected by the amount of data (more data = more variation = longer time depth), this is an unlikely explanation for this difference since H. Zhang's tree has the most languages.

Conclusion

In this project, we collected lexical data from eight languages, used digital tools to identify cognates, and leveraged previous resources to estimate a phylogenetic tree of relations between them. The results suggest that the languages diverged within the last 600 years, and the overall structure of the tree shows good agreement with previous linguistic classification. There are now

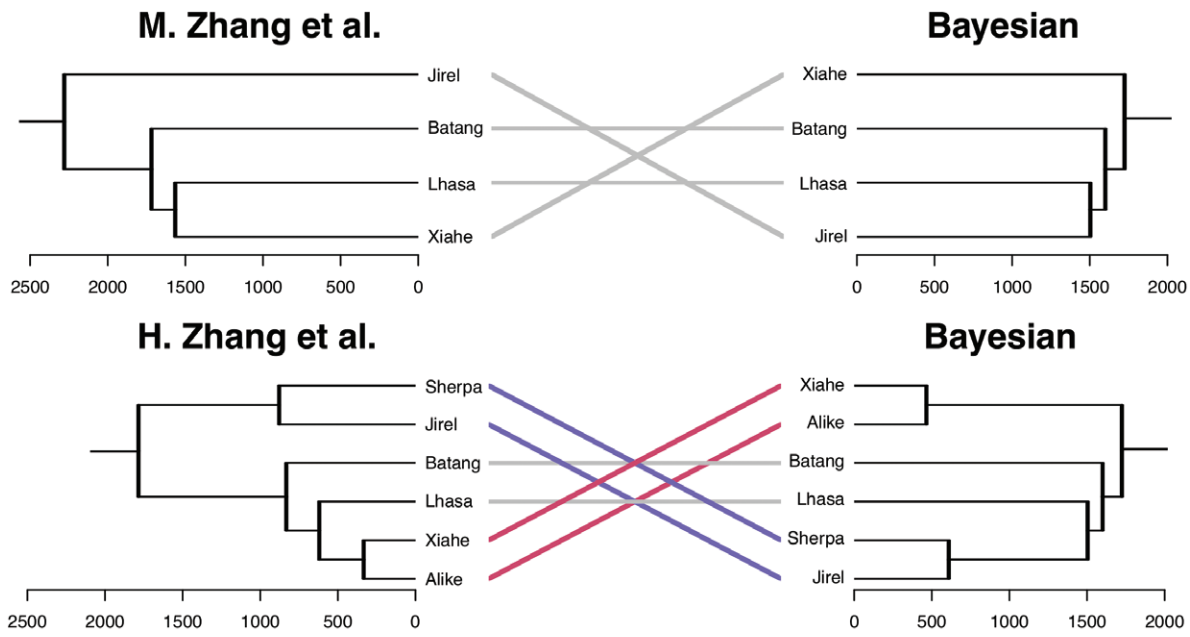

Figure 10. Comparisons between common languages from M. Zhang et al.'s tree (top), H. Zhang et al.'s tree (bottom) and the current Bayesian tree. Numbers indicate years before present. Colours mark matching sub-clades for ease of comparison.

several different phylogenetic analyses of Sino-Tibetan. One of the central disagreements is the placement of Southwestern Tibetic in the divergence from other branches of Tibetic languages. The MCCT estimated in this study agrees more with the tree from Glottolog and Sagart et al. (placing the divergence more recently), rather than the tree from M. Zhang et al. or H. Zhang et al. In fact, the Glottolog tree and the current MCCT tree suggest a linguistic inheritance from North to South, whereas the other two trees suggest a linguistic inheritance from South to North. However, more work needs to be done to support this geographic interpretation. For example, the placement of Central Tibetan, Southwestern Tibetan and Kham-Hor clades is not well resolved in the posterior distribution of this study, suggesting that further data collection and modelling are required to resolve this question. The source of the uncertainty is the amount of conflicting signals in the data, suggesting that different concepts have different patterns of inheritance. This is likely due to language contact, and indeed there are known complex contact dynamics in this region (see Hildebrandt et al. 2023).

Further work still is required to relate this linguistic inheritance to population movement and wider history. Indeed, the history of languages in Nepal provides the critical evidence that differentiates various theories of population movement in Sino-Tibetan: Van Driem (2005) and LaPolla (2001) suggest population movement from North West to South East Nepal, Blench (2009) suggests movement from South of Nepal to the North, and the recent Bayesian phylogenetic trees suggest movement from South East to North West Nepal. Therefore, obtaining more data on the wide variety of languages in Nepal is critical to making progress on these debates (see King et al. 2024 for a similar approach to sampling Philippine languages to clarify Malayo-Polynesian history).

