

A Computer-Assisted Approach to the Comparison of Mainland Southeast Asian Languages

Dissertation
(kumulativ)

Zur Erlangung des akademischen Grades doctor philosophiae
(Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von Mei-Shin Wu-Urbaneck, M.Sc. Bioinformatik, M.A. Computerlinguistik
geboren am 14.09.1985 in Pingtung, Taiwan

Gutachter/Gutachterin:

1. Prof Dr. Johann-Mattis List, Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology; Chair of Multilingual Computational Linguistics, University of Passau
2. Prof. Dr. Volker Gast, Department of English and American Studies, Friedrich Schiller University Jena

Tag der Verteidigung: 13. Februar 2023

Acknowledgement

Working with the Max Planck Institute for Evolutionary Anthropology has allowed me to broaden my perspective on scientific work, appreciate the languages we speak, and think more broadly about many social issues. Both of my supervisors were supportive during the time I worked on the dissertation. I want to thank my supervisors and colleagues for answering my questions without hesitation.

I am grateful to my husband and our family members for their support and care for me throughout the entire process of working on projects and completing my dissertation. I am motivated by the trust they have instilled in me.

Last but not least, I thank my readers for taking the time to read the dissertation. I hope you find the methodology I present interesting. I'm looking forward to hearing from you.

Abstract

This cumulative thesis is based on three separate projects based on a computer-assisted language comparison (CALC) framework to address common obstacles to studying the history of Mainland Southeast Asian (MSEA) languages, such as sparse and non-standardized lexical data, as well as an inadequate method of cognate judgments, and to provide caveats to scholars who will use Bayesian phylogenetic analysis.

The first project provides a format that standardizes the sound inventories, regulates language labels, and clarifies lexical items. This standardized format allows us to merge various forms of raw data. The format also summarizes information to assist linguists in researching the relatedness among words and inferring relationships among languages.

The second project focuses on increasing the transparency of lexical data and cognate judgments with regard to compound words. The method enables the annotation of each part of a word with semantic meanings and syntactic features. In addition, four different conversion methods were developed to convert morpheme cognates into word cognates for input into the Bayesian phylogenetic analysis.

The third project applies the methods used in the first project to create a workflow by merging linguistic data sets and inferring a language tree using a Bayesian phylogenetic algorithm. Furthermore, the project addresses the importance of integrating cross-disciplinary studies into historical linguistic research.

Finally, the methods we proposed for managing lexical data for MSEA languages are discussed and summarized in six perspectives. The work can be seen as a milestone in reconstructing human prehistory in an area that has high linguistic and cultural diversity.

Abstract

Diese kumulative Dissertation basiert auf drei separaten Projekten, die sich auf einen computergestützten Sprachvergleich (CALC) stützen, um häufige Hindernisse bei der Erforschung der Geschichte der südostasiatischen Festlandssprachen (MSEA) zu beseitigen, wie z. B. unzureichende und nicht standardisierte lexikalische Daten, eine unzureichende Methode zur Beurteilung von Verwandtschaftsanalysen sowie Wissenschaftlern Hilfestellung bei der Bewertung von Bayesian phylogenetischen Analysen zu geben.

Im ersten Projekt wird ein Format zur Verfügung gestellt, welches Sprachlautsammlungen standardisiert, Sprachbezeichnungen regelt und lexikalischen Einträge klärt. Dieses standardisierte Format ermöglicht die Zusammenführung verschiedenster Formen von Rohdaten. Zudem fasst das Format auch Informationen zusammen, die Linguisten bei der Erforschung der Verwandtschaft zwischen Wörtern und der Ableitung von Beziehungen zwischen Sprachen unterstützen können.

Das zweite Projekt konzentriert sich auf die Erhöhung der Übersichtlichkeit lexikalischer Daten und Verwandtschaftsanalysen in Bezug auf zusammengesetzte Wörter. Die Methode ermöglicht es, jeden Teil eines Wortes mit semantischen Bedeutungen und syntaktischen Merkmalen zu annotieren. Darüber hinaus wurden vier verschiedene Konvertierungsmethoden entwickelt, um Morphemverwandtschaften in Wortverwandtschaften umzuwandeln und in der Bayesian phylogenetischen Analyse zu verwenden.

Im dritten Projekt werden die im ersten Projekt verwendeten Methoden angewendet, um einen Arbeitsablauf zu erstellen, welcher linguistische Datensätze zusammenführt und mit Hilfe eines Bayesian phylogenetischen Algorithmus einen Sprachbaum abzuleiten. Darüber hinaus befasst sich das Projekt mit der Bedeutung der Integration interdisziplinärer Studien in die historische Sprachforschung.

Schließlich werden die von uns vorgeschlagenen Methoden zur Verwaltung lexikalischer Daten für MSEA-Sprachen diskutiert und in sechs Gesichtspunkten zusammengefasst. Die Arbeit kann als Meilenstein bei der Rekonstruktion der menschlichen Vorgeschichte in einem Gebiet mit großer sprachlicher und kultureller Vielfalt angesehen werden.

Contents

1	Introduction	1
1.1	The Research Territory	1
1.2	Language Family and Language Area	3
1.3	The SEA Languages	5
1.3.1	Large Vowel Systems	6
1.3.2	Strict Syllable Pattern	8
1.3.2.1	Major and minor syllables	9
1.3.3	Compound Words	10
1.3.4	Complex Tone Systems	11
1.4	The Comparative Method	14
1.4.1	Classical Approach	14
1.4.2	Computational Approach	17
1.4.2.1	Sequence alignment	17
1.4.2.2	Cognate detection	19
1.5	Phylolinguistics	20
1.5.1	Lexicostatistics and Glottochronology	21
1.5.2	Bayesian Phylogenetic Analysis	23
1.5.3	Neighbor-Net Network	25
1.6	Obstacles and Proposed Solutions	25
1.6.1	Lexical Material	26
1.6.2	Linguistically-related Challenges	27
1.6.3	Methodology-related Issues	28
1.6.4	Three Papers and Three Solutions	28
2	Computer—Assisted Language Comparison Workflow	29
2.1	The Hmong-Mien Language Family	30
2.1.1	The Affiliation of the Hmong-Mien Language Family	31

2.1.2	The Internal Structure of the Language Family	31
2.1.2.1	The Evidence Provided by Quantitative Analyses	33
2.2	CALC Framework	34
2.3	Author Contributions	36
2.4	First Paper	36
2.5	Tutorial	51
2.5.1	Code Ocean Capsule	51
2.5.2	Installation Instructions	52
2.5.3	Getting Started	53
2.5.3.1	From Raw Data to Tokenized Data	53
2.5.3.2	From Tokenized Data to Cognate Sets	56
2.5.3.3	From Cognate Sets to Alignments	57
2.5.3.4	From Alignments to Cross-Semantic Cognates	58
2.5.3.5	From Cross-Semantic Cognates to Sound Correspondence Pat- terns	59
2.5.3.6	Validation	61
2.5.4	Conclusion	64
2.6	Retrospective	64
2.6.1	Open Data and the FAIR Principle	65
3	Morpheme Annotation in Phylogenetic Studies	67
3.1	The Benefits of Morpheme Annotation	68
3.2	The Data Structure of a Bayesian Phylogenetic Analysis	70
3.3	Author Contributions	72
3.4	Second Paper	72
3.5	Retrospective	110
3.5.1	Complex Algorithms Require Careful Data Treatment	110
3.5.2	Salient Morpheme Annotation Requires Expert Knowledge	110
3.5.3	Morpheme Annotation Is a Solution for Partial Loanwords	112
3.6	Future Work	112

4	Bayesian Phylogenetic Analysis	113
4.1	The Sino-Tibetan Language Family	113
4.1.1	The Sino-Tibetan paradigm	114
4.1.2	The Tibeto-Burman Paradigm	115
4.2	Author Contributions	116
4.3	Third Paper	116
4.4	The Supplementary Material of the Third Paper	172
4.5	Retrospective	187
4.5.1	Reasons for Data Recompilation	187
4.5.2	Challenges of the Sampling Bias	188
4.5.3	Challenges in Data Interpretation	188
4.5.3.1	Anatomically Modern Human Diaspora	188
4.5.3.2	Language Family Differentiation During the Neolithic Period .	189
4.5.3.3	Debates About the Sino-Tibetan Language Family's Homeland	190
4.6	Future Work	192
5	Discussion and Conclusion	193
5.1	Standardization	193
5.1.1	Graphemes to Phonemes	194
5.1.2	Tokenization	196
5.1.3	Tones	199
5.1.4	Pros and Cons of an Orthographic Profile	202
5.2	Aggregation	202
5.3	Annotation	203
5.3.1	Morpheme Annotation	204
5.3.1.1	The Coordinative Type of Compound Words	204
5.3.1.2	Subordinate Type of Compound Words	205
5.3.1.3	Reduplicated Types of Compound Words	205
5.3.2	Considering the Semantic Shift	205
5.4	Transformation	207

5.4.1	Partial Cognates	207
5.4.2	Cross-Semantic Cognates	208
5.5	Application	209
5.5.1	Distance-based Study	209
5.5.2	Character-based Study	210
5.6	Interpretation	211
5.6.1	The Reports on Bayesian Phylogenetic Analyses	211
5.6.2	Selecting Calibration Dates	211
5.6.3	Bayesian Phylogeny	212
5.7	Conclusion	213

Chapter 1 Introduction

Southeast Asia (SEA) is a geographical region with rich linguistic, cultural, and genetic diversity (Enfield, 2011), and has been the cradle of great ancient civilizations. SEA can be further divided into Mainland Southeast Asia (MSEA) and Insular Southeast Asia (ISEA). MSEA was a crossroads of human migration between Eurasia and East and Southeast Asian islands in prehistoric times. Nevertheless, archaeologists and geneticists have only paid serious attention to this region in recent years—not to mention that diachronic linguistics studies remain in their infancy. This Ph.D. dissertation aims to provide a workflow using a computer-assisted language comparison (CALC) framework in combination with a Bayesian phylolinguistic analysis (Greenhill et al., 2020), which is a state-of-the-art approach to infer a timed language phylogeny. The common obstacles to studying the history of MSEA languages using comparative methods are thus addressed by applying the framework above, which was developed in three projects, as explained in Chapters 2, 3, 4, each of which contains detailed descriptions of the obstacles and the proposed methods.

The remainder of the introduction is arranged as follows. First, Section 1.1 provides a broad description of the geography and of the language families spoken in MSEA. The language family and the language area are two important concepts in historical linguistics. Therefore, Section 1.2 explains language family theory and clarifies the differences between the language family and the language area. Section 1.3 then reviews the typological linguistic features of languages in the MSEA area. The comparative method is the core methodology in historical linguistics. The relatedness of languages is determined by comparing phonology and lexicon. Section 1.4 provides a concise description of the classical comparative method and the computational approaches. Language classification, the etymology of words, and proto-language reconstruction are all part of the comparative method. As an increasing number of computational algorithms have been developed to infer languages families' internal relationships, a field that used to be considered to be part of the comparative method is shifting to a new domain: *phylolinguistics*. The language family tree can be inferred via either the distance-based method or the character-based method. Section 1.5 presents the principles in the two methods. Finally, Section 1.6 summarizes the obstacles that were stated in each of the sections according to three aspects, namely data management, linguistics, and methodology. The significance of this dissertation is highlighted in the same section.

1.1 The Research Territory

The range of MSEA is not clearly defined, and some areas of MSEA even overlap with so-called East Asia (EA). Geographically, MSEA contains present-day southern China (the Yangtze River is treated as the northern borderline of the MSEA), Laos, Thailand, Myanmar, Vietnam, and Cambodia (Enfield and Comrie, 2015). However, the range of MSEA can be expanded

to include modern-day northeast India and southwest China (Sidwell and Jenny, 2021) if non-geographical factors, such as historical ethno-linguistic and political relations and origins, are taken into account. The former is the core of MSEA, and the latter is often referred to as “greater MSEA”. Figure 1.1 presents the map of the greater MSEA area. Of note, the term “greater MSEA” in this dissertation is interchangeable with the term “MSEA language area” due to the shared linguistic features resulting from geographical proximity and language contact.



Figure 1.1: Greater Mainland Southeast Asia language area

The five language families that are spoken within the defined area are Sino-Tibetan, Hmong-Mien, Tai-Kadai, Austronesian, and Austroasiatic. Of these five language families, Sino-Tibetan is the largest language family in the greater MSEA in terms of language varieties and speakers. The speakers live across a wide range of landscapes, including the mountainous area, the high altitude plateau, the lowland, and the coastal line. The Sino-Tibetan language family contains highly differentiated languages, many of which are understudied. Notwithstanding the extensive comparative linguistic studies that have presented the shallow-level subgroups, the internal structure is still highly disputed (Matisoff, 2015; Sagart, 2011b; van Driem, 2015). A state-of-the-art approach called the Bayesian phylolinguistic analysis to infer the date of origin of Sino-Tibetan and its internal structure has been applied (Sagart et al., 2019; Zhang et al., 2020; Zhang et al., 2019).¹ Nevertheless, the large-scale studies have the limitation of having overlooked several geographical regions. The issue of language sampling and the Bayesian phylolinguistic approach will be discussed in Chapters 4 and 5.

The majority of Hmong-Mien speakers are distributed across southern China, Vietnam, Laos, and northern Thailand. Most of the speakers live in the mountainous area, with the exception of

¹A Bayesian phylogenetic study of the Austronesian language family was published prior to the three Bayesian phylogenetic studies of Sino-Tibetan. Nevertheless, the Austronesian languages are mainly spoken outside of the Eurasian continent. Therefore, Sino-Tibetan is the first language family in MSEA to which scholars have applied the Bayesian phylolinguistic method to infer the time depth and the internal structure.

speakers of Kim Mun, which is spoken on the Hǎinán island (海南), and some Iu Mienic dialects that are spoken along the coastline of southern China. Tai-Kadai languages are also spoken in southern China (including Hǎinán island), Vietnam, Laos, and Thailand. The Hmong-Mien and Tai-Kadai people have a long history of co-habitation with Sino-Tibetan speakers in southern China. As a result, comparative linguistic studies show that Hmong-Mien languages contain many Sino-Tibetan and Tai-Kadai loanwords in different time depths. These old loanwords may make it difficult to identify the internal and external relationships among Hmong-Mien languages. The Hmong-Mien language family was thought to be a branch of the Sino-Tibetan family, and this argument continues today. The higher-level structure of the Hmong-Mien language family is bipartite, as the family contains Hmongic and Mienic groups. Phylogenies have been proposed by Ratliff (2010), Chen (2012), and other linguists. However, further details about the shallow subgroups remain subject to debate. Since the writing system of Hmong-Mien languages only developed in modern times, proto-Hmong, proto-Mien, and proto-Hmong-Mien are all reconstructed languages instead of being attested languages.

Given the limited amount (if not the absence) of archaic inscriptions to pinpoint the possible time depths of the language family, a timed phylogeny is thus difficult to derive based solely on references to linguistic data. In addition to the aforementioned issue, there is also a sampling issue when studying Hmong-Mien languages. Most of the linguistic data are focused on the Hmong-Mien languages spoken in China. The studies of Hmong-Mien languages spoken in the neighboring countries are either micro-scale surveys or are of poor quality. The issue of scattered lexical data and the other typological linguistic issues create difficulties when attempting to integrate the Bayesian phylolinguistic approach. We will return to these issues in later chapters.

Tai-Kadai, Austroasiatic, and Austronesian do not constitute the focus of this dissertation. However, the issues mentioned above are shared by these three language families. Therefore, the proposed methods in this dissertation can be applied to the other language families in the MSEA language area.

1.2 Language Family and Language Area

“Language family” describes a set of *genealogically* related language varieties; language varieties are thought to belong to the same family as long as a common *ancestor* can be found. The relatedness (genealogical relationship) among languages is based on common linguistic features, including phonological, morphological, or syntactical features. Historical linguistic research aims to reconstruct a proto-language via attested languages and, ultimately, to reconstruct human pre-history.

Using *language family* or *genealogy* to depict languages’ relatedness stems from the basic model of language differentiation, which appears to be somewhat similar to the basic assumption of population diversification. A common belief is that, when a speaker population separates from its parent group, there is no further contact between the two groups. Thereafter, the language that

the speaker population used changes due to in-situ *innovations*. Such changes occur in various forms, namely sound, morphology, semantics, and syntax. The changes, which accumulate over subsequent generations, resulting in decreasing intelligibility between the language and the parent language. Eventually, the languages are no longer mutually intelligible. At this point, the splinter language is considered to be a different language from the original one. Linguists estimate that the formation of a new language takes around 1,000 years. In this example, the parent language is the *ancestor* of the new language, and is also known as the *proto-language*.

Repeating the process mentioned above would eventually create a large set of related languages. The oldest languages gave rise to a few ancient languages, and the modern languages can be differentiated from some ancient languages. The result can be represented via a tree model with roots, branches, and leaves. Therefore, linguists call a set of related languages a language family, and arrange them as a phylogeny according to their genealogical relationships (*Stammbaum* in German). From top to bottom, the hierarchy is arranged in the order of the language family (the root, or proto-language), groups (interchangeable with branches in this dissertation), subgroups, and language varieties. The level of language varieties is sometimes further separated into dialects and languages. However, differentiating between *dialect* and *language* is the subject of a long-term linguistic debate; linguists have not yet answered this question. Fortunately, the methods proposed in this dissertation do not discriminate between dialects and languages.

We can say that “language differentiation is based on separation” is the simplest presumption for modeling language diversification, but it is somewhat unrealistic. Language contact is known to be another essential mechanism in triggering language change. It is difficult to believe that a given language has never been in contact with any other language over the course of thousands of years. There are certainly reasons for a language to be in contact with other languages, disregarding the speakers’ willingness (or lack thereof); for example, trading, politics, or religion (DeLancey, 2013). These activities do not always involve massive population movements within a short time. For example, the trading of merchandise between two societies using different languages triggers long-term language contact without a large population influx into another region. These prolonged language contacts enable some linguistic features to enter other languages. This phenomenon is also described as *horizontal transmission*, which contrasts with the features inherited from the ancestor languages (also described as *vertical transmission*). Loanwords are the most significant outcomes of language contact.

In summary, language diversification does not always depend on the accumulation of vertical transmissions and in-situ innovations, as it also relies on *contacts*. Furthermore, language contacts often obscure the shared linguistic traits among a group of genealogically related languages. Therefore, it is challenging to classify languages into subgroups of a language family in which language contact occurs frequently.

A language behaves in a similar way to a living being. It can be born, change, differentiate, and

die. Language extinction has also occurred frequently throughout human (pre-)history. The death rate is much faster than is the estimated birth rate (one per 1,000 years). The reasons for initiating language contact are also the reasons that a society would cease to speak the native language and shift to another language if situations became extreme. These extinct languages are suspected of being the missing links between *language isolates* and certain language families. Sadly, many extinct or endangered languages were or are spoken languages; therefore, the documentation is either non-existent or sparse. For example, Kusunda, a language isolate, had been suspected of being a Sino-Tibetan language. Nevertheless, linguists could not find a valid basis for establishing a solid link between Kusunda and the language family.

Even though language differentiation is often compared to population differentiation, the terms *genealogy* and *ancestor* need to be detached from biology, as it is clear that population diversification does not coincide with language differentiation. The shared linguistic features should only determine languages' genealogical relationships. Therefore, the term *ancestor* also has to be discussed within the linguistic realm. Section 1.4 elaborates on historical linguistics, the comparative method, and language classification. Furthermore, the argument for the co-evolution of language and population genetics will be elaborated on in Chapter 5.

Language area (*Sprachbund* in German) defines the *geographical range* within which languages share a set of typological features. In contrast to the definition of a language family in which the shared features are transmitted vertically, the common typological features shared among languages within a defined area are derived from both vertical and horizontal transmissions. Therefore, linguists assert that a language area is defined by geographical, historical ethno-linguistic factors, political relations, and origins (Sidwell and Jenny, 2021).

1.3 The SEA Languages

Summarizing the shared typological features from numerous highly diversified languages is challenging. Nevertheless, Enfield and Comrie (2015, p. 18) were able to summarize a list of ten common features of phonological systems and eleven shared characteristics of morphosyntax-semantic systems. A few points in Enfield and Comrie (*ibid.*, pp. 18–19) are relevant to almost all the topics in this dissertation, including a large vowel system (point (1)), a strict syllable pattern (corresponding to points (2) and (3)), word compounding (corresponding to point (4)), and complex tone systems (point (5)). Examples taken from languages spoken in the greater MSEA language area are provided in each subsection.

1. Vowel systems are large, and show many distinctions.
2. Many more consonants are possible in the initial position than in the final position.
3. There is a preference for one major syllable per word, with many languages featuring minor syllables or pre-syllables in an iambic pattern.

4. No inflectional morphology; note that derivational morphology is widespread, and is sometimes highly productive in the Austroasiatic languages of MSEA.
5. Tone systems are complex (often with around six distinct tones; the tone counts for a language depend on the selected analysis).

We agree with Sidwell and Jenny's (2021) critics that some of the features listed in Enfield and Comrie (2015, pp. 18–19) are uncommon, and only appear in a subset of MSEA languages; some exceptions to such arguments are presented in the corresponding subsections. However, although one could consider the list of common typological features to constitute a generalization about MSEA languages, it is a useful starting point for the study of these languages.

1.3.1 Large Vowel Systems

Consider the following statement by Dryer and Matthew S. (2013, Ch.13): “There are concentrations of larger than average vowel inventories in the interior SEA area and southern China”. The statement appears to be justified based on a quick survey of three languages from Tai-Kadai, Hmong-Mien, and Sino-Tibetan: central Tai (Glottolog: deba1238) has nine monophthongs (Diller et al., 2008), the Southern Guizhou Chuanqiandian variety (Glottolog: cent1394) has ten monophthongs (Chen, 2012), and the Tani language (Glottolog: tani1259) has thirteen monophthongs (including nasalized monophthongs).

The description of the “large vowel system” can be made more precise if Enfield and Comrie (2015) the standards for classifying *large*, *average*, and *small* vowel systems are specified. However, it is difficult to evaluate this statement if the classification criteria are not known. For example, the Ho Nte language (Glottolog: shee1238) has five monophthongs (Chen, 2012), as does Standard Mandarin (Glottolog: mand1415) (Dryer and Matthew S., 2013, Ch.13). Therefore, the question is whether or not five monophthongs can be considered to constitute a large vowel system.

Dryer and Matthew S. (ibid.) suggested criteria for classifying languages according to three categories (bullet points below, also see feature 2A: vowel quality inventories). The classification by Dryer and Matthew S. (ibid.) shows that Ho Nte and Standard Mandarin are exceptions.

- Small: fewer than four vowels.
- Average: between five and six vowels.
- Large: more than seven vowels.

Furthermore, many languages in other parts of the world have large vowel systems. For example, Standard German has eight monophthongs, Standard Italian has seven monophthongs, and Hindi has ten monophthongs. Therefore, stating that MSEA languages have large vowel systems does not appear to be a sufficient description of MSEA languages.

Instead of stating that “MSEA languages have large vowel systems”, describing “MSEA languages as having complex vowel and diphthong systems” might be more appropriate. As Enfield and Comrie (2015) stated, “it is sometimes difficult to determine how many vowels a system has, as there are alternative analyses of features such as diphthongs and phonation splits”. The concept of a “diphthong” is not clearly defined. However, the descriptions in different articles are somewhat similar, with most contending that a diphthong uses two vowels to describe the tongue’s temporal and spatial movement within the oral cavity. There are debates regarding whether a diphthong is counted as one phonetic unit (Catford, 1977), a sequence of two vowels (Ladefoged, 1982), or both: “diphthongs are complex phenomena that show both unity and duality” (Sánchez-Miret, 1998, p. 48).

In MSEA linguistic studies, one also finds that the analysis of diphthongs is tending toward unity (Catford, 1977), duality (Ladefoged, 1982), or a dynamic conception (Sánchez-Miret, 1998); such an analysis is language specific, and depends on the structure of diphthongs. For example, a phonetic study of Chengde Mandarin Chinese pointed out that rising diphthongs, such as *au*, *ai*, and *ei*, as well as monophthongs, could be seen as phonetic units. Falling diphthongs, such as *ia*, *ua*, or *ya*, are sequences of articulations (Zhang and Hu, 2019). Chen (2012, p. 62) stated that there were diphthongs and triphthongs in Hmong-Mien languages, and that the sequence of the vowels represented the tongue’s movement from one place to another, thus producing a sequence of transition sounds.² His view appears to be that the diphthongs in Hmong-Mien are simply vowels that line up in the order of articulations, which appears to be closer to the description provided by Ladefoged (1982).

Applying these viewpoints to the CALC framework makes it even more complex. Assuming that we are studying a topic related to both Hmong-Mien languages and to Chengde Mandarin Chinese, the diphthongs will be tokenized differently. For example, the rising diphthong *au* will be considered to be a single vowel in Chengde Mandarin Chinese, while the same diphthong in the Hmong-Mien language will be divided into *a* in the nucleus and *u* in the coda. However, consistency is the key factor in the CALC framework. Since the analysis of diphthongs is complex and language dependent, adopting one universal model is essential to ensure that the analysis is consistent throughout the entire workflow.

To achieve the consistency of all the computational analyses, the treatment of diphthongs in this paper proceeds according to the cross-linguistic transcription system (CLTS) rather than being based on language-dependent guidelines. A further elaboration on the treatment of vowels and diphthongs will be provided in Chapter 2.

²複元音韻母並不是由兩個或三個獨立的原音組合兩成的，而是發原音時舌位由某一部位向另一部位滑動，聲帶不停止震動，因此在滑動過程中產生了一連串的過度音 (Chen, 2012, p. 62)。

1.3.2 Strict Syllable Pattern

A syllable is a “[b]asic phonetic-phonological unit of a word or of speech that can be identified *intuitively*, but for which there is no uniform linguistic definition” (Bußmann and Bußmann, 2006). Even though there is no clear definition of a syllable, most scholars agree that a syllable is made up of a nucleus.

The most common analysis of a syllable structure entails onset (聲母 *shēngmǔ*) and rime (韻母 *yùnmǔ*). Rime consists of a nucleus and a coda, and the nucleus is usually a vowel. Words in MSEA languages can also be analyzed via the onset-rime syllable structure, but MSEA languages tend to follow a more fine-grained underlying pattern.

The onset can be further analyzed as an initial consonant and a medial approximate. The rime consists of an on-glide position, a vocalic nucleus, and a final consonant (also known as a coda). A lexical tone (聲調 *shēngdiào*) is mainly associated with the nucleus of the rime (Ratliff, 2010). Another analysis of the template is that the medial approximate is merged with the on-glide position, and the entire medial approximate is counted as being part of rime. The assignment of syllable-internal glides to either the medial position (part of the onset) or to the on-glide position (part of the rime) is an obstacle for both synchronic and diachronic phonology (*ibid.*). Therefore, the syllabic template is generally described as having five subdivisions—*initial* (聲母 *shēngmǔ*), *medial* (介音 *jièyīn*), *nucleus* (主要元音 *zhǔyào yuányīn*), *coda* (韻尾 *yùnwěi*), and *tones* (聲調 *shēngdiào*) (Baxter, 1992)—instead of six segments. The syllable templates are presented in Table 1.1. The onset can be further analyzed as an initial consonant and a medial approximate. The rime consists of an on-glide position, a vocalic nucleus. The five-subdivision template is mainly used in this dissertation. The template is called the IMNCT template in the CALC framework, and assists in aligning the phonetic strings cross-linguistically among MSEA languages. Subsequently, it detects the sound correspondences among various languages based on the alignments (see Chapter 2).

	(C)C	{j/w/l}		{i/ü}	T	(V)V	(C)
	onset				rime		
Analysis 1	initial	medial		(on-glide)	Tone	nucleus	coda
Analysis 2	initial			medial	Tone	nucleus	coda

Table 1.1: The syllable template for MSEA languages. The details of the two analyses can be found in Ratliff (2010) and Baxter (1992), respectively.

Words in MSEA tonal languages have two essential subdivisions, namely nucleus and tone. For example, the word 椅 “chair” in Mandarin Chinese, which is spelled *yǐ*, is pronounced as *i* with a

falling-rising tone. The tone cannot be ignored because it determines the semantic meaning.³

Ratliff (2010, p. 10) stated that the “Hmong-Mien language is the *typical* southeast Asian type”. Therefore, we use the word “nose” in Hmongic and Mienic languages to illustrate the sounds corresponding to the IMNCT template (see Table 1.2). Note that all the elements in IMNCT are presented as being on an equal hierarchical layer because the hierarchical structure is not an essential feature for sequence alignments, cognate judgments, or Bayesian phylolinguistics.

Doculect	Subgroup	Form	Initial	Medial	Nucleus	Coda	Tone
Eastern Baheng	Hmongic	mpjau ³¹	mp	j	au	—	31
Zao Min	Meinic	tɕaŋ ⁵³	tɕ	—	a	ŋ	53

Table 1.2: The word “nose” in the Eastern Baheng and Zao Min language variety and the alignment with the IMNCT template.

1.3.2.1 Major and minor syllables

Enfield and Comrie (2015) stated that MSEA languages prefer to have one major syllable per word, with many languages featuring minor syllables or pre-syllables. A standard structure is *Cə.CVC*, in which *Cə* is the minor syllable and *CVC* is the major syllable. Alternatively, Michaud (2012) defined the minor syllable as a simple consonant plus an optional nucleus, and stated that the nucleus did not need to be a schwa.

Sinitic languages do not possess the phenomenon of major and minor syllables since most of the morphemes in Sinitic languages are monosyllabic. However, many Hmong-Mien words have pre-syllables (prefixes). Chen (2012, p. 142) stated that the tones of pre-syllables were all neutral, but that the tone values were heavily influenced by the major syllable. Hence, his data set usually annotates the tones in a prefix using the form ^{0x}, such as *ta⁰²na³¹* “person”, *tə⁰²lo⁵⁵* “old (adj.)”, *qa⁰³qaj³⁵* “star”. When the prefix’s tone is not neutral, that is, the tone is not annotated using the form ^{0x}, this means that the prefix’s rime is assimilated by the major syllable. Hence, the tone value of the prefix is no longer neutral; for example, *qi¹³pli¹³* “wildcat” or *ta³⁵la⁴⁴* “rabbit”. The pre-syllables can be divided into lexical and grammatical pre-syllables (Strecker, 2021). For example, the *lo⁵⁵* in the word *tə⁰²lo⁵⁵* “old (adj.)” is a noun: “old”. The *tə⁰²* prefix has changed the noun into an adjective.

Linguists have argued about whether a minor syllable is a complete syllable. Matisoff (1973) labeled the phenomenon of major and minor syllables as *sesquisyllable*, which means one and a half syllables. Other linguists consider words with minor and major syllables to be disyllabic. In the CALC framework, the computer programs perform the shallow-level analysis. These programs treat the minor syllable as a whole syllable because both the segments in each of the major and minor syllables can be fitted into the IMNCT syllabic template. Table 1.3 shows the correspondence between phonemes and the syllable template.

³Mandarin Chinese is 椅子 *yǐzi*, but *yǐ* on its own is sufficient for the word “chair”.

Doculect	Form	Prefix					+	Root				
		I	M	N	C	T		I	M	N	C	T
Central Guizhou Chuanqiandian	tə ⁰² lo ⁵⁵	t	—	ə	—	⁰²	+	l	—	o	—	⁵⁵

Table 1.3: The phonemes in the word “old (adj.)” correspond to the IMNCT template. The plus symbol is the boundary between two syllables, and can also be seen as the morpheme boundary.

1.3.3 Compound Words

Enfield and Comrie (2015) stated that “MSEA languages lacked inflectional morphology, but that derivational morphology was widespread in Austroasiatic languages”. Unfortunately, this statement is not entirely accurate. It is true that inflection is not the primary type of morphological process in MSEA languages, and is largely absent from the languages spoken in the core MSEA language area. However, exceptions can be found in the greater MSEA language area.

Inflection means modifying a lexeme to fit into a particular position within a sentence; for example, marking the gender, the number, or the tense. As an example, the Duhumbi language (Glottolog: chug1252), a Sino-Tibetan language spoken in Arunachal Pradesh, uses suffixes to alter personal pronouns from the singular to the dual or to the plural. Table 1.4, which is taken from the book *Grammar of Duhumbi* (Bodt, 2020, p. 107), provides evidence that inflection can be found in the languages spoken in the greater MSEA language area.

	First person	Second person	Third person
singular	ga	naŋ	woj (wuj)
dual	gaziŋ	naziŋ	waziŋ
plural	gar (galu)	nar (nalu)	war (walu)
anaphoric	—	—	bi
egophoric	raŋ (laŋ)	—	—

Table 1.4: Example taken from Bodt (2020).

Although Enfield and Comrie (2015)’s generalization regarding MSEA languages’ lack of inflections is not without exceptions, the second half of the statement, which states that derivation morphology is widespread among MSEA languages, is accurate. In fact, compound words and derivations are the two main strategies that are used to enrich MSEA languages’ lexical inventories. Prefixes in Hmong-Mien languages can be used to distinguish non-living from living objects. For example, the prefix *qɔ*³⁵ in Western Xiangxi (Glottolog: west2430) is used to describe non-living objects, and the prefix *ta*³⁵ denotes animals. Prefixes can also be used to distinguish human beings. For example, the words “father” and “husband” in the Central Guizhou Chuanqiandian variety (Glottolog: nort2749) are *pa*¹³ and *qa*⁰³*pa*¹³. The *qa*⁰³ is a prefix that is used to differentiate between human characters, such as between “father” and “husband” (Chen, 2012, p. 141). Chen (ibid.) provided a table of the prefixes in 18 Hmong-Mien languages.

Word compounding is possibly the most significant typological feature of MSEA languages that springs to mind. A *compound word* is a lexical item that is produced by combining two or

more free forms (Bloomfield, 1933). Free forms usually refer to morphemes. However, in this dissertation, the concept of free forms is equivalent to monosyllabic words, and is interchangeable with morphemes. Take Standard Mandarin Chinese as an example: The word 太陽 “sun” can be expressed via two monosyllabic words 太 *tài* “greatest” and 陽 *yáng* “sun, masculine, or bright”. Analyzing a compound word requires taking the grammatical features and the semantics of each free form into account. The arrangement of the forms and the relationships among them should also be considered.

In MSEA languages, the majority of monosyllabic words belong to the categories of nouns, verbs, or both. Therefore, if we discuss the types of compound words based on their grammatical features, there are only four different combinations: noun-noun, verb-noun, noun-verb, and verb-verb. However, this classification does not indicate the “weight” of each compound. For example, if we put two nouns 花 *huā* “flower” and 草 *cǎo* “grass” together, 花草 *huā cǎo* is a noun-noun compound word, a collective term for flower and grass, but 草花 *cǎo huā* is also a noun-noun compound word, a certain flower genre. The meanings of *huā cǎo* and *cǎo huā* change when the order of the two monosyllabic words change. Moreover, both parts of *huā cǎo* contribute to the semantic meaning, but the *huā* determines the semantic meaning of *cǎohuā*. The relationships of the two monosyllabic words in *huācǎo* and *cǎo huā* are different. The relationship between the two parts is not fixed. We can insert additional words into *huā cǎo*; for example, 奇花異草 *qí huā yì cǎo* “rare species of flower and grass” or 拈花惹草 *niān huā rě cǎo* “being flirty”. The word *cǎo* in *cǎo huā* is used to modify the word *huā*. We cannot insert any word between *cǎo* and *huā*; therefore, the two parts of *cǎo huā* are linked more strongly than are *huā cǎo*.

Linguists have summarized the compound words in Modern Chinese according to the four categories of coordinate, subordinate, reduplicated, and stump (Kratochvíl 1970, pp. 73–82; Cui et al. 2018). These four classifications can also be applied to other MSEA languages (Enfield and Comrie, 2015). Linguists may use different terms for the four categories, or may re-group the four categories. For example, Chen (2012, pp. 147–149) suggested that the Hmong-Mien compound words should be categorized according to the categories of *copulative* (聯合式), *endocentric* (修飾式), *complement* (補充型), and *subject-predicate* (陳述式). The copulative is the same as the coordinative compound type, while the other three are the subordinate compound type. Hmong-Mien languages also have reduplicated compounds (Máo, 2004).

A compound word is not only an areal feature, but is also a phenomenon that is shared world-wide. For example, compound words can be found in languages in Africa, South America, and India. More details about the types of compound words are provided in Chapter 3.

1.3.4 Complex Tone Systems

Most of the world’s languages make use of some type of pitch or intonation. Intonation generally applies to sentences, to contrast questions, and statements, or implies an ending (Maddieson, 2013). The commonly cited definition of tones refers to the use of pitch patterns to differenti-

ate the core semantic meanings of words (Yip, 2002, p. 1). Tables 1.5 and 1.6 present the tone systems in Mandarin Chinese and in Vietnamese.⁴ One can see that the semantic meanings of *ma* have changed due to tone variations. It is estimated that 60% to 70% of the world’s natural languages are tonal languages according to the definition by Yip (ibid.), and there appears to be a higher concentration of tonal languages in the MSEA area compared to other parts of the world (Maddieson, 2013). Hence, tones are a significant typological feature of the MSEA languages. However, this is not to say that we cannot find atonal languages in the MSEA language area. In the core MSEA language area, Khmer (Glottolog: cent1989) and Mnong (Glottolog: mnon1259) are not considered to be tonal languages because Khmer uses vowel height to change registers (Brunelle and Kirby, 2016, p. 194), and Mnong is an atonal language (ibid., p. 196). We also find several languages in the greater MSEA area: Puroik (Remsangpuia, 2008, p. 90) and Garo (Burling, 1961) are two atonal Sino-Tibetan languages that are spoken in northeast India.⁵ In addition, Japhug (Glottolog: japh1234) and Amdo Tibetan (Glottolog: amdo1237) are atonal languages that are spoken in China.

Pinyin	tones	Chinese example (Meaning in English)
mā	first tone; high tone	媽 (mother)
má	second tone; rising tone	麻 (使... 麻痺 to numb)
mǎ	third tone; falling and rising tone	馬 (horse)
mà	fourth tone; falling tone	罵 (罵... to blame)

Table 1.5: The tonal markings in Mandarin Chinese Pinyin.

Vietnamese	tones	Meaning in English
ma	Thanh Ngang; mid-level tone	a ghost
mà	Thanh Huyền; low falling tone	that
má	Thanh Sắc; high rising tone	cheek
mả	Thanh Hôi; low rising tone	tomb
mã	Thanh Ngã; high broken tone	horse
mạ	Thanh Nặng; heavy tone	a new born rice plant

Table 1.6: The tonal markings in the modern-day Vietnamese writing system.

Several MSEA tonal languages appear to be complex in terms of the number of contrasts. According to Maddieson’s (2009) system—atonal, simple (less than or equal to two tones), complex (more than two tones)—standard Vietnamese, White Hmong, and standard Thai are classified as having complex tonal systems because they all have more than five tones. However, simply counting the number of tones does not make a tonal system in MSEA languages “complex”. There are tonal languages in other parts of the world that can be classified as having complex tone systems, namely Triqui, (Glottolog: triq1251) which has eight tones, while Attié (Glottolog: atti1239) has

⁴The neutral tone is not shown in the table because it was gradually replaced by other tones in Mandarin Chinese (Taiwan).

⁵The Puroik language was once recognized as a tonal language by Sun (1993). However, Remsangpuia (2008) stated that no minimal pair could be found to show that the Puroik language was a tonal language.

six tones, and Mixtec (Glottolog: mixt1427) has three. Therefore, the measurement of “complexity” should be further defined.

Maddieson (2009) stated that measuring the complexity of a tone system should take both the number of contrasts and the tone sandhi rules into account. For example, Hmong-Mien languages generally have more tone sandhi rules than Mandarin Chinese. However, using these two measurements appears to be insufficient. Imagine a scenario in which language A has five tones and only two tone sandhi rules, but language B has four tones and three tone sandhi rules: Which language should be considered to be more “complex”? Ratliff classified tones as being of the Asian type or the African type; according to this classification, Asian-type tones are usually bounded by segments, while African type-tones tend to spread over the segments. Given these circumstances, which system is more complex? In addition to the above arguments, Brunelle and Kirby (2016, p. 199) provided several considerations in the current studies of SEA tonation.

Tones are a difficult topic in historical linguistics. Apart from the difficulty stemming from linguistics, the inconsistent tonal marking of the lexical material adds another layer of difficulty. As the materials in this dissertation were taken from Sino-Tibetan and Hmong-Mien languages, the tones are of the Asian type according to Ratliff’s definition. Many languages attach diacritics to the nuclear vowel to reflect the lexical tone’s position in the segment (for example, the Mandarin Chinese Pīnyīn system and the modern Vietnamese writing systems; also see the examples provided in Enfield and Comrie (2015) for Thai tones). Other tonal marking strategies, such as numbers (see the examples in Table 1.2), symbols, or alphabets (Heimbach, 1969) can also be found in many lexical data sets. Due to there being various approaches to expressing tones, combining sparse lexical data sets into one large data set has been a challenge for computational algorithms. Therefore, standardizing tonal markings is an important step in preparing MSEA tonal languages’ lexical data sets.

Due to the special attributes of tones, computation programs that can successfully incorporate the tone information into a computational comparative analysis have not yet been developed. In addition to the linguistic attributes, a large-scale and well-curated data set that can be used as training data or study material is also missing. Furthermore, computational linguists have not yet thought about cross-linguistic analyses that involve different types of tones (such as grammatical tones or lexical tones, bounded types or spreading types, and so on), or ways of annotating them. As a result, tones are usually ignored in the computational analysis phase.

This dissertation touches upon the topic of standardizing lexical tone annotation in MSEA in Chapter 2 and in Chapter 5 in the hope that the method can assist linguists to generate some finely curated data sets as the testing material for future computational programs.

1.4 The Comparative Method

1.4.1 Classical Approach

The comparative method is a set of principles (Campbell, 2013, pp. 6–7) for reconstructing proto-languages based on the patterns of phonological and semantic correspondences between two (or more) attested languages. The core of the comparative method is based on the concept of sound change, which is possibly the most rigorously studied area in historical linguistics. Sound change has two categories, namely sporadic and regular (Campbell, 1999, p. 17). As is indicated in the names, sporadic sound change only occurs in a small portion of the words in a language, and regular sound change is a uniform change in the vocabulary. In the two categories, regular sound change is a presumption that is derived from the hypothesis that language differentiation is based on separation. Assuming that the voiced bilabial stop **b* existed in a proto-language, and that the daughter language A changed it into voiceless bilabial stop *p* while language B retained the original voiced bilabial stop *b*, we should be able to find *b* in a word in language B whenever *p* appears in a word in language A (Campbell, 2013, pp. 6–7).

The steps in the principles summarized by Durie and Ross (1996) and Jäger (2019) are now presented to assist readers to navigate the dissertation:

1. Assume relatedness among languages based on diagnostic linguistic evidence (Durie and Ross, 1996, pp. 6, 48). The evidence stems from various sources, mainly phonology.
2. Collect and identify homologous words, also known as cognates. For example, *daughter* in English and *tochter* in German are cognates, and the nearest common ancestor is **dokhter* in the proto-Germanic language (ibid., p. 7)
3. Derive sound correspondence from cognate sets. Irregular cognate sets are not involved in the process of summarizing sound correspondence (ibid.).
4. Reconstruct the family's proto-languages, including proto-sound and proto-morphemes (ibid.).
5. Discover and reconstruct more diagnostic evidence (ibid., pp. 7, 48).
 - (a) Group languages according to the innovations, including phonological, lexical, semantic, morphological, and morphosyntactic features (ibid., p. 7).
 - (b) Tabulate the innovation to arrive at an internal diversification of the family; that is, a language phylogeny (ibid.).
6. Construct an etymological dictionary by tracing borrowings, semantic changes, and so forth for the lexicon of the family (or of one language in the family). (ibid.)

As in every other scientific field, scholars identify some phenomena and then establish a hypothesis. The first step in the comparative method also begins by presuming that a set of languages can be traced back to a common ancestor. Building on this presumption, linguists begin to collect lexical items in each language variety.

The second and third steps are to search for homologous words among the sampled languages and, subsequently, to summarize the *sound correspondences* in the identified cognates. Table 1.7 compares the words in eleven Hmong-Mien languages using three different glosses (for example, “to know”, “molar tooth”, and “hundred”) (Ratliff, 2010). Ratliff (ibid.) found nine “to know” words in eleven Hmong-Mien languages that were all related to each other. These words are called *cognates*. However, the author only found three “molar tooth” words among the eleven Hmong-Mien languages. The “-” in the table means either the other eight languages do not have the word “molar tooth”, or that the words in the eight languages are not cognates of the other three words. The same example shows that the sound *p* in the initial position (represented as *p*-) in languages one to 10 corresponds to the sound *b* (represented as *b*-) in language 11. The *p*-~*b*- correspondence set is thus inferred among the eleven Hmong-Mien languages.

	1	2	3	4	5	6	7	8	9	10	11
to know *pei	pu ¹	-	pau ¹	pɔ ^{1a}	-	pe ¹	pi ¹	pei ¹	pei ¹	pəi ¹	pɛi ¹
molar tooth *pæ	-	pæ ²	pua ¹	-	-	-	-	-	-	-	ba ¹
hundred *pæk	pa ⁵	pa ⁵	pua ⁵	pa ^{5a}	pi ^c	pa ⁵	pe ⁵	pɛ ⁷	pe ⁷	pɛ ⁷	ba ⁷

Table 1.7: An example of a sound correspondence set summarized from eleven Hmong-Mien languages. The “-” represents missing values.

Linguists frequently alternate between the second and third steps because sound correspondence sets and cognate sets are mutually corroborated. In addition, the two sets are influenced by the languages and by the lexical items that are sampled.