However, the current results also point to a more complex historical relationship between the languages than simple binary branching varieties. The partial cognate coding at the morpheme level was critical for identifying cognates. It also had a considerable influence on the final cognate coding. Compared to our initial (non-partial) cognate coding, 37% of concepts had some change with a total of 70 splits and mergers, and the MCCT had a slightly different topology and less overall posterior support. It may be the case that other Sino-Tibetan analyses could benefit from this more detailed approach. More generally, rather than trying to reach further back in time and wider in geographic scope, it may be more informative, and more sustainable, to concentrate on collecting and compiling linguistic data on sampling families in more depth. In this way, we hope that small-scale projects like this one can engage with the cutting edge of the field and contribute to the wider picture of linguistic history.

Supplementary Data

Supplementary data is available at *Journal of Language Evolution Journal* online.

Acknowledgements

S.G.R. and D.N.D. were supported by an International Strategic Fund grant from the University of Bristol. S.G.R. was additionally supported by an AHRC grant AH/T006927/1. J.M.L. was supported by the Max Planck Society Research Grant CALC3 (https://digling.org/calc/) and the ERC Consolidator Grant ProduSemy (grant no. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Data availability

The linguistic data and cognate codings are available to view on the EDICTOR website: https://edictor.org/edictor.html?file=tibetic-combined.tsv. The data and code for the project are available online: Zenodo archive: https://doi.org/10.5281/zenodo.13868534; Github repository version 1.0: https://github.com/lexibank/dhakalsouthwesttibetic; Code and data for running the Bayesian phylogenetic analyses on Github: https://github.com/seannyD/SouthWestTibeticLexicons

References

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., and List, J.-M. (2019) A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4: 21–53. https://doi.org/10.2478/yplm-2018-0002

Auderset, S. et al. (2023) 'Subgrouping in a 'Dialect Continuum': A Bayesian Phylogenetic Analysis of the Mixtecan Language Family', *Journal of Language Evolution*, 8: 33–63. https://doi.org/10.1093/jole/lzad004

Barido-Sottani, J. et al. (2018) 'Taming the BEAST—A Community Teaching Material Resource for BEAST 2', *Systematic Biology*, 67: 170–4.

Birchall, J., Dunn, M., and Greenhill, S. J. (2016) 'A Combined Comparative and Phylogenetic Analysis of the Chapacuran Language Family', *International Journal of American Linguistics*, 82: 255–84. https://doi.org/10.1086/687383

Blench, R. (2009) 'If Agriculture Cannot be Reconstructed for Proto-Sino-Tibetan What are the Consequences?' in Takerngrangsarit P. (ed.), *42nd Conference on Sino-Tibetan Language and Linguistics*. SIL, Chiang Mai: 1.

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., and Atkinson, Q. (2022) Global language diversification is linked to socio-ecology and threat status. *SocArXiv*. https://doi.org/10.31235/osf.io/f8tr6

Bouckaert, R. et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337: 957–60. https://doi.org/10.1126/science.1219669

Bouckaert, R. et al. (2019) 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 15: e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bowern, C., and Atkinson, Q. (2012) 'Computational Phylogenetics and the Internal Structure of Pama-Nyungan', *Language*, 88: 817–45. https://doi.org/10.1353/lan.2012.0081

Bradley, D. (1997) 'Tibeto-Burman Languages and Classification', in *Papers in Southeast Asian Linguistics No. 14: Tibeto-Burman languages of the Himalayas*, pp. 1–72. Canberra: Pacific Linguistics (series A-86).

Chang, W. et al. (2015) 'Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis', *Language*, 91, 194–244.

Da Silva, S. G., and Tehrani, J. J. (2016) 'Comparative Phylogenetic Analyses Uncover the Ancient roots of Indo-European Folktales', *Royal Society Open Science*, 3: 150645. https://doi.org/10.1098/rsos.150645

Dhakal, D. N. (2017a) 'Lowa Case Markers in Comparative Perspective', *The Journal of University Grants Commission*, 6: 16–28.

Dhakal, D. N. (2017b) 'Noun Phrase Structure in Tsum', *Interdisciplinary Journal of Linguistics*, 10: 73–84.

Dhakal, D. N. (2018) 'Morphosyntax of Adjectives in Gyalsumdo (Nepal)', *Interdisciplinary Journal of Linguistics*, 11: 87–96.

Dhakal, D. N. (2019) Nubri - English -Nepali Dictionary. Report submitted to National Foundation for Development of Indigenous Nationalities, Lalitpur, Ms.