The sound correspondence and cognate sets are two important areas of evidence for reconstructing the proto-sounds and the proto-morphemes. As shown in Table 1.7, the proto-sound in the onset position *p- is reconstructed through the correspondence set. The proto-Hmong-Mien words *pei “to know”, *pæ “molar tooth”, and *pæk “hundred” are derived from the combination of p-~b- in the onset position and the other sound correspondence sets in the rime position.

The study of the internal relationship in the language family also relies on the cognates and the sound correspondences (step 5). The phonology and word cognacy provide the first evidence of language subgroups. Linguists will then search for other shared linguistic features to support the subgrouping. Although the language subgroups do not depend solely on the number of cognates, the cognates and the tendency for regular sound correspondences are given more weight than are the other linguistic features.

Reconstructing proto-languages and language phylogeny can also provide linguistic evidence to infer the prehistorical lifestyles of the speakers’ ancestors. For example, Ratliff (ibid.) recon-

structed paddy-rice-related proto-words, and indicated that the Hmong-Mien speakers originated in southern China where the weather allows people to grow paddy rice. Sagart et al. (2019) pointed out that linguists have reconstructed words such as millet, horse, pig, and so on in the proto-Sino-Tibetan language. Therefore, the Sino-Tibetan speakers originated in the mid-Yellow River area. Combining the evidence from historical linguistics and archaeology can also help to determine the time depth of the language family and the expansion process throughout history.

The last stage in the comparative method is the presentation of an etymological dictionary to show the changes in words and sounds from the proto-language to the daughter languages, as well as the mechanisms that triggered the changes, including borrowings and semantic changes.

In Sinology, the method used to identify the original Chinese character (考本字 *kǎo běn zì*) (Mei, 1995) is a unique application of the comparative method. Chinese characters may be used in different Sinitic languages with different pronunciations. As the language changes over time, the Sinitic languages may retain the pronunciation but replace the original word with other words. Therefore, Sinologists use the outcome of the comparative method plus three different approaches—looking for words (覓字 *mì zì*), searching for sound (尋音 *xún yīn*), and discussing the original meaning (探義 *tàn yì*)—to identify the real word (Mei, 1995; Yang, 1999). In theory, the outcome of this method can assist with the morpheme annotation in Sinitic languages. However, the sounds of the Chinese characters have been changing over thousands of years. Identifying the sound of the characters in different historical periods is challenging. Therefore, the research outcomes of this method are rare and controversial, not to mention that the method has only been applied to a handful of Sinitic languages.

The goal of the comparative method is to provide evidence of language changes; the above principles apply to phonology, to semantics, and to morphosyntactic levels. However, language changes are not limited to these levels: For example, English grammar is different in Middle English and in Modern English. Nonetheless, the above principles do not consider grammar extensively when reconstructing proto-languages. Due to different research purposes, some linguists may have different breakdowns of the stages. This dissertation does not touch on any topic related to grammatical changes; therefore, the principles suggested by Durie and Ross (1996) are a good fit for the study.

Overall, the comparative method is a labor-intensive and time-consuming method. It requires linguists working on different language varieties within the same language family to provide a wide range of examples to supplement the cognate identification and the sound correspondence. Depending on the range of language samples and lexical items that are being included in the project, the outcome of cognate sets and sound correspondence sets may be different. Therefore, the above principles are not sequential. Linguists constantly alternate between one step and another, with different languages or different words being included or excluded each time. In addition, various factors influence the accuracy of cognate decisions, such as mistaking sporadic

sound changes for regular sound changes, loanwords, insufficient samples of languages, and missing cognates due to semantic shifts. Therefore, linguists need to rely on their experience and to be flexible when making decisions about cognates. As the demand for quantitative research in the scientific world and the volume of linguistic data has been increasing in recent decades, the amount of data that needs to be handled is humanly impossible. Comparative methods need to change to a large-scale and efficient orientation. Thus, computational historical linguistics was born to boost the efficiency of historical linguistic studies.

1.4.2 Computational Approach

Some steps in the principles (Durie and Ross, 1996) are quite mechanical and can be replaced by computational algorithms; for example, sequence alignment, cognate detection, and phylogenetic inference (Jäger, 2019).

1.4.2.1 Sequence alignment

Sequence alignment is an essential stage prior to making cognate judgments. Linguists align the words' phonetic strings and then determine the similarities among the groups of words. This process takes place internally in the linguists' minds. Consider Table 1.7 as an example. The representation in linguists' minds is somewhat similar to Table 1.8. Linguists divide the phonetic sequences according to the template in their minds: The categories that are usually used are onset and rime. They align the phonetic strings position by position. The similarity and sound correspondences are determined based on the outcome of the alignments.

Doculect	Value	Onset	Rime	Tone
1	pu ¹	p	u	1
3	pau ¹	p	au	1
4	pɔ ^{1a}	p	ɔ	1a
6	pe ¹	p	e	1
7	pɪ ¹	p	ɪ	1
8	pei ¹	p	ei	1
9	pei ¹	p	ei	1
10	pəi ¹	p	əi	1
11	pei ¹	p	ɛi	1

Table 1.8: Take the word “to know” as an example.

Computer programs make use of the same method. The strings are first tokenized, and are then aligned. There are three issues to consider during the steps. First, computers do not know how to tokenize phonetic strings as segments. For example, the sound *ei* is a diphthong, but computers will consider *e* and *i* to be two independent sounds unless otherwise specified. Second, the alignment does not always follow the order of the strings; the prosodic structure needs to be taken into consideration to avoid a consonant being confused with a vowel. Third, how is the “similarity” after alignment quantified? Imagine that we have cross-linguistic data for three words *gæp*, *ⁿkap*, and *kep*. How do we reach a result as in Table 1.9? Is *ⁿk* closer to *k* than to *g*, or vice

versa? The same question also applies to vowels.

Doculect	Value	c	v	c
A	gæp	g	æ	p
B	ⁿ kap	ⁿ k	a	p
C	kɛp	k	ɛ	p

Table 1.9: The desired tokenization of the three words. The c and v in the header are abbreviations for “consonant” and “vowel”. This is the simplest syllable structure analysis.

Computational linguists suggested using the idea of sound class, as first proposed by Dolgopolsky (Dolgopolsky, 1964; Dolgopolsky, 1986), who categorized sounds according to ten types; a sound is assumed to have a higher probability of changing to another sound within the same category than to a different category (Dolgopolsky, 1964; Dolgopolsky, 1986; List, 2012b)⁶. Different sound classes have since been developed, with the most representative sound class systems being the Sound-Class-Based Phonetic Alignment (SCA, List, 2012b) and the Automated Similarity Judgment Program (ASJP, Wichmann et al., 2016). Take ASJP as an example: The three imaginary words can be converted into *gEp*, *kEp*, and *kEp*, respectively.⁷ Subsequently, we can tokenize the three converted strings and align them according to the consonants and the vowels.

The two alignment approaches are pairwise and multiple alignment methods. The *pairwise sequence alignment* (PSA) only compares two strings at a time. In our example, three pairs are needed to be aligned and compared: (*gæp*, ⁿ*kap*), (*gæp*, *kɛp*), and (ⁿ*kap*, *kɛp*). The *multiple sequence alignment* (MSA) aligns all the strings at the same time. Although the MSA algorithms appear to be more efficient, aligning multiple sequences simultaneously requires sophisticated algorithms. The multiple sequence alignment is already well incorporated in bioinformatics research due to high-throughput sequencing technology. The Python library *LingPy* also implements the MSA method, but there is room for improvement to the MSA method in historical linguistic research (see the discussion in List et al., 2018).

As the phonetic strings become longer, there may be more than one possibility for aligning two words. For example, the medial position in Hmong-Mien languages sometimes disappears from the phonetic strings. The Needleman-Wunsch (NW) algorithm is commonly used for optimal global alignment. When aligning two sequences, each matched sound segment (phonetic symbol) increases the similarity by 1 if matched; otherwise the similarity is reduced by 1. The NW algorithm does not treat all types of mismatches as equal weights, but introduces the gap-opening penalty to address one base deletion, as well as the gap-extension penalty to manage continuous deletions.

The results of the sequence alignments can then be used to infer similarity among words. The

⁶In phonology, sounds in the same category share common features

⁷Conversion according to the indication on CLTS.

most straightforward measure is the Levenshtein distance, also known as the edit distance, which calculates how many times a string needs to be edited in order to become another string. For example, changing the word *gap* to the word *kɛp* in our example requires two edits: (1) change *g* to *k*, and (2) change *æ* to *ɛ*. Another interesting method is the *pointwise mutual information* (PMI) method.

$$PMI(a, b) \doteq \log \frac{s(a, b)}{q(a)q(b)}$$

The PMI formula is shown above. The $q(a)$ is the probability of a appearing in a string, and $q(b)$ represents the same meaning; $s(a, b)$ is the probability that a is aligned with b in the correct alignment (Jäger, 2019). The similarity score for the alignment pair is the sum of all the segments' PMIs.

There are more ways to determine the similarity among word pairs and to subsequently derive the machinery cognates. Two cognate detection methods are particularly highlighted in the following paragraph — **Sound Class Alignment** (SCA) (List, 2012b) and **LexStat** (List, 2012a) —because these two methods are implemented in *LingPy* (List et al., 2019), the core Python library in the workflow.

1.4.2.2 Cognate detection

Several cognate detection methods are available at present. The cognate detection methods provided in *LingPy* were used in the workflow because this is the only library that provides the option of detecting words' *partial cognacy*. As mentioned previously, word compounding is a major word-formation mechanism in MSEA languages; thus, discussing the cognates at the morpheme level is more appropriate than is discussing the word cognates.

The input data format is a multilingual word list and the process includes four steps to produce machine cognates (List, 2012a):

1. Sequence conversion
2. Scoring-scheme creation
3. Similarity (distance) calculation
4. Sequence clustering

SCA and LexStat are the two main models that are used to evaluate the similarity between two sequences. In the sequence clustering process, *LingPy* used the Infomap algorithm.

The scoring scheme of sound correspondences is based on a permutation method to compare the attested distribution of residue pairs in the phonetic alignment analyses of a given data set to the expected distribution (ibid.). The alignment process was conducted using the SCA method.

Overall, the SCA and the LexStat methods use the same strategy for clustering, but the distances for the SCA model are computed with the assistance of the SCA alignment method, and the similarity scores for the LexStat model are obtained from previously identified regular sound correspondences.

The outcomes of the cognate judgments can be used to determine the languages' relatedness on the lexical level. Subsequently, languages are classified into subgroups based on their relatedness. In recent years, historical linguistic studies have focused extensively on reconstructing and interpreting the internal structure of a language family. In particular, the amount of studies is increasing rapidly as a result of the integration of computational algorithms and the possibility of combining inputs from multiple disciplines in language phylogeny reconstruction. Therefore, phylolinguistics is discussed in an independent section, even though reconstructing language trees is part of the comparative method (Durie and Ross, 1996, pp. 6–7).

1.5 Phylolinguistics

The comparative method does not dictate how a “language tree” should be displayed, as long as the languages' relatedness is shown; the best option would be to provide some evidence to indicate the degree of relatedness. The representation can be as simple as a table that lists shared cognates in language pairs, or the table could accompany a hand-drawn tree. The majority of language trees present the language subgroups in hierarchy charts or dendrograms: The node is the common ancestor, and the edges link two or more related languages (also known as sister languages) to a common ancestor. The dendrogram, or hierarchy structure, helps people to identify the relationships among languages quickly.

Apart from linguists' hand drawings, two categories of statistical methods are used to infer a language phylogeny from the cognate sets, namely the distance-based and the character-based methods.

Both the distance-based and the character-based methods are based on cross-linguistic cognate sets. However, the questions pertain to the sufficient number of words needed to determine relatedness, and how the words should be selected in order to represent a language. The average number of basic vocabulary words, which are words that are used for communication in daily life in a language, is about 2,000 to 3,000 words. The actual number in the lexical inventory of a language is much greater than 3,000 words; for example, there are 600,000 entries in the latest version of the Oxford English Dictionary. If we were to compare all the words across all the sampled languages in order to construct language subgroups, the comparison would never be complete.

In statistics, sampling methods are developed based on the idea of using a portion of samples to represent the population, such as stratified sampling, random sampling, clustered sampling, or systematic sampling. Sampling can also be applied to historical linguistics; for example, the core

vocabulary that is used for lexicostatistics and glottochronology.

1.5.1 Lexicostatistics and Glottochronology

Lexicostatistics and glottochronology (詞源統計分析法) are two *distance-based* methods. Both approaches were developed based on the idea that two closely related languages would share more cognates than would two distantly related languages. Both methods are based on the assumption that core vocabulary changes follow a constant rate, which is analogous to the carbon dating technique in archaeology.

The difference between the two methods is that lexicostatistics merely computes the languages' genealogical relationships, while glottochronology not only computes the relatedness among languages but also estimates the amount of time needed for languages to differentiate. One can also say that glottochronology is a subtype of lexicostatistics, since glottochronology is based on lexicostatistics.

In order to compare languages' relatedness systematically, Morris Swadesh suggested a list of 215 words as the test list. He and his team members found that words used in daily life—such as words describing human body parts, the natural environment and phenomena, or numbers—had a constant rate of change. Moreover, these words are thought to be resistant to borrowing or lending with regard to other languages. Based on his observations, he created a word list, which is also known as the *Swadesh list*, based on the following criteria:

- Universal: words exist in all the sampled languages
- Non-cultural: due to the assumption of a constant rate of change
- Unambiguous: words are easily identifiable and can be expressed in simple terms
- Known to all the speakers: words that are not used only by a specific section of people in a society.

Not including the cultural words in the core vocabulary means that culturally specific words will have a different rate of change in different languages. Swadesh also suggested removing a word when there were too many languages that did not possess the lexical item. He subsequently changed the word list a number times, and the final version contained 100 words. There have been other attempts to create word lists based on the similar guidelines. The ASJP database even went a step further and extracted only 40 words from the Swadesh lists, stating that the list of 40 words was sufficient for inferring language phylogeny. The words that were included in the word list are called the *core vocabulary*.

Linguists create cross-linguistic data sets based on the core vocabulary to compare languages' relationships. The lexicostatic method iterates through all the language pairs and computes the *similarity* of each language pair. The outcomes are eventually shown as a pairwise matrix (see Example 1.10). The result is a symmetrical matrix; the numbers are the amount of cognates that are shared between the two languages to represent the similarity. Some scholars tend to only

provide the upper triangle of the matrix. The similarity score does not necessary have a fixed range, but it is usual to present the similarity score as a percentage. The way to read the numbers is the higher the number, the greater the similarity between the two languages.

	Language A	Language B	Language C
Language A	100	90	40
Language B	90	100	20
Language C	40	20	100

Table 1.10: An example of lexicostatistics assuming that a linguist selected 100 lexical items from the Swadesh list.

Linguists can use the similarity method to draw a tree: For example, Chen (2012) presented the tree in Figure 1.2 based on the similarity matrix that he computed.

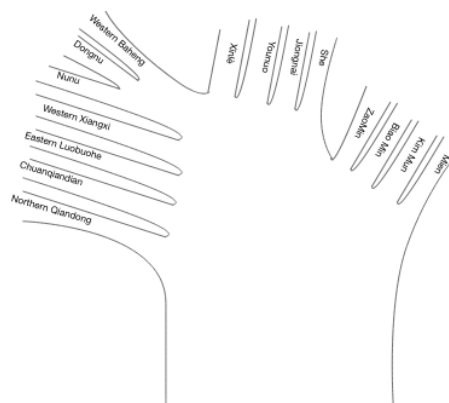


Figure 1.2: The Hmong-Mien phylogeny constructed based on the lexicostatistics in Chen (2012).

The pairwise similarity matrix can be used to inspect the relatedness when the set of languages is small. As more languages are included, the similarity matrix may be too large for the human eye to discern the pattern of relatedness. Instead, one can measure the distance between two languages in the cognate sets and then use distance-based phylogeny reconstruction methods to infer a tree-like structure from the distance matrix; this is called glottochronology.

These three formulas are commonly used: The first and third can be applied directly to the similarity matrix, while the second is the Jaccard index (or Jaccard distance), which is only used when we can access the cognate sets directly. The distance ranges from 0 to 1, with 0 indicating that the languages are identical, and 1 meaning that the two languages are completely unrelated.

$$D(A, B) = 1 - \frac{A \cap B}{\text{Total lexical items}} J(A, B) = \frac{A \cap B}{A \cup B} D(A, B) = -\log s(A, B)$$

To reconstruct the phylogeny from a distance matrix, one can use either a Neighbor-Join (NJ) tree (Saitou and Nei, 1987) or the Unweighted Pair Group Method with Arithmetic Mean (UP-GMA) tree (Rédei, 2008). Both tree types use nodes and edges to show the relationships among

languages, with the edge lengths (also called the branch length in a tree-like structure) representing the differentiation time.

The difference between the two algorithms is that the NJ tree generates an unrooted tree, which means that no evolution directions are inferred. Figure 1.3, shows the conversion of the similarity matrix in Chen (2012) into a distance matrix; an NJ algorithm was used to infer the tree. A further interpretation of this figure can be found in Chapter 2.

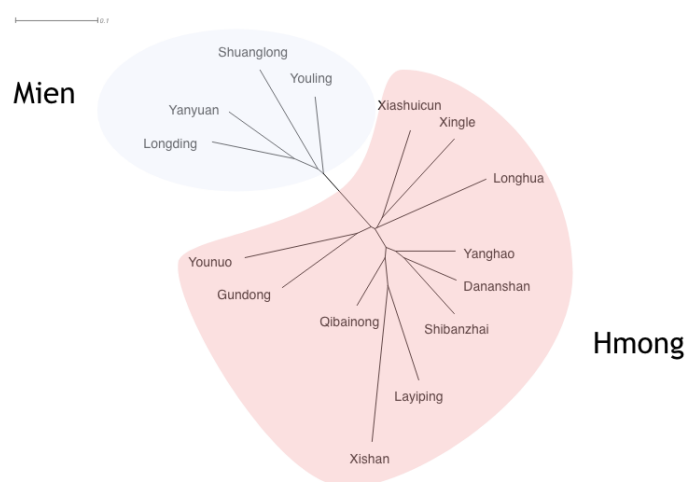


Figure 1.3: The Neighbor-Join tree inferred from the similarity matrix in Chen (2012)

One can assign an outgroup in the tree in order for the tree to indicate directional phylogeny. There are algorithms that infer the root of the NJ tree automatically but, as these are all additional inferences, the individual branch lengths may be skewed by the algorithm. If the evolution direction is desired, UPGMA is a good choice to infer a rooted tree without an additional layer of inference.

1.5.2 Bayesian Phylogenetic Analysis

Alternatively, the phylogeny can be inferred directly from the cognate sets. Each language is represented as a binary vector. The phylogeny can be inferred via maximum likelihood or via maximum parsimony algorithms. A state-of-the-art, character-based phylogeny inference is the Bayesian phylolinguistic analysis, which was used in the third project in this dissertation (see Chapter 4).

Previous methods were based on the presumption of a strict molecular clock, according to which the lineages developed at a constant rate. Swadesh proposed that the rate of lexical change was fixed in the core vocabulary. He noticed that culture-related words might have a different rate of change; thus, linguists also know that the presumption of lexicostatistics is unrealistic. However, until Bayesian phylogenetic analysis was developed, there was no other way to incorporate the presumption that each language had its own rate of evolution.

A Bayesian phylogenetic analysis is a data-driven method, which means that it requires large-

scale data sets. Since it does not necessarily follow the presumption of the strict molecular clock, the lexical items can be expanded to include more words, including culture-related words, in the lexical data sets. Cross-linguistic data sets in MSEA languages usually have their own collections of words; for example, Chen (2012) provided a Hmong-Mien lexical data set containing 888 lexical items, and Huáng and Dài (1992) presented a large-scale cross-linguistic Sino-Tibetan word list with 1,800 glosses (lexical items). These word lists usually include religion-related words (such as religious items for praying), agricultural equipment and supplies, or animals and plants that are found in the MSEA area.

A Bayesian phylogenetic analysis has different models for inferring the transition state of characters. The models estimate the cognates' state from "present" to "absent", and vice versa. One can assume that a cognate can appear and disappear from a language with the same likelihood. Users can also make the assumption that, once a cognate appears in the language, it is unlikely to disappear.

To infer the lineage evolution rate, users can select either a strict molecular clock or a relaxed molecular clock. The relaxed molecular clock shows that each language can have a different evolution period.

In glottochronology, languages could only be "born". However, the reality is that a language can become extinct after a certain point in time. Therefore, a Bayesian phylogenetic analysis allows users to presume that the sampled languages can be born at any point in time, and that the sampled languages can also die at some point. This is called the birth-death model.

The greatest advantage of a Bayesian phylogenetic analysis is that it can incorporate calibrations from archaeology or genetic studies. If we know that a language became extinct X years ago, we can input the known information into the model. However, a limitation of Bayesian phylogeny is due to the calibration points. If the languages or the speaker populations are all well documented throughout history, it is not difficult to provide the calibrations for the differentiation points. Nevertheless, we often do not have the luxury of detailed documentation to support the calibrations. When faced with such a difficulty, one has to estimate a tree height, which is a presumption regarding when the oldest proto-language in the language family may have been born.

The Bayesian phylogenetic analysis relies on the presumptions that users determine and run a huge number of iterations; each iteration generates a tree. Following the extensive computations, the algorithm summarized from the sampled trees generates a consensus tree and assigns each internal node a probability; that is, how many of the sampled trees show this diversification event.

A Bayesian phylogenetic analysis has a few limitations apart from the calibration points sometimes being difficult to find. These issues include word compounding and semantic shift. Linguists usually judge cognates based on entire words. We will return to these challenges in later chapters (see Chapter 4).

1.5.3 Neighbor-Net Network

The basic assumption of a language phylogeny is separation-based language differentiation. Therefore, in theory, loanwords are excluded from the data sets. However, MSEA languages, particularly the core MSEA languages, have had long-term and intense language contact with each other. This prolonged language contact resulted in words entering (that is, being borrowed by) MSEA languages in different periods. A possibility is that old loans may be mistaken for native words because they have been integrated into the languages so well. Therefore, we cannot guarantee that the lexical data sets are entirely “loanword-free”. It is concerning that the undetected loanwords may introduce some noise or conflicting signals into the distance matrix. The phylogenetic algorithm introduced above only reports the optimal tree; therefore, this noise is not shown in the phylogeny.

The splits decomposition algorithm was proposed to evaluate the degree of conflict in a given distance matrix (Bandelt and Dress, 1992). Bryant and Moulton (2004) developed the Neighbor-Net network algorithm based on the splits decomposition algorithm to present the conflicting signals in a given matrix via a split graph. Because the end product of the Neighbor-Net network does not necessarily form a tree-like structure, it is used as a means of evaluating whether a given distance matrix can be represented more appropriately as a tree or as a network. The Neighbor-Net network algorithm is not only an evaluation tool: It can also be applied to explain the alternative evolution processes (Fitch, 1997; Gray et al., 2010).

The alternative evolution process in historical linguistics refers to contact-induced language change. If the web-like structure is clearly evident in the Neighbor-Net network, this means that intensive language contact is occurring among the selected languages. If not, the languages can be arranged as a tree-like structure. We used a distance matrix that we converted from the similarity matrix presented in Chen (2012) as the input for the Neighbor-Net network algorithm (Huson, 1998). Figure 1.4 shows that there is a web-like structure among the Hmong languages, as well as between the Hmong and Mien languages. The Delta score and the Q-residual are two indications of the “tree-ness” of the splits graph.

1.6 Obstacles and Proposed Solutions

Even though historical linguistics has been developing since the late eighteenth century, applying comparative linguistics to the study of MSEA languages is still in its infancy. Many languages in MSEA are understudied. Even for a widely discussed language family such as Sino-Tibetan, the shallow-level subgroups and the relationships among subgroups have not yet attained consensus. A few obstacles were observed when applying computational historical linguistic methods to the study of MSEA languages. The challenges stemmed from various aspects, including study materials, linguistics, and methodology. These issues, which are the focus of three separate studies, are presented in the following paragraph. Other difficulties of which we were aware but on which

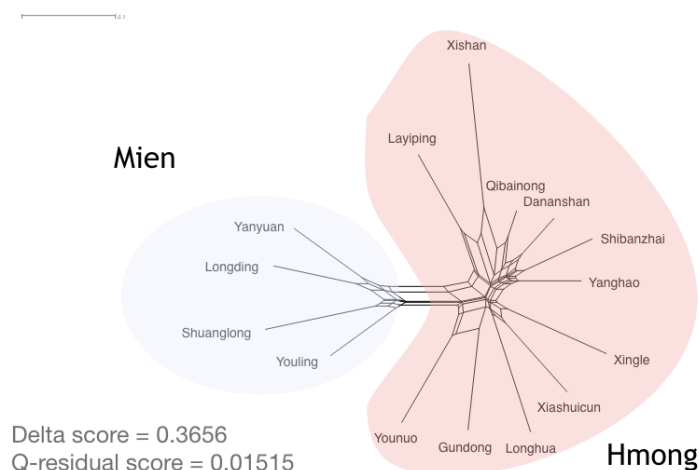


Figure 1.4: The Neighbor-Net network inferred from the similarity matrix in Chen (2012).

we were not able to work will be explored in the discussion section.

1.6.1 Lexical Material

There are three major difficulties when collecting lexical material: The first is having a balanced-sample cross-linguistic data set, the second is inconsistent lexical data formats, and the third is ambiguous metadata documentation.

A balanced sample means that each of the subgroups in a language family is represented by at least one language. The best scenario is that all the subgroups are represented by an equal number of varieties, and no individual subgroup is overly represented. However, a balanced sample is not feasible. There are large and small language subgroups in terms of the number of identified varieties in one language family. In addition, some languages have been well documented and studied carefully throughout history. A typical example is the Sinitic languages. It is easy to collect a large amount of lexical material from such languages. However, there are several understudied languages or language subgroups. A classic example is the languages spoken in northeast India. The data for such cases are sparse and usually have a limited amount of glosses. Since obtaining a balanced sample for a cross-linguistic lexical data set is impossible, merging individual data sets into a large cross-linguistic data set is essential.

The inconsistent data formats increase the difficulty of expanding or merging the existing cross-linguistic data sets efficiently. Take Hmong-Mien languages as an example. Because the field work focuses on Hmongic and Mienic languages in China, the Hmong-Mien languages outside of China are often understudied. To present a clear picture of the Hmong-Mien phylogeny requires merging the data sets of Hmong-Mien languages within and outside of China. Nevertheless, data sets are often presented in different formats, which slows down the merging of data sets.

Lastly, Swadesh indicated that the glosses in a word list needed to be represented using short

words with additional descriptions in brackets (Swadesh, 1964). Some cross-linguistic data sets follow this recommendation; however, this regulation creates more confusion than clarity. Take a commonly used gloss 香 *xiāng* as an example (Chen, 2012). The word means aromatic or fragrant, or can be a noun meaning the incense used when praying. Both meanings have the same Chinese gloss, and the only way to differentiate between the two glosses is to check the order of the glosses on a word list: One is placed in the category of a human's sense of smell, and the other is placed in the category of culture. Similar issues include the inconsistent labeling of language varieties, idiosyncratic tonal annotation, and the fact that linguists tend to delete affixations arbitrarily. To resolve the issues related to data presentation, we made use of the cross-linguistic data format to standardize the data format across different data sets.

1.6.2 Linguistically-related Challenges

The linguistically related challenges are cognate detection and sound correspondences. First, the identification of cognates in MSEA languages is not straightforward because the customary practice in the comparative method is to view a word as an entire unit and to determine these units' relatedness. The outcome is usually binary, which means that word A and word B are or are not cognates. Nevertheless, compound words can be seen as either one unit or as multiple units depending on the semantic meanings. Each part of the compound words could experience a different evolution. As a result, the compound words may only be partially related (Hill and List, 2017). Assuming that compound words are complete entities when identifying cognacy may simplify the process of language change significantly.

As an alternative, linguists tend to judge the cognacy among compound words by placing more weight on a certain part than on the other parts; for example, emphasizing a morpheme that the linguist considers to be salient. The degree of a morpheme's salience can be ascribed to the semantic meaning, the function of the words, or it could simply be dynamic because the other words in the same gloss all share the same morpheme. Some linguists even make this internal analysis transparent by presenting the morphemes in the data set and omitting the raw forms (e.g., Ratliff, 2010). However, this approach cannot preserve the completeness of the words' raw forms, which limits the re-usability of the data.

To overcome the issues arising due to compound words, List (2016) proposed the concept of *partial cognacy* to address compound words. Chapter 2 presents a workflow to assist linguists to make *partial cognate* judgments while preserving the words' raw forms. The data annotation project described in Chapter 3 also departed from the idea to enhance the transparency of morpheme cognacy.

The sound correspondence is summarized based on the cognate sets. The larger the lexical data set involved, the more accurate the sound correspondence. The perfect scenario is that linguists are able to apply the comparative method to survey all the languages in one language family: The sound correspondence sets will be the most accurate in this scenario. However, it is humanly

impossible to work on such a project because the number of languages and the volume of lexical items exceeds human capacity. Therefore, computational programs are designed to assist linguists by summarizing the sound correspondence sets quickly.

1.6.3 Methodology-related Issues

The first critical issue is the alignment of phonetic sequences, which determines the accuracy of cognate decisions and sound correspondences. The sequence alignment requires a guiding template; thus, we made use of the tendency toward a strict syllable structure in MSEA languages. Even though there is a strict syllable structure to follow, there are still different levels of analysis. The sequences are tokenized differently depending on the chosen analysis. We understand that there is no single template that can attain universal agreement among linguists. Hence, we valued consistency over other factors at this stage. The phonetic sequence may undergo a certain degree of modification in order to be tokenized and to fit into the chosen template.

The second issue also pertains to compound words. We introduced partial cognates as a treatment for words that are only partially related. At present, there is no suitable algorithm to reconstruct dated phylogeny from the binary vectors that are converted directly from the partial cognates. Therefore, we developed an annotation scheme and introduced four different conversion methods to transform the partial cognate sets and to pipe with the distance-based phylogenetic algorithms.

The third issue is the integration of multidisciplinary knowledge into a historical linguistic study. Archaeology and archaeogenetic studies in the MSEA area are just beginning to flourish, and there are still several areas awaiting further study. A Bayesian phylolinguistic analysis can integrate several lines of evidence to support a dated phylogeny. However, we need to examine the results more carefully in order to not fall into the trap of overly interpreting the dated phylogeny.

1.6.4 Three Papers and Three Solutions

The aforementioned issues cannot be fully addressed with a single Ph.D. dissertation. Therefore, this cumulative dissertation is an initiative that centers on the MSEA language area and addresses three fundamental obstacles in three distinct projects:

- Develop a workflow to standardize the format for lexical data digitization,
- propose an annotation scheme to enhance the transparency of studying the etymology of compound words, and
- address the importance of combining shreds of evidence from multiple disciplines to infer a timed phylogeny.

Finally, the FAIR principle has been considered to be a cornerstone of research in many disciplines. The benefits of open data are demonstrated in Chapter 2, 3, and 4. We include a brief introduction to the FAIR principles in Chapter 2 and a retrospective on the topic of open data in the same chapter.

Chapter 2 Computer—Assisted Language Comparison Workflow

This chapter introduces the workflow that was constructed based on the CALC framework to convert the given lexical material from its raw form into a standardized format. The standardization operates on both the data and the metadata. The enhancement of the metadata documentation is a response to the FAIR data principle. The workflow is the first step toward efficiently and accurately merging multiple lexical data sets into a large-scale data set, with the aim of creating a data set that can be flexibly expanded to incorporate other data sets.

The research is centered on the concept of establishing an MSEA-specific Bayesian phylolinguistic analysis. Our analysis required well-curated data as its foundation. However, the lexical data to which we had access were in a number of different file formats. We created a data standardization phase for each freshly acquired data set to enhance the comparability of the lexical data sets, as this is essential for enhancing the precision of a computational analysis. In addition, we substituted some repetitive procedures in traditional comparative methodologies with computer programs to increase the productivity of our research. Therefore, we utilized existing Python tools to automate the repetitive tasks. Our workflow also enables the post-editing of computer algorithm outputs by specialists.

To transform the acquired lexical data to be the same set of standards, we made use of well-established databases to normalize our data sets. The preprocessing phase involved transforming the various formats into a desired template, standardizing the usages of phonetic symbols, ensuring the lexical items' definitions were equivalent across all the data sets, and so forth. We have a collection of tools and data bases to assist linguists to complete these tasks more rapidly and consistently; we will elaborate on these databases in the sections that follow.

Following standardization, the lexical data will enhance not only the precision of a computational analysis, but also the comparability of diverse data sets. Standardized content can be combined with other data sets to increase the quantity of lexical items or to improve the quality of phonetic transcriptions. Take the Hmong-Mien language families as an example; linguists have concentrated on documenting the Hmong-Mien languages spoken in China, but lexical databases for the Hmong-Mien languages spoken in peripheral areas typically contain fewer vocabulary items. These incomplete aspects cannot be ignored if we are to study the history of the spread of Hmong-Mien languages. Therefore, we implemented a procedure to standardize the Hmong-Mien lexical data sets and to elevate each of them to a level at which they could be combined. We will demonstrate the procedure used to merge the data sets in the last section of this chapter. We encountered further challenges in the course of our research on the Hmong-Mien language

families; we will detail the obstacles other than data formats in later chapters.

Not only does our method transform data sets into the same format, it also generates an initial set of cognate judgments based on the standardized results. We provide computational cognate sets because observing the patterns of phonological sequences and then correcting them via linguists' knowledge is a repeated effort. It is the most time-consuming task. In addition, this phase is characterized by the greatest degree of obscurity, as few linguists are accustomed to explaining their rationale for classifying words into cognate sets. Our technique divides the words into sets of cognates based on phonetic tokenization. The methodology may not generate the most precise cognate sets, but it clarifies why computer systems infer cognate sets in a particular manner.

Our workflow can produce good results with the vast majority of data, but we encourage specialists to exert extra effort to make the process more efficient and the outcome more accurate. We explain our workflow in detail in the published paper (Wu et al., 2020) and the later section; it is our hope that the tools we provide in this work can assist other researchers to prepare their data.

2.1 The Hmong-Mien Language Family

The Hmong-Mien languages provide useful examples for establishing our workflow. As stated previously, lexical data sets are provided in a variety of forms and sizes. In addition, conventions for naming languages are not standardized, and the same language is frequently labeled differently using the geographical name or language subgroups in various sources. Moreover, the phonological inventory of each Hmong-Mien language is inconsistent in the data resources. Hence, the Hmong-Mien transcripts are frequently provided in linguists' personal transcripts rather than as a consensus of transcriptions. The underlying explanation for these inconsistencies may be that the language family is understudied and the data sets are fragmentary. In view of this, we provide a summary of the disputes in the field of Hmong-Mien language studies in an effort to increase comprehension of the language family. In addition, the lexical data sets that were processed in our workflow have been published online in response to the movement for open data.

The Hmong-Mien language family (also known as 苗瑤語系 *Miáo-Yáo* in China) is spoken by the *Miáo* and *Yáo* people, two ethnic groups in SEA who are native to China, northern Thailand, Laos, and Vietnam (Figure 2.1). Apart from SEA, a diaspora of Miao and Yao speakers migrated to North America in recent centuries. The language family comprises 39 language varieties according to Glottolog (v4.4). Linguists treat the labels *Hmong-Mien* and *Miáo-Yáo* as being interchangeable; however, we argue that they are two different terms. *Yáo* people all speak Mienic languages, while *Miao* people speak either Hmongic or Mienic languages. Using *Miáo-Yáo* to describe the language family actually confuses ethnology with linguistics. Therefore, the terms *Miáo* and *Yáo* in this dissertation refer to the populations of speakers. The labels *Hmong*, *Mien*, and *Hmong-Mien* are used to describe the language varieties and the language family.



Figure 2.1: Hmong-Mien language distributions

2.1.1 The Affiliation of the Hmong-Mien Language Family

Hmong-Mien languages were included in the Sino-Tibetan language family (Chen, 1996; Chen, 2012; Klaproth, 1823; Leyden, 1808; Li, 1937). The languages are still included in the Sino-Tibetan language family by some linguists at present. For example, Li (1937) placed the Hmong-Mien languages in the same group as Sinitic and Tai languages; he later promoted the position of Hmong-Mien languages in the Sino-Tibetan phylogeny to a higher layer as a language group (Li, 1973). Chen (2012, p. 8) presented 166 word cognates among Sino-Tibetan and Hmong-Mien languages.¹ Nevertheless, many linguists argue that the similarity between ST and HM words may be attributable to loanwords that entered Hmong-Mien languages in different periods, or may be mere coincidence (Gong, 2006).

At present, the majority of historical linguists consider HM languages to be a different language family from the Sino-Tibetan language family.

2.1.2 The Internal Structure of the Language Family

Linguists have not yet proposed a detailed topology for Hmong-Mien language phylogeny, despite decades of historical linguistic studies. Currently, linguists only agree that the higher-level structure of the language family is a bipartite structure involving the Hmongic and the Mienic

¹The author stated that he identified 166 words in “The comparative study of Sinitic, Miao and Yao dialects” (Chen, 2002), and presented the 166 words again in his work in (Chen, 2012).

groups (Chen, 1984; Li, 1937; Li, 1973; Ratliff, 2010; Strecker, 1987).

As shown in Figure 2.2, Strecker (1987) suggested that the Hmong-Mien language family consisted of seven groups: Hmongic, Baheng, Hm Nai, Jiongnai, Younuo, Mienic, and Ho Nte (also known as She). This shows that linguists have identified Qiangdong, Xiangxi, and Chuanqiandian as Hmongic languages since 1987. The Mienic group contains Mien-Kim, Zao Min, and Biao Min.

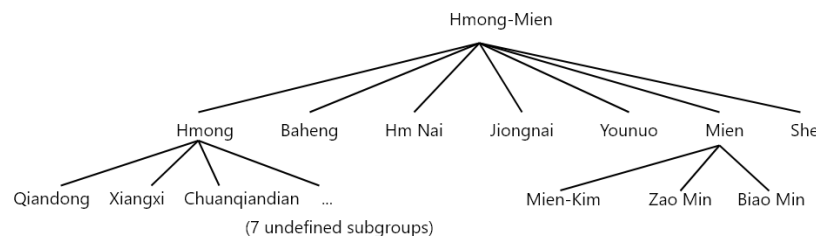


Figure 2.2: Hmong-Mien language phylogeny by Strecker (1987)

The bipartite structure was also supported by Chen (1984). However, Chen placed the Ho Nte language under the Mienic branch rather than under the Hmongic branch.

Wang and Mao (1995) also proposed a bipartite structure for the Hmong-Mien language family, with four different layers. The first layer contains the Hmongic and Mienic groups. The Hmongic group contains the Hmong subgroup, Bunu, Baheng-Younuo, and Jiongnai-She. As Strecker (1987) suggested, Xiangxi, Qiangdong, and Chuanqiandian are placed under the Hmong subgroup. However, the authors did not specify the internal structure of Mienic languages.

Finally, Ratliff (2010) reconstructed the proto-Hmong, proto-Mien, and proto-HM words from eleven HM language varieties and proposed a HM language phylogeny. The phylogeny is a bipartite structure involving the Hmongic and the Mienic groups. The Hmongic group can be further divided into five subgroups: Pahang (Glottolog: paha1256), Jiongnai/Ho Nte, East Hmongic (Glottolog: east2369), North Hmongic (Glottolog: nort2748), and West Hmongic (Glottolog: west2430). The Mienic group contains three subgroups: Zao Min, Biao Min, and Mien-Min. The phylogeny proposed by Ratliff is shown in Figure 2.3.

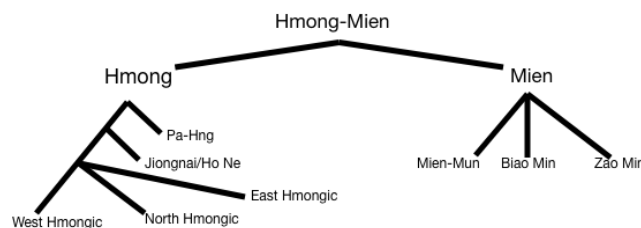


Figure 2.3: The internal structure of the Hmong-Mien language phylogeny proposed by Ratliff (2010)

Linguists once proposed that the language family had been a tripartite structure in the past.

The discussion was led by Wang and Mao (1995); the two linguists addressed the ambiguous position of Ho Nte (also known as She). In their tripartite phylogeny, Hmongic, Mienic, and Ho Nte (also known as She) branched from the proto-Hmong-Mien at the same time. Hmong, Bunu, Baheng, and Jiongnai are placed under the Hmongic group. In addition, the Younuo language is placed in the Baheng subgroup. Nevertheless, (ibid.) did not specify further relationships among Mien, Zao Min, Kim Mun, and Bio Min.

2.1.2.1 The Evidence Provided by Quantitative Analyses

To date, few linguistic studies in the field of Hmong-Mien language studies have made use of quantitative analyses. In the following paragraphs, we discuss two studies by Deng and Wang (2003) and by Chen (2012), which applied quantitative analyses.

Deng and Wang (2003) used the glottochronology (詞源統計分析法) method, and compared the relatedness of language pairs using the percentage of shared cognates among 111 core vocabulary words. The authors inferred the phylogeny via NJ methods, and then used the mid-point method to root their tree. The results can be found in their article Deng and Wang (ibid., p. 6), showing that Mienic and Hmongic are clearly two groups. The She language is closer to the Hmongic languages than it is to Mienic languages.

Chen (2012) offered a pairwise matrix based on cognate annotation. However, he did not use any further mathematical methods to construct a phylogeny using the pairwise matrix. The classification of the Hmong-Mien languages spoken in China based on dialect intelligibility is presented in the book. Nevertheless, a few criticisms of the figure should be mentioned. First, the number of lexical items provided in the book exceeds the number that the author claimed should be used for classification. We also could not identify the concepts that were used in the intelligibility test. Second, the classification is illustrated by hand, which makes some of the lower branches' classifications somewhat confusing. We suspect that there were not many user-friendly programs available to assist linguists to deduce their pairwise matrices with regard to the representation of languages' classifications. For those who want to reconstruct phylogeny using their data today, there are a few possibilities. We will revisit this topic in later chapters.

While it is encouraging to see academics provide data based on quantitative research rather than on subjective judgments, it is a pity that they only shared the pairwise matrix and not the cognate sets. If scholars were to share/publish their data sets, this would be advantageous in terms of increasing the replicability of their studies and increasing data re-use.

We could examine Deng and Wang's or Chen's cognate annotations if the cognate sets were made available to the general audience. Being unable to inspect the raw data may induce some skepticism; for example, we cannot identify the lexical items that were used to make the cognate judgments in Deng and Wang (2003) or the intelligibility test in Chen (2012). As a result, we cannot discount the possibility that the authors cherry-picked concepts to construct the classifications in the form that they expected. In addition, without the cognate sets, we could not use their data to

evaluate their classifications using various phylogenetic techniques, such as maximum likelihood or Bayesian phylogeny.

In addition to the issue of low replicability, we encountered the problem of re-usability. We needed to begin our cognate annotation from scratch without their cognate judgments; we also could not compare our cognate judgments to theirs. Consider the time that could have been saved by using the external data sets. We do not intend to criticize Deng and Wang (2003); in fact, academics failing to share their data with the public appears to be a common problem. Many academics are concerned about this custom because some academic subjects are developing more slowly as a result.

The issue of scientific studies not releasing data gave rise to the open data movement. At present, there are websites that provide data archiving services. As the use of open data was an advantage in this dissertation, we will introduce the idea in the section that follows. The significance of open data will be mentioned several times in this dissertation.

2.2 CALC Framework

The CALC framework has been proven to be an efficient strategy in the domains of comparative linguistics (Wu et al., 2020), typology (Dryer and Matthew S., 2013), and other linguistic fields. The framework has also helped to shed light on existing theories (Chechuro et al., 2021; Rzymiski et al., 2020; Sagart et al., 2019). The CALC framework is designed to incorporate experts’ knowledge and computing power to achieve the four main goals of consistency, flexibility, efficiency, and accuracy. In recent years, several web applications (List, 2017; List et al., 2017), programming packages (*LingPy* by List et al. 2019, *CLDFbench* by Forkel and List 2020), and databases (Concepticon by List et al. 2016; List et al. 2021b, CLTS by List et al. 2021a) have been developed within the framework to assist in the process of data inspection, curation, analysis, and management. In addition, in response to the FAIR principles of Findability, Accessibility, Interoperability, and Re-usability (Wilkinson et al., 2016), a standardized format (Forkel et al., 2018) has been designed to facilitate data management and sharing.

The cross-linguistic data format (CLDF) has been developed by the research group “Computer-Assisted Language Comparison (CALC)” at the Max Planck Institute in an attempt to increase the FAIRness of linguistic data. The CLDF is a sustainable ecosystem consisting of a set of general guidelines, as well as three databases and software packages accompanying these databases. The databases and software packages are revised and updated regularly.

The CLDF format not only standardizes linguistic data, but also highlights the importance of describing the data in a systematic way. The general guidelines include using text-based formats (*.csv*, *.tsv*) and a narrow table format (see the proposed table format in Forkel et al. 2018). The structure of the table should be as follows: Each type of information should be coded in a separate column, and each column should correspond to only one type of information. Each cell in the table

should only contain one value. Each row in the table should correspond to a single entry. Each entry should have a unique identifier. Linguistic data should be separated from metadata and linked via identification numbers (IDs).

The metadata following the CLDF standard should be described in a separate table, and should follow the same requirements as for the linguistic data. The CLDF metadata must include information about the languages in the sample, the explanation of the glosses and, most importantly, an orthography profile for standardizing phonetic symbols. Users are also free to add additional information that is important for their data set. For the readers' convenience, we provide an example of a CLDF data set in the supplementary materials (S1). A detailed description of and instructions for the example data are also provided.