Dhakal, D. N. (2020) 'Kinship Terms in Gyalsumdo, Nubri and Tsum', *Linguistic Society Of India*, 81.

Dhakal, D. N. (2023) *A Sketch of Nubri Morphosyntax*. Muenchen: Lincom GmbH.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88. https://doi.org/10.1371/journal.pbio.0040088

Ferraz Gerardi, F., and Reichert, S. (2021) 'The Tupí-Guaraní Language Family: A Phylogenetic Classification', *Diachronica*, 38: 151–88. https://doi.org/10.1075/dia.18032.fer

Forkel, R. and Hammarström H. (2022). 'Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information', In: J. Bosque-Gil, M. Dojchinovski, P. Cimiano, J. Bosque-Gil, P. Cimiano and M. Dojchinovski (eds.) *Semantic Web*, 13: 917–24. https://doi.org/10.3233/sw-212843

Forkel, R. et al. (2018) 'Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics', *Scientific Data*, 5: 1–10. https://doi.org/10.1038/sdata.2018.205

Forkel, R. et al. (2022) *PyLexibank [Software Library, Version 3.4]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/pylexibank

Forkel, R., List, J-M, and Rzymski, Christoph (2023). CLDFBench. *Tooling to create CLDF datasets from existing data [Software library, Version 1.14.0]*. https://pypi.org/project/cldfbench/

Fortunato, L., and Jordan, F. (2010) 'Your Place or Mine? A Phylogenetic Comparative Analysis of Marital Residence in Indo-European and Austronesian societies', *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 365: 3913–22. https://doi.org/10.1098/rstb.2010.0017

Gautam, B. L. (2020) Sociolinguistic survey of Nepalese languages. *Language Ecology*, 3: 189–208. https://doi.org/10.1075/le.19004.gau

Genetti, C. 2016. 'The Tibeto-Burman Languages of South Asia'. In: H. H. Hock and E. Bashir (eds) *The Languages and Linguistics of South Asia: A comprehensive guide*, pp. 130–54. Berlin/Boston: Walter de Gruyter.

Gray, R. D., and Atkinson, Q. D. (2003) 'Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin', *Nature*, 426: 435–9. https://doi.org/10.1038/nature02029

Greenhill, S. J., and Hoffmann, K. (2019) Language Phylogenies: Using Babel to Analyse Linguistic Data Background. https://taming-the-beast.org/tutorials/LanguagePhylogenies/

Hammarström, H. et al. (2018). Simultaneous visualization of language endangerment and language description.

Hammarström, H., Forkel, R., Haspelmath, M., Bank, S. (2023) *Glottolog 4.8*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 0–1. https://doi.org/10.5281/zenodo.8131084

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014) 'The Fossilized Birth–Death Process for Coherent Calibration of Divergence-Time Estimates', *Proceedings of the National Academy of Sciences*, 111: E2957–66.

Heggarty, P. et al. (2023) 'Language Trees with Sampled Ancestors Support a Hybrid Model for the Origin of Indo-European Languages', *Science*, 381: eabg0818. https://doi.org/10.1126/science.abg0818

Helfrich, P. et al. (2018). 'TreeAnnotator: Versatile Visual Annotation of Hierarchical Text Relations' in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association, Miyazaki, Japan.

Hildebrandt, K. A., Bond, O., and Dhakal, D. N. (2023) 'A Micro-Typology of Contact Effects in Four Tibeto-Burman Languages', *Journal of Language Contact*, 15: 302–40. https://doi.org/10.1163/19552629-15020003

Hildebrandt, K. A., and Perry, J. J. (2011) 'Preliminary Notes on Gyalsumdo, an Undocumented Tibetan Variety in Manang District, Nepal', *Himalayan Linguistics*, 10: 167–185.

Hill, N. W., List, J.-M. (2017) Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznan Linguistic Meeting*, 3: 47–76. https://doi.org/10.1515/yplm-2017-0003

Hoffmann, K. et al. (2021) 'Bayesian Phylogenetic Analysis of Linguistic Data Using BEAST', *Journal of Language Evolution*, 6: 119–35. https://doi.org/10.1093/jole/lzab005

Holm, H. J. (2017) 'Steppe Homeland of Indo-Europeans Favored by a Bayesian Approach with Revised Data and Processing', *Glottometrics*, 37: 54.

Jäger, G. (2018) 'Global-scale Phylogenetic Linguistic Inference from Lexical Resources', *Scientific Data*, 5: 1–16.