Metadata curation in CLDF relies on two databases, namely Glottolog (Hammarström et al., 2020) and Concepticon (List et al., 2020b). The regularization of phonetic symbols and segmentation in the CLDF workflow depends on CLTS (List et al., 2021a). All three databases aim to create reference catalogs for various sources.

Glottolog aims to provide comprehensive information about languages across the world. The database gathers metadata about languages, such as the language families, the language status, and relevant studies. It serves a similar purpose to ISO 639 (that is, the language's ISO code) as it disambiguates the labels for language variants by assigning a unique identification number to a language that had previously been assigned different names in the existing literature. For example, the White Hmong language in the World Loanword Database (WOLD) is the same language variety as the Hmong Daw language in Ratliff (2010); therefore, both varieties are assigned the same glottocode, *hmong1333*.

Concepticon was established based on the same philosophy in an attempt to unify the annotation of lexical entries (also known as glosses or concepts) across different sources. For example, an entry labeled “mortar” may refer to two different concepts, the first being “a bowl used to crush and grind ingredients with a pestle”, and the second being a “paste made of a mixture of a binder (cement, plaster, or lime), sand, and water used in masonry to make bricks, stones, etc. stick together”. If no additional information is provided in the data set, it is impossible to establish which of the two meanings was intended. Linking the data to the Concepticon database allows one to disambiguate such cases easily and to ensure the correct reading in each particular case.

The Concepticon database is constantly growing, and currently features about 3,800 commonly used concepts (also known as the Concepticon concepts) taken from various concept lists, including the Swadesh list and its variants (Holman et al., 2008; Swadesh, 1955; Swadesh, 1964), as well as large concept lists that contain over 800 unique glosses (Chen, 2012; Huáng and Dài, 1992). The database provides a unique identification number for each concept, as well as a detailed description and additional information. Mapping the vocabularies in a data set onto the Concepticon database can be seen as transforming the implicit glosses into explicit concepts.

This step also creates a link to numerous existing data sets that are also linked to Concepticon, thus significantly increasing the potential benefits one may extract from a single data set.

The CLTS database features 15 different transcription data sets and provides catalogs of five different transcription systems. The Broad International Phonetic Alphabet (BIPA), a universal transcription system that is regularly updated by experts, is used as a reference system for the other transcription systems and data sets. All the data sets in the CLTS database have the same structure: The graphemes in the data sets are matched to the BIPA graphemes. For users' convenience, the website displays the summary of 5,371 conventional graphemes and their BIPA counterparts as of 2021. The CLTS database is accompanied by a Python Application Programming Interface (API). Its applications include (but are not limited to) looking up the BIPA counterpart of a given grapheme and transliterating a given orthography in another transcription system.

2.3 Author Contributions

Mei-Shin Wu (MSW), Nathan W. Hill (NWH), and Johann-Mattis List (JML) initiated the study. MSW, NWH, JML, and Timotheus A. Bodt (TAB) drafted the workflow. MSW and JML implemented the workflow. Nathanael E. Schweikhard (NES) wrote the glossary. TAB, NWH, and NES tested the workflow on different datasets. MSW and JML wrote the accompanying tutorial. MSW and JML wrote the first manuscript. NES, NWH, and TAB helped in revising the manuscript. All authors agree with the final version of the manuscript. The article is published in *Journal of Open Humanities Data*.² The code and data are available in the online repository.³

2.4 First Paper

The paper appeared in the *Journal of Open Humanities Data* in 2020 (Wu et al., 2020).

²DOI:<https://doi.org/10.5334/johd.12>

³<https://github.com/lingpy/workflow-paper>

RESEARCH PAPER

Computer-Assisted Language Comparison: State of the Art

Mei-Shin Wu¹, Nathanael E. Schweikhard¹, Timotheus A. Bodt², Nathan W. Hill² and Johann-Mattis List¹¹ Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, DE² SOAS, University of London, London, UKCorresponding author: Mei-Shin Wu (wu@shh.mpg.de)

Historical language comparison opens windows onto a human past, long before the availability of written records. Since traditional language comparison within the framework of the comparative method is largely based on manual data comparison, requiring the meticulous sifting through dictionaries, word lists, and grammars, the framework is difficult to apply, especially in times where more and more data have become available in digital form. Unfortunately, it is not possible to simply automate the process of historical language comparison, not only because computational solutions lag behind human judgments in historical linguistics, but also because they lack the flexibility that would allow them to integrate various types of information from various kinds of sources. A more promising approach is to integrate computational and classical approaches within a *computer-assisted framework*, “neither completely computer-driven nor ignorant of the assistance computers afford” [1, p. 4]. In this paper, we will illustrate what we consider the current state of the art of computer-assisted language comparison by presenting a workflow that starts with raw data and leads up to a stage where sound correspondence patterns across multiple languages have been identified and can be readily presented, inspected, and discussed. We illustrate this workflow with the help of a newly prepared dataset on Hmong-Mien languages. Our illustration is accompanied by Python code and instructions on how to use additional web-based tools we developed so that users can apply our workflow for their own purposes.

Keywords: computer-assisted; language comparison; historical linguistics; Hmong-Mien language family

1 Introduction

There are few disciplines in the humanities that show the impact of quantitative, computer-based methods as strongly as historical linguistics. While individual scholarship and intuition had played a major role for a long time, with only minimal attempts to formalize or automate the painstaking methodology, the last twenty years have seen a rapid increase in quantitative applications. Quantitative approaches are reflected in the proposal of new algorithms that automate what was formerly done by inspection alone [2], in the publication of large cross-linguistic databases that allow for a data-driven investigation of linguistic diversity [3], and in numerous publications in which the new methods are used to tackle concrete questions on the history of the world’s languages (for recent examples, see [4, 5]).

While it is true that – due to increasing amounts of data – the classical methods are reaching their practical limits, it is also true that computer applications are still far from being able to replace experts’ experience and

intuition, especially in those cases where data are sparse (as they are still for many language families). If computers cannot replace experts and experts do not have enough time to analyze the massive amounts of data, a new framework is needed, neither completely computer-driven nor ignorant of the assistance computers provide. Current machine translation systems, for example, are efficient and consistent, but they are by no means accurate, and no one would use them in place of a trained expert. Trained experts, on the other hand, do not necessarily work consistently and efficiently. In order to enhance both the quality of machine translation and the efficiency and consistency of human translation, a new paradigm of computer-assisted translation has emerged [6].

Following the idea of computer-assisted frameworks in translation and biology, scholars have begun to propose frameworks for *computer-assisted language comparison* (CALC), in which the flexibility and intuition of human experts is combined with the efficiency and consistency of computational approaches. In this study, we want to

introduce what we consider the state of the art¹ in this endeavor, and describe a workflow that starts from raw, cross-linguistic data. These raw data are then consistently lifted to the level of an etymologically annotated dataset, using advanced algorithms for historical language comparison along with interactive tools for data annotation and curation.

2 A workflow for computer-assisted language comparison

Our workflow consists of five stages, as shown in **Figure 1**. It starts from *raw data* (tabular data from field-work notes or data published in books and articles) which we re-organize and re-format in such a way that the data can be automatically processed (Step 1). Once we have lifted the data to this stage, we can infer sets of etymologically related words (*cognate sets*) (Step 2). In this first stage, we only infer cognates inside the same *meaning slot*. That means that all cognate words have the same meaning in their respective languages. Once this has been done, we *align* all cognate words *phonetically* (Step 3). Since we only infer cognate words that have the same meaning in Step 2, we now use a new method to infer cognates *across meanings* by employing the information in the aligned cognate sets (Step 4). Finally, in Step 5, we employ a recently proposed method for the detection of correspondence patterns [7] in order to infer sound correspondences across the languages in our sample.

Our workflow is strictly *computer-assisted*, and by no means solely *computer-based*. That means that during each stage of the workflow, the data can be manually checked and modified by experts and then used in this modified form in the next stage of the workflow. Our goal is not to replace human experts, but to increase the efficiency of human analysis by providing assistance especially in those

tasks which are time consuming, while at the same time making sure that any manual input is checked for internal consistency.

Our study is accompanied by a short tutorial along with code and data needed to replicate the studies illustrated in the following. The workflow runs on all major operating systems. In addition, we have prepared a Code Ocean Capsule² to allow users to test the workflow without installing the software.

3 Illustration of the workflow

3.1 Dataset

The data we use was originally collected by Chén (2012) [8], later added in digital form to the SEALANG project [9], and was then converted to a computer-readable format as part of the CLICS database (<https://clics.clld.org>, [10]). Chén's collection comprises 885 concepts translated into 25 Hmong-Mien varieties. Hmong-Mien languages are spoken in China, Thailand, Laos and Vietnam in Southeast Asia. Scholars divide the family into two main branches, Hmong and Mien. The Hmong-Mien languages have been developing in close contact with neighboring languages from different language families (Sino-Tibetan, Tai-Kadai, Austroasiatic, and Austronesian [11, p. 224]). Chén's study concentrates on Hmong-Mien varieties spoken in China.

In order to make sure that the results can be easily inspected, we decided to reduce the data by taking a subset of 502 concepts of 15 varieties from the dataset. While we selected the languages due to their geographic distribution and their representativeness with respect to the Hmong-Mien language family, we selected the concepts for reasons of comparability with previous linguistic studies. We focus both on concepts that are frequently used in general studies in historical linguistics (reflecting the so-called **basic vocabulary** [12–15]), and

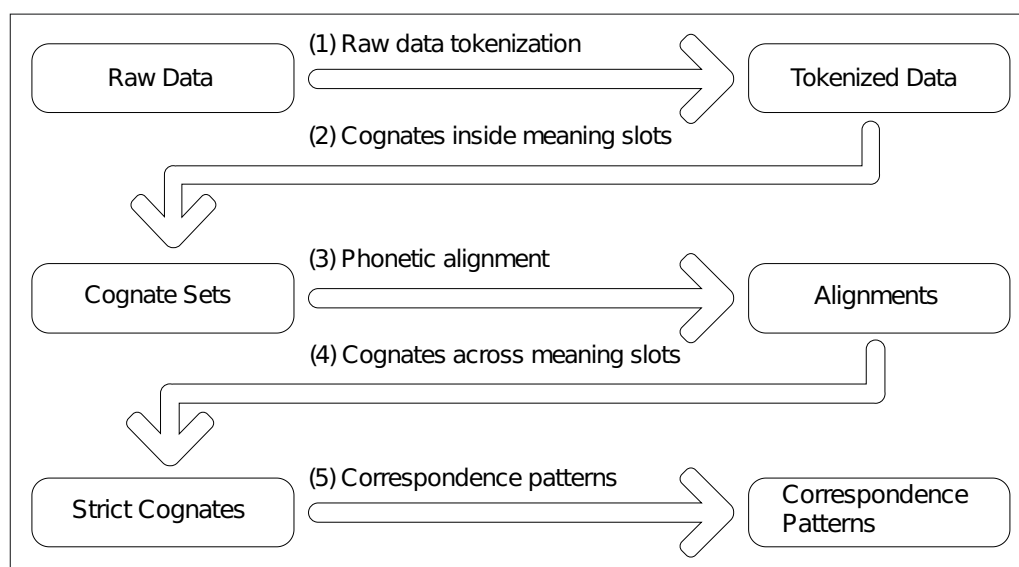


Figure 1: An overview of the workflow.

concepts that have been specifically applied in studies on Southeast Asian languages [4, 16–19]. The 15 varieties are shown in their geographic distribution in **Figure 2**. While the reduction of the data is done for practical reasons, since smaller datasets can be more easily inspected manually, the workflow can also be applied to the full dataset, and we illustrate in the tutorial how the same analysis can be done with all languages in the original data sample.

3.2 Workflow

3.2.1 From raw data to tokenized data

As a first step, we need to lift the data to a format in which they can be automatically digested. Data should be human- and machine-readable at the same time. Our framework works with data in *tabular form*, which is usually given in a simple text file in which the first line serves as table header and the following lines provide the content. In order to apply our workflow, each word in a given set of languages must be represented in one row of the data table, and four obligatory values need to be supplied: an identifier (ID), the name of the language variety (DOCULECT), the elicitation gloss for the concept (CONCEPT), and a phonetic transcription of the word form, provided in tokenized form (TOKENS). Additional information can be flexibly added by placing it in additional columns. **Table 1** gives a minimal example for four words in Germanic languages.

As can be seen from **Table 1**, the main reference of our algorithms is the phonetic transcription in its *tokenized form* as provided by the column TOKENS. Tokenized, in this context, means that the transcription explicitly marks what an algorithm should treat as one sound segment. In **Table 1**, for example, we have decided to render *diphthongs* as one sound. We could, of course, also treat them as two sounds each, but since we know that diphthongs often evolve as a single unit, we made this explicit decision with respect to the tokenization.

Transcriptions are usually not provided in tokenized form. The tokenization thus needs to be done prior to analyzing the data further. While one can easily manually tokenize a few words as shown in **Table 1**, it becomes tedious and error-prone to do so for larger datasets. In order to increase the consistency of this step in the workflow, we recommend using *orthography profiles* [22]. An orthography profile can be thought of as a simple text file with two columns in which the first column represents the values as one finds them in the data, and the second column allows to convert the exact sequence of characters that one finds in the first column into the desired format. An orthography profile thus allows tokenizing a given transcription into meaningful units. It can further be used to modify the original transcription by replacing tokenized units with new values.³ How an orthography profile can be applied is illustrated in more detail in **Figure 3**.

Our data format can be described as a *wide-table format* [23–25] and conforms to the strict principle of entering only *one value per cell* in a given data table. This contrasts with the way in which linguists traditionally code their data, as shown in **Table 2**, where we contrast the original data from Chén with our normalized representation. To keep track of the original data, we reserve the column VALUE to store the original word forms, including those

Table 1: A minimal example for four words in four Germanic languages, given in our minimal tabular format. The column VALUE (which is not required) provides the orthographical form of each word [20, 21].

ID	DOCULECT	CONCEPT	VALUE	TOKENS
1	English	house	house	h au s
2	German	house	Haus	h au s
3	Dutch	house	huis	h ui s
4	Swedish	house	hus	h ʉ: s

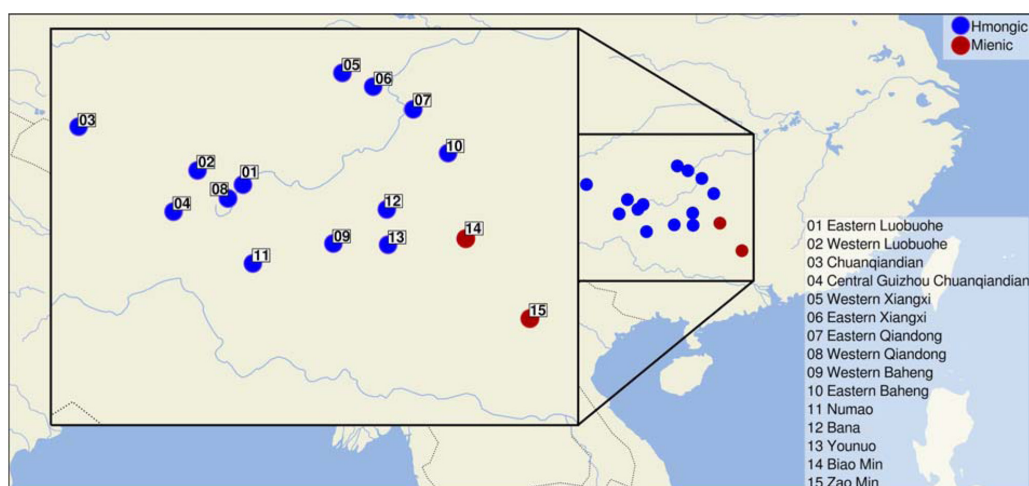
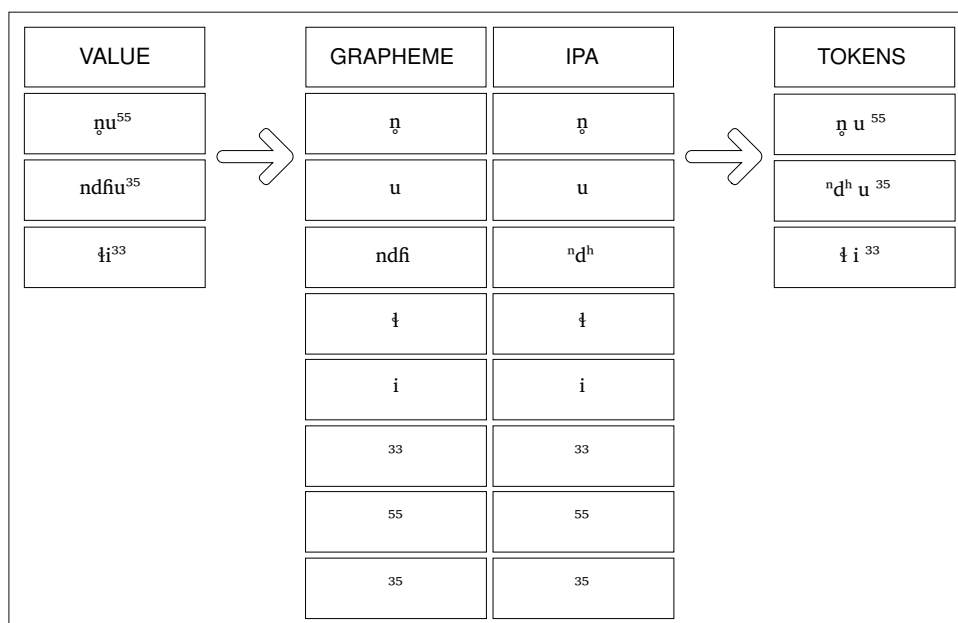


Figure 2: The geographic distribution of the Hmong-Mien languages selected for our sample.

Art. 2, p. 4 of 14

Wu et al: Computer-Assisted Language Comparison

**Figure 3:** An example to illustrate the usage of orthography profiles to tokenize the phonetic transcriptions.**Table 2:** The transformation from raw to machine-readable data. As illustrated in Table 1, the VALUE column displays the raw form. The tokenized forms are added to the TOKENS column.

English	Chinese	Bana	Numao	Zao Min	Biao Min
moon	月亮	la ⁰⁴ la ³⁵	ɬo ⁴⁴	lo ⁴²	la ⁵³ gwan ³³
sun	太陽	la ⁰⁴ ni ¹³	ma ⁴² ŋaŋ ³³	ʔa ⁵³ nai ⁴⁴	ŋi ²¹ tau ³¹
mother	母親	ʔa ⁰⁴ ŋa ³¹³	mai ³³	ni ⁴⁴ ; ze ⁴⁴	ŋa ³¹

a) Raw data as given in the digitized version of Chéns (2012) book.

ID	DOCULECT	SUBGROUP	CONCEPT	VALUE	TOKENS
1	Bana	Hmongic	moon	la ⁰⁴ la ³⁵	l a ^{0/4} + l a ³⁵
2	Numao	Hmongic	moon	ɬo ⁴⁴	ɬ o ⁴⁴
3	ZaoMin	Mienic	moon	lo ⁴²	l o ⁴²
4	BiaoMin	Mienic	moon	la ⁵³ gwan ³³	l a ⁵³ + g w a ŋ
5	Bana	Hmongic	sun	la ⁰⁴ ni ¹³	l a ^{0/4} + n i ¹³
6	Numao	Hmongic	sun	ma ⁴² ŋaŋ ³³	m a ⁴² + ŋ a ŋ
7	ZaoMin	Mienic	sun	ʔa ⁵³ nai ⁴⁴	ʔ a ⁵³ + n ai ⁴⁴
8	BiaoMin	Mienic	sun	ŋi ²¹ tau ³¹	ŋ i ²¹ + t au ³¹
9	Bana	Hmongic	mother	ʔa ⁰⁴ ŋa ³¹³	ʔ a ^{0/4} + ŋ a ³¹³
10	Numao	Hmongic	mother	mai ³³	m ai ⁵³
11	ZaoMin	Mienic	mother	ni ⁴⁴ ; ze ⁴⁴	n i ⁴⁴
12	ZaoMin	Mienic	mother	ni ⁴⁴ ; ze ⁴⁴	z e ⁴⁴
13	BiaoMin	Mienic	mother	ŋa ³¹	ŋ a ³¹

b) Long-table format in which tokenized forms (TOKENS) have been added, and language names have been normalized.

cases where multiple values are placed in the same cell. The separated forms are placed in the column FORM, which itself is converted into a tokenized transcription with the help of orthography profiles.

In order to make sure that our data is comparable with other datasets, we follow the recommendations by the Cross-Linguistic Data Formats initiative (CLDF, <https://cldf.cldf.org>, [24]) and link our languages to the Glottolog database (<https://glottolog.org>, [26]), our concepts to the Concepticon (<https://concepticon.cldf.org>, [27]), and follow the transcription standards proposed by the Cross-Linguistic Transcription Systems initiative (CLTS, <https://clts.cldf.org>, [28]).

In the accompanying tutorial, we show how the data can be retrieved from the CLDF format and converted into plain tabular format. We also show how the original data can be tokenized with the help of an orthography profile (TUTORIAL 3.1).

3.2.2 From tokenized data to cognate sets

Having transformed the original data into a machine-readable format, we can start to search for words in the data which share a common origin. These *etymologically related* words (also called *cognates*) are the first and most crucial step in historical language comparison. The task is not trivial, especially when dealing with languages that diverged a long time ago. A crucial problem is that words are often not entirely cognate across languages [29]. What we find instead is that languages share *cognate morphemes*⁴ (word parts). When languages make frequent use of *compounding* to coin new words, such as in Southeast Asian languages, *partial cognacy* is rather the norm than the exception, which is well-known to historical linguists working in this area [30]. We explicitly address partial cognacy by adopting a numerical annotation in which each morpheme, instead of each word form, is assigned to a specific cognate set [31], as shown in **Figure 4**.

In order to infer partial cognates in our data, we make use of the partial cognate detection algorithm proposed by List et al. [32], which is, so far, the only algorithm available that has been proposed to address this problem. In the tutorial submitted along with this paper, we illustrate in detail how partial cognates can be inferred from the data and how the results can be inspected (TUTORIAL 3.2). In addition, the tutorial quickly explains how the web-based EDICTOR tool (<https://digling.org/tsv/>, [33]) can be used to manually correct the partial cognates identified by the algorithm (TUTORIAL 3.2).

3.2.3 From cognate sets to alignments

An **alignment** analysis is a very general and convenient way to compare sequences of various kinds. The basic idea is to place two sequences into a matrix in such a way that corresponding segments appear in the same column, while placeholder symbols are used to represent those cases where a corresponding segment is lacking (**Figure 5**) [34]. As the core of historical language comparison lies in the identification of regularly recurring sound correspondences across cognate words in genetically-related languages, it is straightforward to make use of alignment analyses once cognates have been detected in order to find patterns of corresponding sounds. In addition to building the essential step for the identification of sound correspondences, alignment analyses also make it easier for scholars to inspect and correct algorithmic findings.

Automated **phonetic alignment analysis** has greatly improved during the last 20 years. The most popular alignment algorithms used in the field of historical linguistics today all have their origin in alignment applications developed for biological sequence comparison tasks, which were later adjusted and modified for linguistic purposes [34].

DOCULECT	CONCEPT	TOKENS	COGID	COGIDS
Chuanqiandian	SUN	ŋ o ⁴³	1	①
Numao	SUN	m a ⁴² + ŋ a ŋ ³³	2	② ①
ZaoMin	SUN	? a ⁵³ + n ai ⁴⁴	3	③ ①
EasternBaheng	SUN	l a ^{0/3} + ŋ e ³⁵	4	④ ①

Figure 4: The comparison of full cognates (COGID) and partial cognate sets (COGIDS). While none of the four words is entirely cognate with each other, they all share a common element. Note that the IDs for full cognates and partial cognates are independent from each other. For reasons of visibility, we have marked the partial cognates shared among all language varieties in red font.

Art. 2, p. 6 of 14

Wu et al: Computer-Assisted Language Comparison

(a)				(b)
DOCULECT	TOKENS	COGIDS	ALIGNMENT	
Chuanqiandian	ŋ o ⁴³	①	ŋ o - 43	
Numao	m a ⁴² + ŋ a ŋ ³³	② ①	ŋ a ŋ 33	
ZaoMin	? a ⁵³ + n ai ⁴⁴	③ ①	n ai - 44	
EasternBaheng	l a ^{0/3} + ŋ e ³⁵	④ ①	ŋ e - 35	

Figure 5: The alignment of 'sun' (cognate ID 1) among 4 Hmong-Mien languages, with segments colored according to their basic sound classes. The table on the left shows the cognate identifiers for cognate morphemes, as discussed in Figure 4. The table on the right shows how the cognate morphemes with identifier 1 (basic meaning 'sun') are aligned.

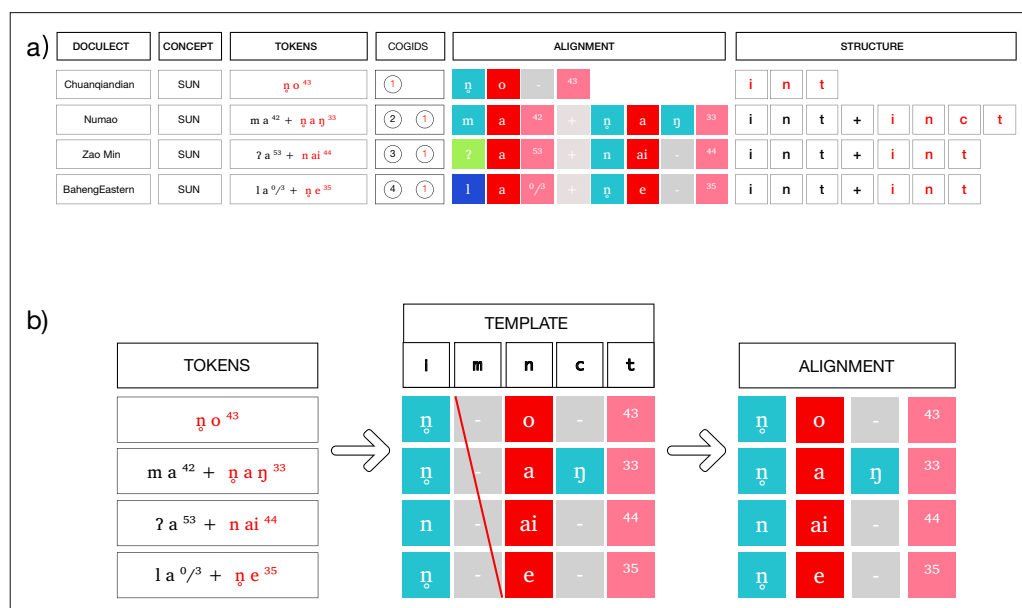


Figure 6: Illustration of the template-based alignment procedure. **a)** Representing prosodic structure reflecting syllable templates for each morpheme in the data. **b)** Aligning tokenized transcriptions to templates, and deleting empty slots.

While the currently available alignment algorithms are all very complex, scholars often forget that the same amount of algorithmic complexity is not needed for all languages. Since most Southeast Asian languages have fixed *syllable templates*, alignments are often predicted by the syllable structure. As a result, one does not need to employ complicated sequence comparison methods in order to find the right matchings between cognate morphemes. All one needs to have is a template-representation of each morpheme in the data.

As an example, consider the typical template for many Southeast Asian languages [35]: syllables consist maximally of an initial consonant (i), a medial glide (m), a nucleus vowel (n), a coda consonant (c), and the tone (t). Individual syllables do not need to have all these positions filled, as can be seen in the following example in **Figure 6a**.⁵

Once the templates of all words are annotated, aligning any word with any other word is extremely simple. Instead of aligning the words with each other, we simply align

them to the template, by filling those spots in the template which have no sounds with gap symbols (“.”). We can then place all words that have been aligned to a template in our alignment and only need to delete those columns in which only gaps occur, as illustrated in **Figure 6b**.

Our accompanying tutorial illustrates how template-based alignments can be computed from the data (TUTORIAL 3.3). In addition, we also show how the alignments can be inspected with the help of the EDICTOR tool (TUTORIAL 3.3).

3.2.4 From alignments to cross-semantic cognates

As in many Southeast Asian languages, most morphologically complex words in Hmong-Mien languages are *compounds*, as shown in **Table 3**. The word for ‘fishnet’ in Northeast Yunnan Chuanqiandian, for example, is a combination of the morpheme meaning ‘bed’ [dz^hau³⁵] and the morpheme meaning ‘fish’ [p^hə³³].⁶ The word for ‘eagle’ in Dongnu is composed of the words [po⁵³] ‘father’ and [təŋ⁵³] ‘hawk’. As can be seen from the word for ‘bull’ in the same variety, [po⁵³və²³¹], [po⁵³] can be used to denote male animals, but in the word for ‘eagle’ it is more likely to denote strength [8, p. 328]. As a final example, Younuo lexicalizes the concept ‘tears’ as [ki⁵⁵mo³²ŋ⁴⁴], with [ki⁵⁵mo³²] meaning ‘eye’ and [ŋ⁴⁴] meaning ‘water’.

An important consequence of the re-use of word parts in order to form new words in highly isolating languages of Southeast Asia, is that certain words are not only cognate *across* languages, but also *inside* one and the same language. However, since our algorithm for partial cognate detection only identifies those word parts as cognate which appear in words denoting the same meaning, we need to find ways to infer the information on *cross-semantic cognates* in a further step.

As an example, consider the data for ‘son’ and ‘daughter’ in five language varieties of our illustration data. As can be seen immediately, two languages, Chuanqiandian and

East Qiandong, show striking partial *colexifications* for the two concepts. In both cases, one morpheme recurs in the words for the two concepts. In the other cases, we find different words, but if we compare the overall cognacy, we can also see that all five languages share one cognate morpheme for ‘son’ (corresponding to the Proto-Hmong-Mien *tɕen in Ratliff’s reconstruction [11]), and three varieties share one cognate morpheme for ‘daughter’ (corresponding to *mphje^D in Ratliff’s reconstruction), with the morpheme for ‘son’ occurring also in the words for ‘daughter’ in East Qiandong and Chuanqiandian, as mentioned before.

While a couple of strategies have been proposed to search for cognates across meaning slots [36, 37], none of the existing algorithms is sensitive to partial cognate relations, as shown in **Table 4**. In order to address this problem in our workflow, we propose a novel approach that is relatively simple, but surprisingly efficient. We start from all *aligned cognate sets* in our data, and then systematically compare all alignments with each other. Whenever two alignments are *compatible*, i.e., they have (1) at least one morpheme in one language occurring in both aligned cognate sets, which is identical, and there are (2) no shared morphemes in two alignments which are not identical, we treat them as belonging to one and the same cognate set (see **Figure 7**). Note that this approach can – by design – only infer *strict cognates* with different meanings, since not the slightest form of form variation for colexification inside the same language are allowed. We iterate over all alignments in the data algorithmically, merging the alignments into larger sets in a greedy fashion, and re-assigning cognate sets in the data.

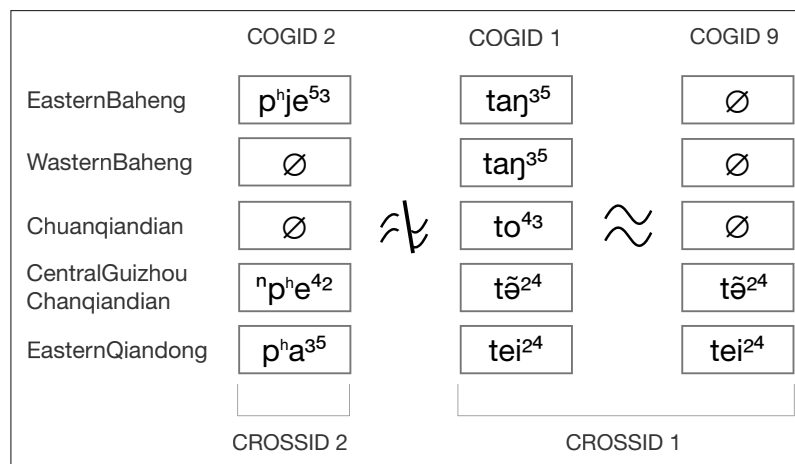
The results can be easily inspected with the help of the EDICTOR tool, for example, by inspecting cognate set distributions in the data, as illustrated in detail in the tutorial (TUTORIAL 3.4). When inspecting only those cognate sets that occur in at least 10 language varieties in our sample,

Table 3: Examples of *compound words* in Hmong-Mien languages. The column MORPHEMES uses morpheme glosses [31] in order to indicate which of the words are cognate inside the same language. The form for ‘net’ in the table serves to show that ‘bed’ and ‘net’ are not colexified, and that instead ‘fishnet’ is an analogical compound word.

DOCULECT	GLOSS	VALUE	TOKENS	MORPHEMES
Northeast-Yunnan-Chuanqiandian	fishnet	dzfau ³⁵ mpə ³³	dz ^h au ³⁵ + p ^h ə ³³	bed fish
	fish	mpə ³³	p ^h ə ³³	fish
	bed	dzfau ³⁵	dz ^h au ³⁵	bed
	net	dzfio ³³	dz ^h o ³³	net
Dongnu	bull	po ⁵³ və ²³¹	p o ⁵³ + v ə ²³¹	father cow
	eagle	po ⁵³ təŋ ⁵³	p o ⁵³ + t ə ŋ ⁵³	father hawk
	father	po ⁵³	p o ⁵³	father
	bovine	və ²³¹	v ə ²³¹	cow
	hawk	təŋ ⁵³	t ə ŋ ⁵³	hawk
Younuo	tear	ki ⁵⁵ mo ³² ŋ ⁴⁴	k i ⁵⁵ + m o ³² + ŋ ⁴⁴	ki-suffix eye water
	water	ŋ ⁴⁴	ŋ ⁴⁴	water
	eye	ki ⁵⁵ mo ³²	k i ⁵⁵ + m o ³²	ki-suffix eye

Table 4: Two glosses, ‘son’ and ‘daughter’, in [8] are displayed here as an example to compare the differences between cognates inside and cognates across meaning slots.

DOCULECT	CONCEPT	FORM	Cognacy	Cross-Semantic
EasternBaheng	SON	taŋ ³⁵	1	1
EasternBaheng	DAUGHTER	p ^h je ⁵³	2	2
WesternBaheng	SON	ʔa ^{3/0} + taŋ ³⁵	3 1	3 1
WesternBaheng	DAUGHTER	ta ⁵⁵ + qa ^{3/0} + t ^h jei ⁵³	4 5 6	4 5 6
Chuanqiandian	SON	to ⁴³	1	1
Chuanqiandian	DAUGHTER	ⁿts ^h ai ³³	7	7
CentralGuizhouChuanqiandian	SON	tə ^{2/0} + tã ²⁴	8 1	8 1
CentralGuizhouChuanqiandian	DAUGHTER	tã ²⁴ + ⁿp ^h e ⁴²	9 2	1 2
EasternQiandong	SON	tei ²⁴	1	1
EasternQiandong	DAUGHTER	tei ²⁴ + p ^h a ³⁵	9 2	1 2

**Figure 7:** Compare alignments for morphemes meaning ‘son’ and ‘daughter’ as an example to illustrate how cross-semantic cognates can be identified. The cognate sets in which the forms in the languages are identical are clustered together and assigned a unique cross-semantic cognate identifier (CROSSID). Those which are not compatible as the cognate sets 2 and 1 in our example are left separate.

we already find quite a few interesting cases of cross-semantic cognate sets: morphemes denoting the concept ‘one’, for example, recur in the words for ‘hundred’ (indicating that hundred is a compound of ‘one’ plus ‘hundred’ in all languages); morphemes recur in ‘snake’ and ‘earthworm’ (reflecting that words for ‘snake’ and ‘earthworm’ are composed of a morpheme ‘worm’); and ‘left’ and ‘right’ share a common morpheme (indicating an original meaning of ‘side’ for this part, such as ‘left side’ vs. ‘right side’).

3.2.5 From cross-semantic cognates to sound correspondence patterns

Sound correspondences, and specifically sound *correspondence patterns* across multiple languages, can be seen

as the *core objective* of the classical comparative method and build the basis of further endeavors such as the reconstruction of proto-forms or the reconstruction of phylogenies. Linguists commonly propose *sound correspondence sets*, that is, collections of sound correspondences which reconstruct back to a common proto-sound (or sequence of proto-sounds) in the ancestor language, as one of the final stages of historical language comparison. In Hmong-Mien languages, for example, Wang proposed 30 sets [38] and Ratliff reduced the quantity of correspondence sets to 28 [11].

An example for the representation of sound correspondence sets in the classical literature [11] is provided in **Table 5**. The supposed proto-sound **ntshj*- in

Table 5: An example of correspondence sets in the classical literature, following Ratliff [11, p. 75], reconstructed forms for Proto-Hmong-Mien are preceded by an asterisk.

	1	2	3	4	5	6	7	8	9	10	11
blood [*ntshjamX]	chan ³	ntchi ³	ntsha ³	ntsua ^{3b}	nʔtshen ^B	θi ³	ne ³	cam ³	sa:m ³	san ³	dzjem ³
head louse [*ntshjeiX]	chu ³	ntchi ³	ntsau ^{3b}	ntsɔ ^{3b}	nʔtshu ^B	—	tchi ³	ceib ³	tθei ³	—	dzei ³
to fear/be afraid [*ntshjeX]	chi ¹	—	ntʂai ⁵	ntse ^{5b}	nʔtshe ^C	ntfei ¹	ne ⁵	dza ⁵	da ^{5r}	da ⁵	dzje ⁵
clear [*ntshjɛŋ]	chi ¹	—	ntʂia ¹	ntsæin ^{1b}	nʔtshe ^A	—	nī ¹	dzan ¹	—	—	—

proto-Hmong-Mien is inferred from the initials of four words in 11 contemporary Hmong-Mien languages.

Although this kind of data representation is typical for classical accounts on sound correspondence patterns in historical language comparison, it has several shortcomings. First, the representation shows only morphemes, and we are not informed about the full word forms underlying the patterns. This is unfortunate, since we cannot exclude that compound words were already present in the ancestral language, and it may likewise be possible that processes of compounding left traces in the correspondence patterns themselves. Second, since scholars tend to list sound correspondence patterns merely in an exemplary fashion, with no intent to provide full frequency accounts, it is often not clear how strong the actual evidence is, and whether the pattern at hand is exhaustive, or merely serves to provide an example. Third, we are not being told where a given sound in a given language fits a general pattern less well. Thus, we can find two different *reflexes* in language 8 in the table, [ɕ] and [dz], but without further information, we cannot tell if the differences result from secondary, conditioned sound changes, or whether they reflect irregularities that the author has not yet resolved.

To overcome these shortcomings, we employ a two-fold strategy. We first make use of a new method for sound correspondence pattern detection [7] in order to identify exhaustively, for each column in each alignment of our data, to which correspondence pattern it belongs. In a second step, we use the EDICTOR tool to closely inspect the patterns identified by the algorithm and to compare them with those patterns proposed in the classical literature.

The method for correspondence pattern identification starts by assembling all *alignment sites* (all columns) in the aligned cognate sets of the data, and then clusters them into groups of compatible sound correspondence patterns. Compatibility essentially makes sure that no language has more than one reflex sound in all partitioned alignment sites (see [7] for a detailed explanation of this algorithm).

Table 6 provides some statistics regarding the results of the correspondence pattern analysis. The analysis yielded a total of 1392 distinct sound correspondence patterns (with none of the patterns being compatible with any of the other 1392 patterns). While this may seem a lot, we find that 234 patterns only occur once in the data (probably reflecting borrowing events,

Table 6: A summary of the result of the sound correspondence pattern inference algorithm applied to our data. The numbers below each item are the quantities of sound correspondence patterns detected at each position in the syllables.

Position	'Regular' Patterns	Singletons
Initial	165	106
Medials	45	23
Nucleus	213	57
Coda	66	13
Tone	164	29
Total	653	228

erroneously coded cognates, or errors in the data).⁷ Among the non-singleton patterns, we find 302 corresponding to initials, 74 to medials, 389 to nucleus vowels, 95 to the codas, and 298 to the tone patterns. These numbers may seem surprising, but one should keep in mind that phonological reconstruction will assign several distinct correspondence patterns to the same proto-form and explain the divergence by means of conditioning context in sound change.⁸ So far, there are few studies on the numbers of distinct correspondence patterns one should expect, but the results we find for the Hmong-Mien dataset are in line with previous studies on other language families [7]. More studies are needed in order to fully understand what one ought to expect in terms of the numbers of correspondence patterns in datasets of various sizes and types.

While the representation in textbooks usually breaks the unity of morphemes and word forms, our workflow never loses track of the words, although it enables users to look at the morphemes and at the correspondence patterns in isolation. Our accompanying tutorial shows not only how the correspondence patterns can be computed (TUTORIAL 3.5), but also how they can be inspected in the EDICTOR tool (TUTORIAL 3.5), where we can further see that our analysis uncovers the correspondence pattern shown in **Table 5** above, as we illustrate in **Table 7**. Here, we can see that our approach confirms Ratliff's pattern by clustering initial consonants of cognates for 'blood' and 'fear (be afraid)' into one correspondence pattern.⁹

Table 7: Cells shaded in blue indicate the initial consonants belonging to a common correspondence pattern, with missing reflexes indicated by a \emptyset .

Language	'blood'		'fear (be afraid)'	
Numao	$^{n}ts^h$	a n ¹³	$^{n}ts^h$	ei ³³
Western Luobuohe	$^{n}ts^h$	e n ⁴⁴	$^{n}ts^h$	e ³⁵
Biao Min	s	a n ³⁵	\emptyset	
Zao Min	$ʐ$	a m ²⁴	$ʐ$	a ⁴²
Younuo	ts^h	u n ³³	ts^h	i ⁴⁴
Western Xiangxi	$^{n}tɕ^h$	i ⁴⁴	$^{n}tɕ^h$	a ⁵³
Eastern Luobuohe	$^{n}ts^h$	e n ⁴⁴	$^{n}ts^h$	e ²⁴
Bana	\emptyset		dʒ	i ¹³
Eastern Xiangxi	ts^h	i ⁵⁵	\emptyset	
Western Qiandong	$ɕ^h$	ẽ ¹³	$ɕ^h$	e ⁴⁴
Eastern Baheng	$^{n}tɕ^h$	e ³¹³	\emptyset	
Chuanqiandian	$^{n}tʂ^h$	a ŋ ⁵⁵	$^{n}tʂ^h$	ai ⁴⁴
Western Baheng	\emptyset		\emptyset	
Central Guizhou Chuanqiandian	$^{n}s^h$	õ ¹³	$^{n}s^h$	e ⁴²
Eastern Qiandong	ɕ	a n ³³	ɕ	a ²⁴

4 Discussion

Although our workflow represents what we consider the current state of the art in the field of computational historical linguistics, it is not complete yet, and it is also not perfect. Many more aspects need to be integrated, discussed, and formalized. Based on a quick discussion of the general results of our study, we will discuss three important aspects, namely, (a) the current performance of the existing algorithms in our workflow, (b) possible improvements of the algorithms, and (c) general challenges for all future endeavors in computer-assisted or computational historical linguistics.

4.1 Current performance

Historical language comparison deals with the reconstruction of events that happened in the past and can rarely be directly verified. Our knowledge about a given language family is constantly evolving. At the same time, debate on language history is never free of disagreement among scholars, and this is also the case with the reconstruction of Hmong-Mien.¹⁰ As a result, it is not easy to provide a direct evaluation of the performance of the computational part of the workflow presented here.

In addition to these theoretical problems, evaluation faces practical problems. First, classical resources on historical language comparison of Hmong-Mien are not available in digital form (and digitizing them would be beyond the scope of this study). Second, and more importantly, however, even when having recent data on Hmong-Mien reconstruction in digital form, we could not compare them directly with our results due to the difference in the workflows. All current studies merely consist of morphemes that were taken from different sources without giving reference to the original words [31]. Full words,

which are the starting point in our study, are not reported and apparently not taken into account. For a true evaluation of our workflow, however, we would need a manually annotated dataset that would show the same completeness in terms of annotation as the one we have automatically produced. Furthermore, since our workflow is explicitly thought of as computer-assisted and not purely computational, the question of algorithmic performance is rather aesthetical than substantial, given that the computational approaches are merely used to ease the labor of the experts.

Nevertheless, to some degree, we can evaluate the algorithms which we assembled for our workflow here, and it is from these evaluations that have been made in the past, that we draw confidence in the overall usefulness of our workflow. Partial cognate detection, as outlined in Section 3.2, for example, has been substantially evaluated with results ranging between 90% (Chinese dialects) and 94% (Bai dialects) compared to expert judgments. The alignment procedure we propose is supposed to work as good as an expert, provided that experts agree on the prosodic structure we assign to all morphemes. For the cross-semantic cognate set detection procedure we propose, we do not yet have substantial evaluations, since we lack sufficient test data. The correspondence pattern detection algorithm has, finally, been indirectly evaluated by testing how well so far unobserved cognate words could be predicted (see also [39]), showing an accuracy between 59% (Burmish languages) and 81% (Polynesian languages) for trials in which 25% of the data was artificially deleted and later predicted.

As another quick way to check if the automated aspects of our workflow are going in the right direction, we can compute a phylogeny based on shared cross-semantic

cognates between all language pairs and see if the phylogeny matches with those proposed in the literature. This analysis, which can be inspected in detail in the accompanying tutorial (TUTORIAL 4.2), shows that the automated workflow yields a tree that correctly separates not only Hmongic from Mienic languages but also identifies all smaller subgroups commonly recognized.

4.2 Possible improvements

The major desideratum in terms of possible improvements is the inclusion of further integration of our preliminary attempts for *semi-automated reconstruction*, starting from already identified sound correspondence patterns. Experiments are ongoing in this regard, but we have not yet had time to integrate them fully.¹¹ In general, our workflow also needs a clearer integration of automatic and manual approaches, ideally accompanied by extensive tutorials that would allow users to start with the tools independently. This study can be seen as a first step in this direction, but much more work will be needed in the future.