Ji, T. et al. (2022) 'A Phylogenetic Analysis of Dispersal Norms, Descent and Subsistence in Sino-Tibetans', *Evolution and Human Behavior*, 43: 147–54. https://doi.org/10.1016/j.evolhumbehav.2021.12.002

Kaiping, G. A., and Klamer, M. (2022) 'The Dialect Chain of the Timor-Alor-Pantar language Family: A New Analysis Using Systematic Bayesian Phylogenetics', *Language Dynamics and Change*, 12: 274–326. https://doi.org/10.1163/22105832-bja10019

King, B. et al. (2024) 'Bayesian Phylogenetic Analysis of Philippine Languages Supports a Rapid Migration of Malayo-Polynesian Languages', *Scientific Reports*, 14: 14967. https://doi.org/10.1038/s41598-024-65810-x

Koile, E. et al. (2022) 'Geography and Language Divergence: The Case of Andic Languages', *PLoS One*, 17: e0265460. https://doi.org/10.1371/journal.pone.0265460

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., and Verkerk, A. (2018) A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5: 171504–171504. https://doi.org/10.1098/rsos.171504

LaPolla, R. J. (2001) 'The Role of Migration and Language Contact in the Development of the Sino-Tibetan Language Family', in A. Aikhenvald and R. M. W. Dixon (eds) *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, pp. 225–54, Oxford University Press.

List, J.-M. (2019) 'Automatic Inference of Sound Correspondence Patterns Across Multiple Languages', *Computational Linguistics*, 45: 137–61. https://doi.org/10.1162/coli_a_00344

List, J.-M. (2021) PyEdictor [Software Library, Version 0.4]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/pyedictor

List, J.-M., Anderson, C., Tresoldi, T., and Forkel, R. (2021) *Cross-Linguistic Transcription Systems*. Version 2.1.0. Max Planck Institute for the Science of Human History: Jena. https://clts.clld.org/

List, J.-M. et al. (2022) 'Lexibank, A Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features', *Scientific Data*, 9: 1–31. https://doi.org/10.20396/liames.v22i00.8669038

List, J.-M. et al. (2023) CLLD Concepticon [Dataset, Version 3.1.0]. Version 3.1.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, J.-M. (2023) *EDICTOR: A Web-Based Interactive Tool For Creating And Editing Etymological Datasets. Version 2.1.0*. Passau: MCL Chair at the University of Passau. https://digling.org/edictor/. Date accessed 01 October 2024.

List, J.-M., and Forkel, R. (2023) *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.13]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/lingpy. Date accessed 01 October 2024.

List, J.-M., and van Dam, K. P. (2024) 'Computer-Assisted Language Comparison with EDICTOR 3' in N. Tahmasebi et al. (eds) Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change. Association for Computational Linguistics, Bangkok.

Nakhleh, L. et al. (2005) 'A Comparison of Phylogenetic Reconstruction Methods on an Indo-European Dataset', *Transactions of the Philological Society*, 103: 171–92. https://doi.org/10.1111/j.1467-968x.2005.00149.x

National Statistical Office. (2023) *National Population and Housing Census/ Census of Population and National Report on Caste/Ethnicity, Language & Religion*. Kathmandu: National Statistical Office.

Needleman, S. B., Wunsch, C. D. (2005) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular*