4.3 General challenges

General challenges include the full-fledged *lexical reconstruction of words*, i.e., a reconstruction that would potentially also provide compounds in etymological dictionaries. This might help to overcome a huge problem in historical language comparison in the Southeast Asian area, where scholars tend to reconstruct only morphemes, and rarely attempt at the reconstruction of real word forms in the ancestral languages [31]. Furthermore, we will need a convincing annotation of sound change that would ideally allow us to even check which sounds changed at which time during language history.

5 Outlook

This article provides a detailed account on what we consider the current state of the art in computer-assisted language comparison. Starting from raw data, we have shown how these can be successively lifted to higher levels of annotation. While our five-step workflow is intended to be applied in a computer-assisted fashion, we have shown that even with a purely automatic approach, one can already achieve insightful results that compare favorably to results obtained in a purely manual approach. In the future, we hope to further enhance the workflow and make it more accessible to a wider audience.

Notes

¹ By “state of the art”, we refer to approaches that have been developed during the past two decades and are available in the form of free software packages that can be used on all major computing platforms and have shown to outperform alternative proposals in extensive tests. These approaches themselves build on both qualitative and quantitative considerations that have been made in the field of historical linguistics during the past two centuries (for early quantitative and formal approaches, compare, for example, Hoenigswald [40] and Kay [41]).

² The permanent link of the Code Ocean Capsule is: <https://codeocean.com/capsule/8178287/tree/v2>.

³ Orthography profiles proceed in a greedy fashion, converting grapheme sequences in the reverse order of their length, thus starting from the longest grapheme sequence.

⁴ Linguistic terms which are further explained in our glossary, submitted as part of the supplementary information, are marked in bold font the first time they are introduced.

⁵ Note that this template of *i(nitial) m(edial) n(ucleus) c(oda)* and *t(one)* is generally sufficient to represent all syllables in the Hmong-Mien data we consider here. Seemingly complex cases, such as *ntsæn*²² “clear”, for example, can be handled by treating *nts* as one (initial) sound, resulting in a phonetic transcription of [ʰts æ n²²].

⁶ We are aware of the fact that the transcriptions by Chén are not entirely “phonetic”, but since they are much less phonologically abstract than, for example, the transcriptions provided by Ratliff [11], we prefer to place them in phonetic rather than phonological brackets.

⁷ In cases of very intensive language contact, one would expect to find recurring correspondence patterns that include borrowings, but in the case of sporadic borrowings, they will surface as exceptions.

⁸ How this step of identifying conditioning context can be done in concrete is not yet entirely clear to us. Computational linguists often use *n-gram* representations in order to handle context of preceding and following sounds, but this would not allow us to handle situations of remote context.

⁹ The other two cognate sets in Ratliff’s data could not be confirmed, because they do not occur in our sample.

¹⁰ Compare, for example, the debate about regular epenthesis in Proto-Hmong-Mien among Ratliff [42] and Ostapirat [43].

¹¹ A specific problem in semi-automated reconstruction consists in the importance of handling conditioning context in sound change. To our knowledge, no approaches that would sufficiently deal with this problem have been proposed so far. This reflects one apparent problem of common alignment approaches, as they cannot handle cases of *structural equivalence* which require information on conditioning context [44].

Supplementary information and material

The appendix that is submitted along with this study consists of two parts. First, there is a glossary explaining the most important terms that were used throughout this study. Second, there is a tutorial explaining the steps of the workflow in detail. In addition to this supplementary information, we provide supplementary material in the form of data and code. The data used in this study is archived on Zenodo (DOI: 10.5281/zenodo.3741500) and curated on GitHub (Version 2.1.0, <https://github.com/lexibank/chenhmongmien>). The code, along with the tutorial, has also been archived on Zenodo (DOI:

Art. 2, p. 12 of 14

Wu et al: Computer-Assisted Language Comparison

10.5281/zenodo.3741771) and is curated on GitHub (Version 1.0.0, <https://github.com/lingpy/workflow-paper>). Additionally, our Code Ocean Capsule allows users to run the code without installing anything on their machine; it can be accessed from <https://codeocean.com/capsule/8178287/> (Version 2).

Acknowledgements

This research was funded by the ERC Starting Grant 715618 “Computer-Assisted Language Comparison” (CALC, <http://calc.digling.org>, MSW, NES, JML), the ERC Synergy Grant 609823 “Beyond Boundaries: Religion, Region, Language and the State” (ASIA, NWH), and the Grant of P2BEP1_181779 “Reconstruction of Proto-Western Kho-Bwa” of the Swiss National Science Foundation (TAB). The workflow was presented in the workshop “Recent Advances in Comparative Linguistic Reconstruction” in SOAS, London. We thank the workshop participants for giving valuable feedback regarding several aspects of the workflow in their studies. In addition, we thank Christoph Rzymiski and Tiago Tresoldi who provided technical support on setting up our Code Ocean Capsule.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

MSW, NWH, and JML initiated the study. MSW, NWH, JML, and TAB drafted the workflow. MSW and JML implemented the workflow. NES wrote the glossary. TAB, NWH and NES tested the workflow on different datasets. MSW and JML wrote the accompanying tutorial. MSW and JML wrote the first manuscript. NES, NWH and TAB helped in revising the manuscript. All authors agree with the final version of the manuscript.

References

1. List J-M. Computer-assisted language comparison: Reconciling computational and classical approaches in historical linguistics [Internet]. Jena: Max Planck Institute for the Science of Human History. 2016. Available from: <https://hcommons.org/deposits/item/hc:25045/>.
2. List J-M, Greenhill SJ, Gray RD. The potential of automatic word comparison for historical linguistics. *PLOS ONE*. 2017; 12(1): 1–18. DOI: <https://doi.org/10.1371/journal.pone.0170046>
3. Dellert J, Daneyko T, Münch A, Ladygina A, Buch A, Clarius N, Grigorjew I, Balabel M, Boga HI, Baysarova Z, Mühlenbernd R, Wahle J, Jäger G. NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation*. 2020; 54(1): 273–301. DOI: <https://doi.org/10.1007/s10579-019-09480-6>
4. Sagart L, Jacques G, Lai Y, Ryder R, Thouzeau V, Greenhill SJ, List JM. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America*. 2019; 116(21): 10317–10322. DOI: <https://doi.org/10.1073/pnas.1817972116>
5. Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert R, Gray RD, et al. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*. 2018; 5(171504): 1–17. DOI: <https://doi.org/10.1098/rsos.171504>
6. Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lgarda A, Ney H, Tomás J, Vidal E, Vilar J-M. Statistical approaches to computer-assisted translation. *Computational Linguistics*. 2008; 35(1): 3–28. DOI: <https://doi.org/10.1162/coli.2008.07-055-R2-06-29>
7. List J-M. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*. 2019; 1(45): 137–161. DOI: https://doi.org/10.1162/coli_a_00344
8. Chén Q. *Miányáo yǔwén* 苗瑶语文 [Mao and Yao Language]. Běijīng 北京: Zhōngyāng Mínzú Dàxué 中央民族大学出版社 [Central Institute of Minorities]. 2012. Available from: https://en.wiktionary.org/wiki/Appendix:Hmong-Mien_comparative_vocabulary_list
9. Cooper D. Data Warehouse, Bronze, Gold, STEC, Software. In: *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. 2014; 91–99.
10. Rzymiski C, Tresoldi T, Greenhill SJ, Wu M-S, Schweikhard NE, Koptjevskaja-Tamm M, Gast V, Bodt TA, Hantgan A, Kaiping GA, Chang S, Lai Y, Morozova N, Arjava H, Hübner N, Koile E, Pepper S, Proos M, Epps B, Blanco I, Hundt C, Monakhov S, Panykh K, Ramesh S, Gray RD, Forkel R, List J-M. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*. 2020; 7(13): 1–12. DOI: <https://doi.org/10.1038/s41597-019-0341-x>
11. Ratliff M. Hmong-Mien Language History. Canberra: Pacific Linguistics; 2010.
12. Swadesh M. Lexico-statistic dating of prehistoric ethnic contacts: With special book to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*. 1952; 96(4): 452–463.
13. Swadesh M. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*. 1955; 21(2): 121–137. DOI: <https://doi.org/10.1086/464321>
14. Comrie B, Smith N. Lingua Descriptive Series: Questionnaire. *Lingua*. 1977; 42: 1–72. DOI: [https://doi.org/10.1016/0024-3841\(77\)90063-8](https://doi.org/10.1016/0024-3841(77)90063-8)
15. Liú L, Wáng H, Bǎi Y. *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cǐjí* 现代汉语方言核心词 特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng 南京: Fènghuáng 凤凰. 2007.
16. So-Hartmann H. Notes on the Southern Chin languages. *Linguistics of the Tibeto-Burman Area*. 1988; 11(2): 98–119.
17. Matisoff JA. Variational semantics in Tibeto-Burman. The “organic” approach to linguistic comparison. *Institute for the Study of Human Issues*; 1978.

18. **Blust R.** Variation in retention rate among Austronesian languages. *Unpublished paper presented at the Third International Conference on Austronesian Linguistics*, Bali, January 1981.
19. **Běijīng Dàxué. Hànyǔ fāngyán cíhuì** 汉语方言词汇 [Chinese dialect vocabularies]. Běijīng 北京: Wénzì Gǎigé 文字改革. 1964.
20. **Baayen RH, Piepenbrock R, Gulikers L.** (eds.). *The CELEX Lexical Database*. Philadelphia: University of Pennsylvania; Linguistic Data Consortium; CD-ROM; 1995.
21. **PONS.Eu Online-Wörterbuch.** Stuttgart: Pons GmbH; [Accessed 2019 October 24].
22. **Moran S, Cysouw M.** The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin: Language Science Press; 2018. Available from: <http://langsci-press.org/catalog/book/176>.
23. **Wickham H, others.** Tidy data. *Journal of Statistical Book*. 2014; 59(10): 1–23. DOI: <https://doi.org/10.18637/jss.v059.i10>
24. **Forkel R, List J-M, Greenhill SJ, Rzymiski C, Bank S, Cysouw M, Hammarström H, Haspelmath M, Kaiping G, Gray RD.** Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*. 2018; 5(180205): 1–10. DOI: <https://doi.org/10.1038/sdata.2018.205>
25. **Broman KW, Woo KH.** Data organization in spreadsheets. *The American Statistician*. 2018; 72(1): 2–10. DOI: <https://doi.org/10.1080/00031305.2017.1375989>
26. **Hammarström H, Haspelmath M, Forkel R.** *Glottolog. Version 4.0*. Jena: Max Planck Institute for the Science of Human History; 2019. Available from: <https://glottolog.org>.
27. **List JM, Rzymiski C, Greenhill S, Schweikhard N, Pianykh K, Tjuka A, Tjuka A, Wu M-S, Forkel R.** Concepticon. A resource for the linking of concept lists (Version 2.3.0) [Internet]. Jena: Max Planck Institute for the Science of Human History; 2020. Available from: <https://concepticon.cld.org/>.
28. **List J-M, Anderson C, Tresoldi T, Rzymiski C, Greenhill S, Forkel R.** Cross-linguistic transcription systems (Version 1.3.0). Jena: Max Planck Institute for the Science of Human History; 2019. Available from <https://clts.cld.org/>.
29. **List J-M.** Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*. 2016; 1(2): 119–136. DOI: <https://doi.org/10.1093/jole/lzw006>
30. **Matisoff JA.** On the uselessness of glottochronology for the subgrouping of Tibeto-Burman. In: Renfrew C, McMahon A, Trask L. (eds.), *Time depth in historical linguistics*. 2000; 333–371. Cambridge: McDonald Institute for Archaeological Research.
31. **Hill NW, List J-M.** Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*. 2017; 3(1): 47–76. DOI: <https://doi.org/10.1515/yplm-2017-0003>
32. **List J-M, Lopez P, Baptiste E.** Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)* [Internet]. Berlin: Association of Computational Linguistics; 2016. 599–605. DOI: <https://doi.org/10.18653/v1/P16-2097>
33. **List J-M.** A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics System Demonstrations* [Internet]. Valencia: Association for Computational Linguistics; 2017. 9–12. Available from: <https://digling.org/edictor/>. DOI: <https://doi.org/10.18653/v1/E17-3003>
34. **List J-M, Walworth M, Greenhill SJ, Tresoldi T, Forkel R.** Sequence comparison in computational historical linguistics. *Journal of Language Evolution*. 2018; 3(2): 130–44. DOI: <https://doi.org/10.1093/jole/lzy006>
35. **Wang WS-Y.** Linguistic diversity and language relationships. In: Huang C-T J. (ed.) *New horizons in Chinese linguistics*. Dordrecht: Kluwer; 1996. 235–267. (Studies in natural language and linguistic theory). DOI: https://doi.org/10.1007/978-94-009-1608-1_8
36. **Arnaud AS, Beck D, Kondrak G.** Identifying cognate sets across dictionaries of related languages. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017; 2509–2518. Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/D17-1267>
37. **Wahle J.** An approach to cross-concept cognacy identification. In: Bentz C, Jäger G, Yanovich I. (eds.) *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: Eberhard-Karls University; 2016. DOI: <https://doi.org/10.15496/publikation-10060>
38. **Wang F.** *Miáoyǔ gǔyīn gòunǐ* 苗语古音构拟 [Reconstruction of the sound system of Proto-Miao]. Tokyo: Institute for the Study of languages; Cultures of Asia; Africa; 1994.
39. **Bodt TA, List J-M.** Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology*. 2019; 4(1): 22–44. DOI: <https://doi.org/10.2218/pihph.4.2019.3037>
40. **Hoenigswald HM.** Phonetic similarity in internal reconstruction. *Language*. 1960; 36(2): 191–192. DOI: <https://doi.org/10.2307/410982>
41. **Kay M.** *The logic of cognate recognition in historical linguistics*. Santa Monica: The RAND Corporation; 1964.
42. **Ratliff M.** Against a regular epenthesis rule for Hmong-Mien. *Papers in Historical Phonology*. 2018 Dec; 3. DOI: <https://doi.org/10.2218/pihph.3.2018.2877>
43. **Ostapirat W.** Issues in the reconstruction and affiliation of Proto-Miao-Yao. *Language and Linguistics*. 2016; 17(1): 133–145. DOI: <https://doi.org/10.1177/1606822X15614522>
44. **List J-M.** Beyond edit distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*. 2019; 45(3–4): 1–10. DOI: <https://doi.org/10.1515/tl-2019-0016>


Art. 2, p. 14 of 14

Wu et al: Computer-Assisted Language Comparison

How to cite this article: Wu M-S, Schweikhard NE, Bodt TA, Hill NW, List J-M. 2020 Computer-Assisted Language Comparison: State of the Art. *Journal of Open Humanities Data* 6: 2. DOI: <https://doi.org/10.5334/johd.12>

Published: 22 May 2020

Copyright: © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 Unported License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS 

2.5 Tutorial

This tutorial supplements the study “Computer-Assisted Language Comparison: State of the Art”. In this tutorial, we explain in detail how our workflow can be tested and applied.

The workflow consists of several Python libraries that interact, one producing the data that can be used by the other. Since the data are available in different stages, each stage allows us to intervene by manually correcting errors that were made in the automated approach.

For users who are interested in testing our workflow on their local machines or applying it further in their own research, some basic knowledge of the Python programming language and the command line will be required. All the software offered here is available for free. For more information about *LingPy*, the main programming library used here, we recommend that users consult the tutorial (<https://github.com/lingpy/lingpy-tutorial>) accompanying the study “Sequence Comparison in Computational Historical Linguistics (<https://academic.oup.com/jole/article/3/2/130/5050100>)” by List et al. (2018).

2.5.1 Code Ocean Capsule

In order to facilitate the rapid testing of our workflows without installing the software, we have established a Code Ocean Capsule that users can use to run the code remotely. Code Ocean is an open-access platform that enables researchers to reproduce their or others’ experiments. For a detailed introduction to the Code Ocean platform, please refer to the website (<https://codeocean.com/>). To see how our experiments can be run from within the Code Ocean Capsule, follow these steps:

- Navigate to the capsule: <https://codeocean.com/capsule/8178287/tree/v2>.
- Press the “Re-Run” button to reproduce the results.
- View the progression in the “Terminal” panel.
- Download all the results and unzip the .zip file for further inspection of <https://digling.org/editor/>.

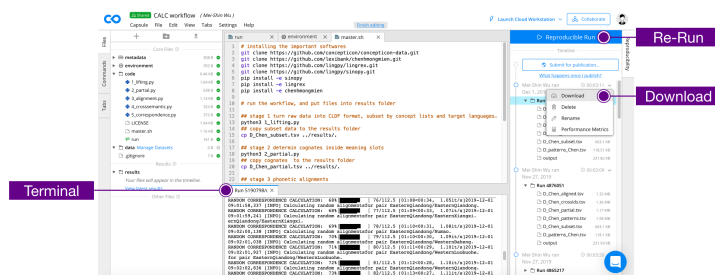


Figure 2.4: The structure of the CALC-workflow Code Ocean capsule.

The following files can be found in the downloaded file:

File	Stage	Section
D_Chen_subset.tsv	From Raw Data to Tokenized Data	2.5.3.1
D_Chen_partial.tsv	From Tokenized Data to Cognate Sets	2.5.3.2
D_Chen_aligned.tsv	From Cognate Sets to Alignments	2.5.3.3
D_Chen_crossids.tsv	From Alignments to Cross-Semantic Cognates	2.5.3.4
D_Chen_patterns.tsv	From Cross-Semantic Cognates to Sound Correspondence	2.5.3.5
D_Chen_distance.dst	Validation	2.5.3.6
D_Chen_tree.tre	Validation	2.5.3.6

2.5.2 Installation Instructions

We assume that users who are interested in running the workflow on their local machines are familiar with the essentials of command-line operations and system administration on either Unix-like systems (such as Linux and MacOS) or Windows systems. Moreover, users should have Python installed (<https://www.python.org/>, version 3.5 or higher), including the package manager **pip**; in addition, the version control system **git** will be required⁴. We strongly encourage users to run this code in a *virtual environment*. A virtual environment is a practical solution for creating independent configurations for testing and experimenting, with no interference on the system-wide installation, and without requiring complex virtualization or containerization solutions. The Python Packaging User Guide⁵ provides clear instructions for setting up a virtual environment on Windows, Linux, and macOS.

We begin by installing the dependencies from the command line. In order to do so, we first download the code that we will use with the help of **git**.

```
git clone https://github.com/lingpy/workflow-paper.git
cd workflow-paper
```

Now that we have done this, we can install all the packages we will need with the assistance of **pip**.

```
pip install -r requirements.txt
```

Now that this has been done, we need to configure the access to reference catalogs, such as Concepticon (<https://github.com/concepticon/concepticon-data>) and CLTS (<https://github.com/cldf-clts/clts/>), in order ensure that they can be accessed readily by the code. This can be done with help of the **catconfig** argument submitted using the *cldfbench* package, which organizes the linguistic datasets.

⁴<https://git-scm.com/>

⁵<https://packaging.python.org/guides/installing-using-pip-and-virtual-environments/>

```
cldfbench catconfig
```

```
cldfbench lexibank.makecldf chenhmongmien
```

You will be prompted to ask if you want to clone actual versions of Concepticon, Glottolog, and CLTS, and the easiest way to address this is to agree and type “y” in all cases.

2.5.3 Getting Started

There are two basic ways in which you can run our workflow.

1. You can run it by downloading a set of Python scripts and running them directly on your computer.
2. You can use the *cldfbench* package to run the commands via the command line without downloading the data directly.

The advantage of Solution 2 is that you do not have to download extra data, as we have integrated the code directly into the **lexibank** version of the data set of Hmong-Mien languages by Chen (2012). Once this data set has been installed (and this is the first package that we installed in the previous section as part of all the dependencies needed), one can type commands on the command line, and the code will be carried out. The disadvantage is that the code example itself is not particularly easy to process for people who are less experienced with Python. Therefore, we will only note the commands in each of the steps we discuss in the following section, and will not explain them in more detail.

2.5.3.1 From Raw Data to Tokenized Data

The first script essentially loads the data from the repository and creates a word list that contains a subselection of all the data that were used. Some aspects of the more difficult “lifting” of data have already been done and distributed, along with the original data package (<https://github.com/lexibank/chenhmongmien>), which specifically also contains the orthography profile in the file **etc/orthography.tsv**, and can be automatically applied with the assistance of the *cldfbench* package.

However, since the data are available in the form of a *cldf* package with the original orthography already tokenized to the formats we needed, one can also skip this step and convert the data into the word list format required by the *LingPy* package.

```
python 1_select.py
```

If you want to test the version from the CLDF repository directly with *cldfbench*, you can type **cldfbench chenhmongmien.wf_select**.

This will select part of the languages and part of the concepts, as indicated in the main study, and write them into a file **D_Chen_subsets.tsv**. In addition, you will see some statistics on the terminal; specifically, a table indicating the coverage for each language. If you want to select all languages, and not just a subset, type:

```
python 1_select.py all
```

The output **A_Chen_subset.tsv** is generated when the argument **all** is used. Once the argument **all** is used in the first stage, it has to be added to the rest of the stages to ensure that the workflows process the correct files.

Doculect	Words	Coverage
Bana	502	1.00
BiaoMin	488	0.97
CentralGuizhouChuanqiandian	454	0.90
Chuanqiandian	501	1.00
EasternBahen	492	0.98
EasternLuobuohe	499	0.99
EasternQiandong	442	0.88
EasternXiangxi	492	0.98
Numao	490	0.98
WesternBaheng	500	1.00
WesternLuobuohe	488	0.97
WesternQiandong	494	0.98
WesternXiangxi	502	1.00
Younuo	500	1.00
ZaoMin	455	0.91

You can now inspect the data with help of the *EDICTOR* tool (<https://digling.org/edictor/>). In order to do so, open the tool’s website at <https://digling.org/edictor/>

and wait until the page is loaded (note that we recommend browsing *EDICTOR* in *Firefox*, but *Google Chrome* should not cause problems).

The data are in the file **D_Chen_subset.tsv**; in order to load it to the tool, press the **Browse** button and select the file. Once this has been done, press the **Open the file** button to examine the data, as illustrated in the following figure.

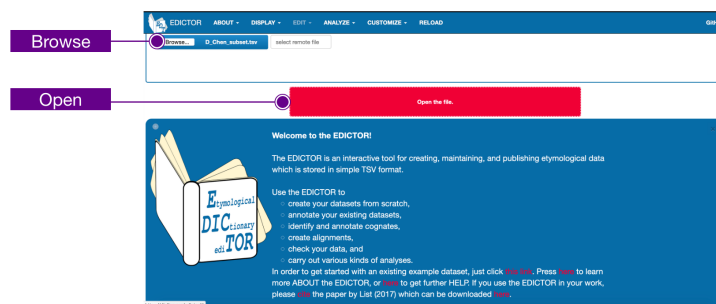


Figure 2.5: The interface of EDICTOR.

The segmented strings are displayed in the **TOKENS** column. Press **Select Columns** to inspect the raw forms and other aspects of the data, as shown in Figure 2.6.