Biology, 48: 443–53. https://doi.org/10.1016/0022-2836(70)90057-4
Rama, T. (2018) 'Three Tree Priors and Five Datasets: A Study of Indo-European Phylogenetics', *Language Dynamics and Change*, 8: 182–218. https://doi.org/10.1163/22105832-00802005
Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901–4. https://doi.org/10.1093/sysbio/syy032
Rácz, P., Passmore, S., and Jordan, F. M. (2019) Social Practice and Shared History, Not Social Scale, Structure Cross-Cultural Complexity in Kinship Systems. *Topics in Cognitive Science*, 12: 744–65. https://doi.org/10.1111/tops.12430
Ritchie, A. M., and Ho, S. Y. (2019) 'Influence of the Tree Prior and Sampling Scale on Bayesian Phylogenetic Estimates of the Origin Times of Language Families', *Journal of Language Evolution*, 4: 108–23.
Robbeets, M., Bouckaert, R. (2018) Bayesian phylolinguistics reveals the internal structure of the Transeurasian family. *Journal of Language Evolution*, 3: 145–62. https://doi.org/10.1093/jole/lzy007
Sagart, L. et al. (2019) 'Dated Language Phylogenies Shed Light on the Ancestry of Sino-Tibetan', *Proceedings of the National Academy of Sciences of the United States of America*, 116: 10317–22. https://doi.org/10.1073/pnas.1817972116
Savelyev, A., and Robbeets, M. (2020) 'Bayesian Phylolinguistics Infers the Internal Structure and the Time-Depth of the Turkic Language Family', *Journal of Language Evolution*, 5: 39–53. https://doi.org/10.1093/jole/lzz010
Shcherbakova, O. et al. (2023) 'Societies of Strangers Do Not Speak Less Complex Languages', *Science Advances*, 9: eadf7704. https://doi.org/10.1126/sciadv.adf7704
Tournadre, N. (2014) 'The Tibetic Language and Their Classification', in N. Hills and T. Owen-Smith (eds) *Trans-Himalayan Linguistics*, pp. 105–30. Berlin: DeGruyter.
USDA (2024) Nepal production. https://ipad.fas.usda.gov/countrysummary/Default.aspx?id=NP
van Driem (2005) 'Tibeto-Burman vs Indo-Chinese', in L. Sagart, Roger B.,; A. Sanchez-Mazas (eds.) *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, pp. 81–106. London: Routledge Curzon.
Wu, B., Zhang, H., and Zhang, M. (2023) 'Phylogenetic Insight into the Origin of Tones', *Proceedings Biological Sciences*, 290: 20230606. https://doi.org/10.1098/rspb.2023.0606
Wu, M. S., Bodt, T. A., and Tresoldi, T. (2022) 'Bayesian Phylogenetics Illuminate Shallower Relationships Among Trans-Himalayan Languages in the Tibet-Arunachal Area', *Linguistics of the Tibeto-Burman Area*, 45: 171–210. https://doi.org/10.1075/ltba.21019.wu
Wu, M. S., and List, J. M. (2023) 'Annotating Cognates in Phylogenetic Studies of Southeast Asian Languages', *Language Dynamics and Change*, 13: 161–97. https://doi.org/10.1163/22105832-bja10023
Zhang, H. et al. (2020) 'Dated Phylogeny Suggests Early Neolithic Origin of Sino-Tibetan Languages', *Scientific Reports*, 10: 20792. https://doi.org/10.1038/s41598-020-77404-4
Zhang, M. et al. (2019) 'Phylogenetic Evidence for Sino-Tibetan Origin in Northern China in the Late Neolithic', *Nature*, 569: 112–5. https://doi.org/10.1038/s41586-019-1153-z

Appendix

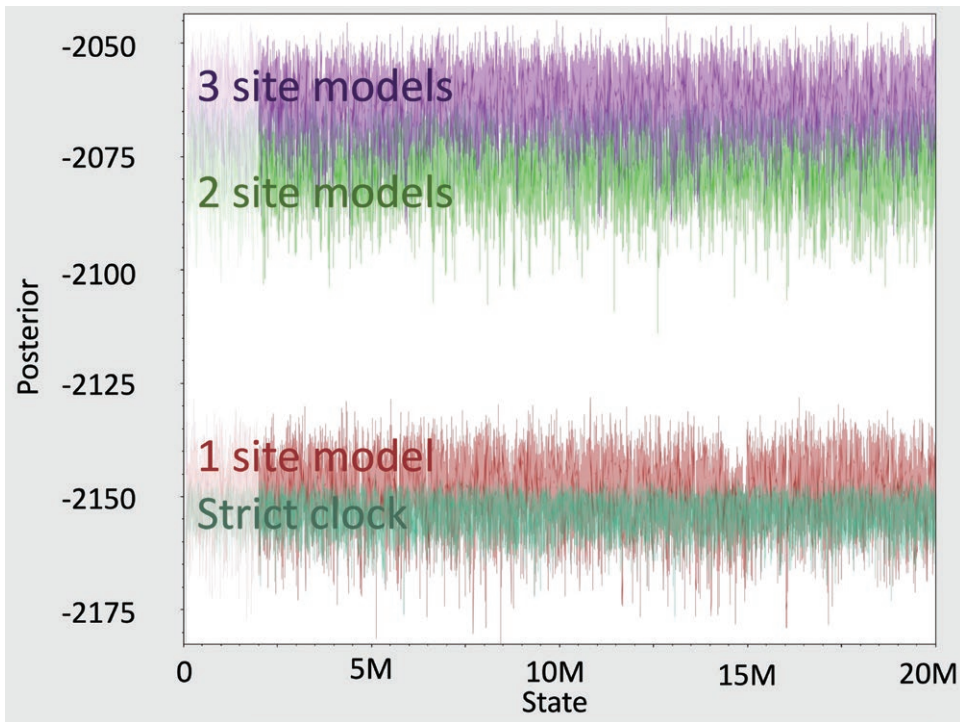

Figure A1. Comparison of the posterior likelihood over the MCMC chain for the four models (strict clock, and three relaxed clock models with one site for all cognates, two sites, and three sites).