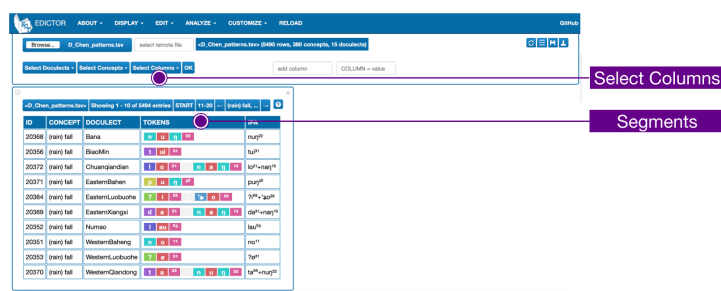


Figure 2.6: The **Select Column** button

In order to save data on your computer, after manual editing, you need to “download” the items. This may be somewhat surprising since you do not effectively download the data but, since the *EDICTOR* is working on a browser, it does not have any access to the data on your computer, and “download” is the only way to communicate with your machine. Thus, in order to save your data and to load it onto your device, you first need to press the **save** icon in the top-right corner in order to store the edited data on the web browser. When pressing the **download** icon at the top right, your browser will either directly download the data and store it in your download folder, or will ask you to specify a particular file destination.

Be careful when editing data in the *EDICTOR* without saving and downloading the items. If you close your browser, all the edits you made will be lost; thus, you should save and download your data when working with the *EDICTOR* regularly. As a shortcut, you could also type **CONTROL+S** to save and **CONTROL+E** to export the data (that is, to download items).

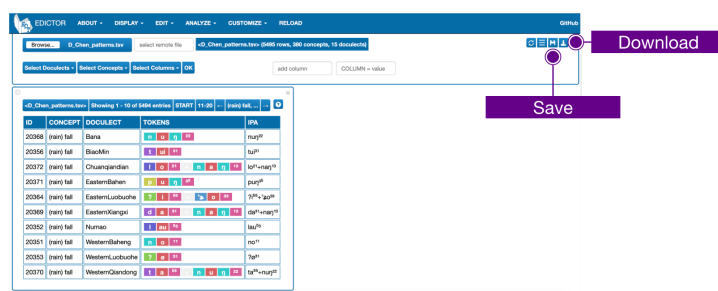


Figure 2.7: Select columns to display on the browser.

2.5.3.2 From Tokenized Data to Cognate Sets

Partial cognate detection is an important task, particularly when working with Southeast Asian language data. The algorithm that we used for this task was first proposed in the study “Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists” by List (2016), in which the algorithm is described in appropriate detail.

To illustrate how the algorithm works, we provide an example with four words for “moon” in the Eastern Baheng, Eastern Qiangdong, Bana, and Biao Min language varieties.

The main steps in the algorithm are the following:

1. Calculate the distances of all morpheme pairs.
2. Create a fully connected network from the distance scores.
3. Filter the network by deleting edges in the following fashion: A. Two morphemes in the same word should not be linked (see the dashed lines in the following figure). B. A morpheme in a word should not be linked to two morphemes in another word (see the yellow edges in the figure).
4. Remove the edges with similarity scores below a given threshold.

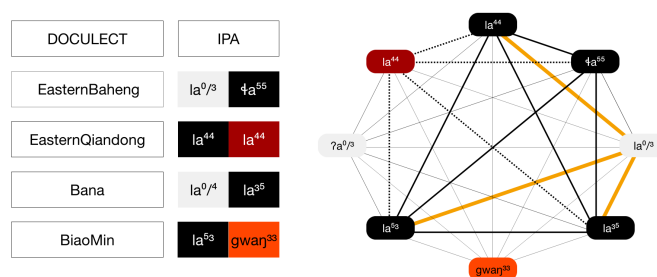


Figure 2.8: A brief introduction to partial cognate detection.

Once this has been done, an algorithm for Community Detection in networks (Rosvall and Bergstrom, 2008) is used to partition the network into “communities”, with each community representing one partial cognate set.

In order to calculate partial cognates, we use the algorithm as provided by the *LingPy* software package and apply it to our subselection of languages.

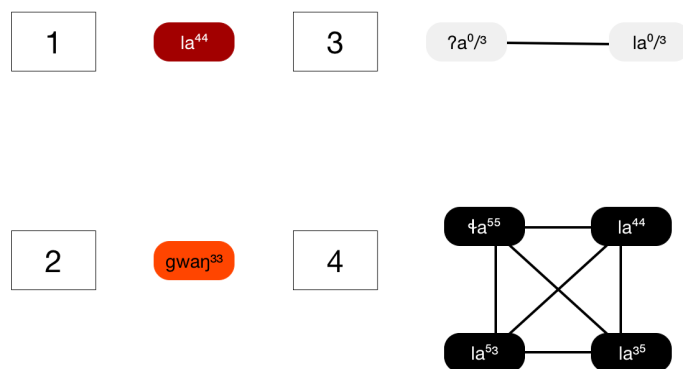


Figure 2.9: Assign a unique identification number (COGID) to each cluster.

```
python 2_partial.py
```

If you want to test the version from the CLDF repository directly with *cldfbench*, you can type **cldfbench chenhmongmien.wf_partial**

This will take some time when you run it the first time. The data can be found in the file **D_Chen_partial.tsv**.

To inspect the data using *EDICTOR*, load **D_Chen_partial.tsv** as shown before. Then press **DISPLAY** to select **SETTINGS** in the drop-down menu. Select **PARTIAL** in the **Morphology and Colexification Mode** entry. Press the **Refresh** button.

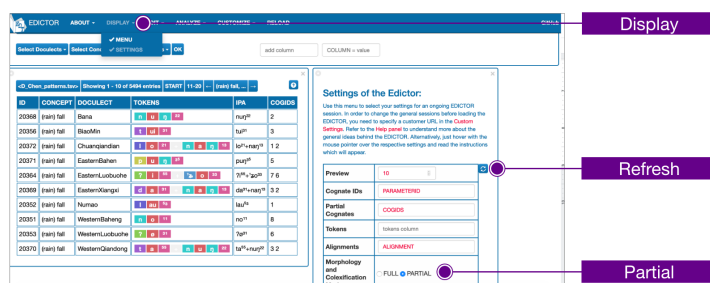


Figure 2.10: The interface of EDICTOR.

In order to investigate the partial cognates, you need to select the column that stores the identifiers. To do so, press **Select Columns** and select **COGIDS** in the drop-down menu. If you right-click on any number in the **COGIDS** column, a pop-up window will open and show all the cognate sets for a given word form in the form of an alignment. Since we have not yet aligned the data, the alignment will be incorrect at this point.

2.5.3.3 From Cognate Sets to Alignments

To align the data, we use the new procedure for template-based alignment, which is available from the *lingrex* package that we have installed as one of the requirements of our workflow, as well as the *sinopy* package, which assisted us to compute syllable templates from all the morphemes in

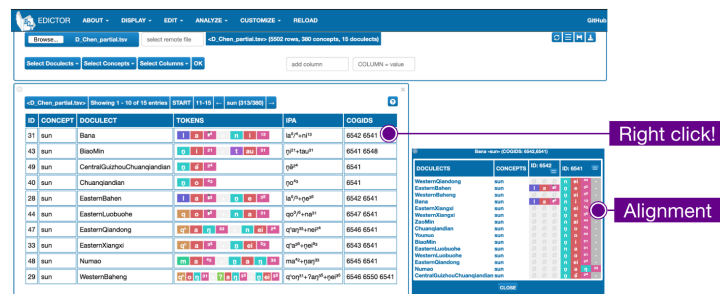


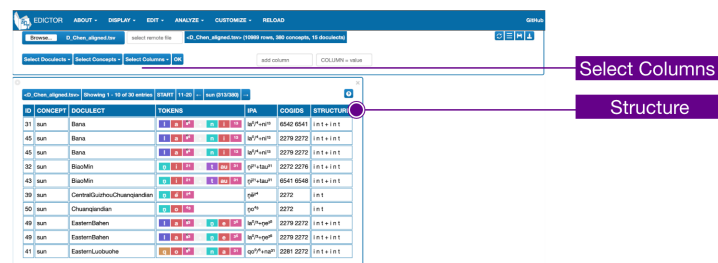
Figure 2.11: Show the alignment.

the data. Running the code is again straightforward.

```
python 3_alignment.py
```

If you want to test the version from the CLDF repository directly using *cldfbench*, you can type **cldfbench chenhmongmien.wf_alignment**.

The aligned data will be stored in the file **D_Chen_aligned.tsv**. To inspect the alignments in *EDICTOR*, load this file and follow the previous steps mentioned in Section 2.5.3.2. In addition to selecting the **COGIDS** column, we now also select the **STRUCTURE** column, since this column provides the templates for each morpheme, which we have automatically added to the data with the assistance of *sinopy*.



As mentioned previously, if you right-click on any number in the **COGIDS** column, a pop-up window will show the alignment. Click on the = sign to modify the alignment. The modification itself is extremely straightforward: Simply click on a sound segment to move it to the right, and click on a gap segment to delete this segment.

2.5.3.4 From Alignments to Cross-Semantic Cognates

The algorithm for cross-semantic cognate detection as we propose it here is illustrated in more detail in the main study. It is implemented as part of the *lingrex* package. Again, running the code is straightforward.

If you want to test the version from the CLDF repository directly using *cldfbench*, you can type **cldfbench chenhmongmien.wf_crosssemantic**.

The output file is **D_Chen_crossids.tsv**, and we load it into the *EDICTOR* tool, just as we

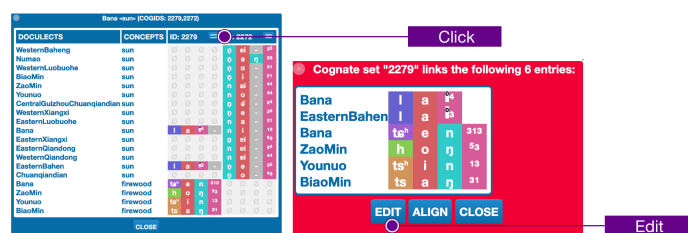


Figure 2.12: Modify the alignment.

```
python 4_crossemantic.py
```

did before but, when checking the **SETTINGS** in the menu this time, we need to specify that the column “CROSSIDS” holds the partial cognates. To do so, simply type **CROSSIDS** in the text field **Partial Cognates** in the settings menu and then press the **refresh** button.

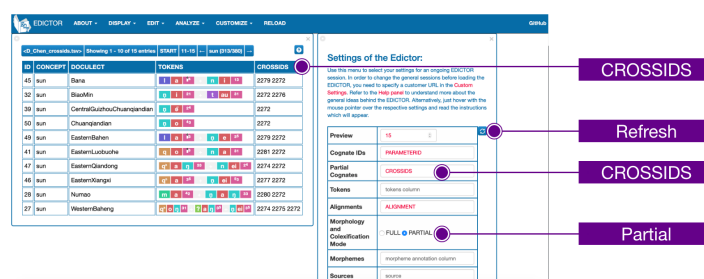


Figure 2.13: Adjust settings.

To inspect the distribution of partial cognates, press **ANALYZE** in the top-level menu and select **Cognate sets** in the drop-down menu.

As a result, a new panel will open, and will show the distribution of all cognate sets across the different language varieties. Pressing the red button with the cognate set identifier on the left will open the alignment. Pressing the yellow buttons with the word identifiers will show the original morpheme. On the right, in the column **CONCEPTS**, you will find the cognate sets that are attested for more than one concept as separated by a comma. Clicking on this field will modify the main word list panel in such a way that only the selected concepts will appear.

2.5.3.5 From Cross-Semantic Cognates to Sound Correspondence Patterns

As a final step, we will attempt to infer the major correspondence patterns in the data using the algorithm by List (2019), which is available from the *lingrex* package. Running the code is straightforward, as previously.

```
python 5_correspondence.py
```

If you want to test the version from the CLDF repository directly using *cldfbench*, you can

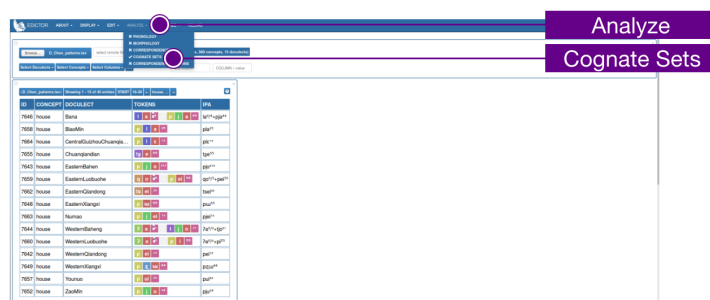


Figure 2.14: Inspect the partial cognates.

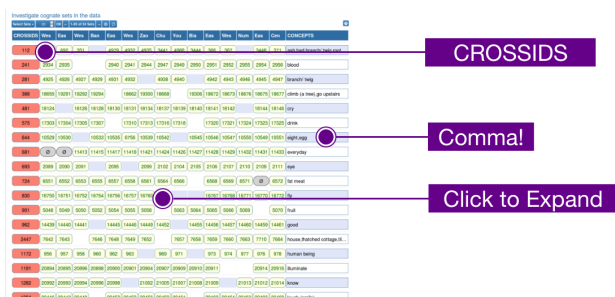


Figure 2.15: The cross-semantic IDs.

type `cldfbench chenhmongmien.wf_correspondence`.

This creates two output files. One, which is called **D_Chen_patterns.tsv**, is the file without a word list that can be loaded by EDICTOR and inspected, and one file contains the patterns that have been inferred alone, called **D_patterns_Chen.tsv**.

In order to inspect the patterns, we recommend using the EDICTOR tool, which requires the same steps that were already applied when loading the cross-semantic cognates. Once this has been done, press the **ANALYZE** button in the top menu and select **CORRESPONDENCE PATTERNS** in the drop-down menu.

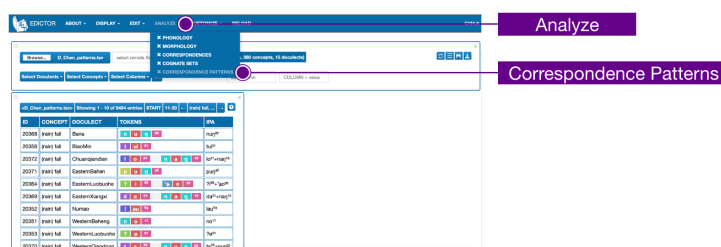


Figure 2.16: Inspect the correspondence patterns.

In order to allow for a good display, the doculect names are all abbreviated. Hovering the mouse cursor over an abbreviation will reveal the full name.

Clicking on a cell in the correspondence pattern panel will allow you to see not only the sound in question, but also the full morpheme in which this sound occurs.

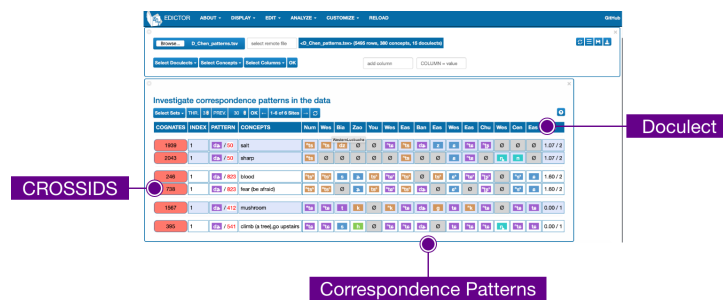


Figure 2.17: Inspect the correspondence patterns in detail.

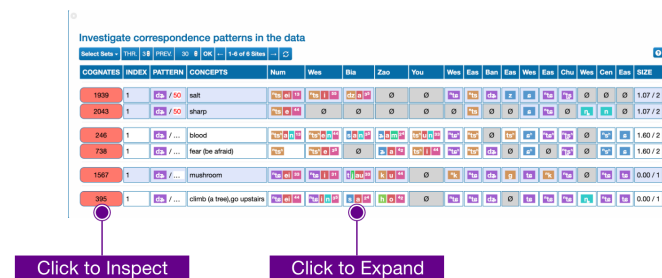


Figure 2.18: Inspect the correspondence patterns in detail.

2.5.3.6 Validation

We calculate the shared cognates between language pairs and output the scores in the form of a pairwise distance matrix. The script `6_phylogeny.py` provides two documents, a distance matrix (**A_Chen_distance.dst** or **D_Chen_distance.dst**), and a tree file based on a neighbor-joining analysis (**A_Chen_tree.tre** or **D_Chen_tree.tre**).

There are many ways to work with the distance matrix; here, we provide one of the approaches to visualizing the matrix as a neighbor-net network with the use of *SplitsTree*.

To get started, first ensure that *SplitsTree* (Huson, 1998) from <https://software-ab.informatik.uni-tuebingen.de/download/splitstree4/welcome.html> is installed, and follow the installation instructions. In order to compute the distance matrix with our code, use the command line (here, we computed it for the entire data set, so we run it using the keyword **all**).

```
python 6_phylogeny.py all
```

To generate a Neighbor-Net from the distance matrix, open the file **A_Chen_distance.dst** or **D_Chen_distance.dst** with any plain text editor and start the *SplitsTree* software. Then click on **File** and **Enter Data**, as shown in Figure 2.19.

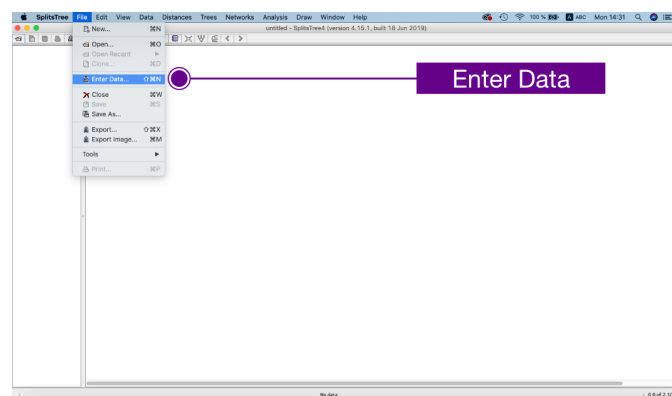


Figure 2.19: Open a dialogue to enter the distance matrix.

Then copy the distance matrix in the paste it into the **Enter Data Dialog**, and press **Execute**.

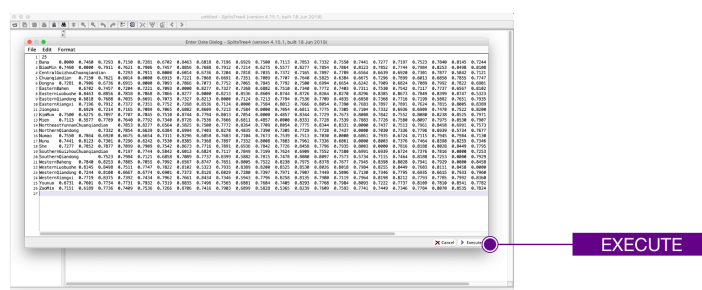


Figure 2.20: Enter the distance matrix.

You can now inspect the network. To analyze the data further, you can compute the delta scores, showing the degree of reticulation in the data, by pressing **Analysis** and then **Compute Delta Score**, as shown in Figure 2.21.

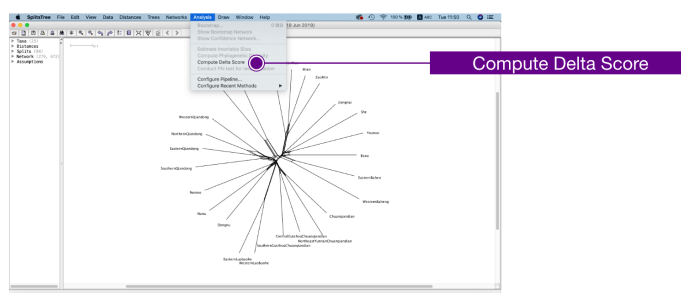


Figure 2.21: Press the button to compute delta scores.

The resulting Neighbor-Net is shown in Figure 2.22. For the purpose of illustration, the Mienic language varieties are colored red, and the Hmongic group is highlighted in blue.

Table 2.2 shows the delta scores we computed from the data.

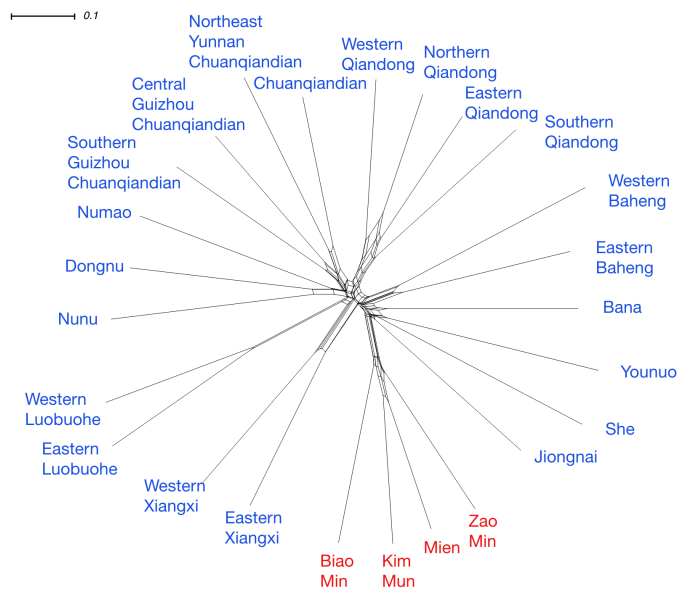


Figure 2.22: The network.

Taxon	Delta score
Bana	0.34706
Biao Min	0.27289
Central Guizhou Chuanqiandian	0.29924
Chuanqiandian	0.29172
Dongnu	0.32416
Eastern Baheng	0.32056
Eastern Luobuohe	0.33529
Eastern Qiangong	0.32083
Eastern Xiangxi	0.33736
Jiongnai	0.32644
Kim Mun	0.26992
Mien	0.25672
Northeast Yunnan Chanqiandian	0.29748
Northern Qiangong	0.28447
Numao	0.34185
Nunu	0.32375

Taxon	Delta score
She	0.31671
Southern Guizhou Chuanqiandian	0.34376
Southern Qiandong	0.30988
Western Baheng	0.35259
Western Luobuohe	0.3211
Western Qiandong	0.31137
Western Xiangxi	0.35174
Younuo	0.2996
Zao Min	0.26797

Table 2.2: Delta scores

The average delta score is 0.313. As mentioned previously, the distances between taxa are calculated via shared cognates. The shorter the distances between two taxa, the greater are the similarities between them. If the taxa share cognates not only within their group but also outside of their groups, the network finds it difficult to determine the best cluster for them. The larger the reticulated structure, or the less tree-like the data, the higher the delta score. Each particular language variety’s delta score means that this specific language contributes to a certain amount of conflict in the data.

2.5.4 Conclusion

In this tutorial, we provided details of how to execute our workflow for computer-assisted language comparison using the scripts we wrote, while simultaneously illustrating how the results can be manually inspected and modified. We have not discussed the details of the code we wrote, but we recommend that users who are proficient in Python have a look at it.

2.6 Retrospective

Generating an orthography profile by relying on the CLTS database implies that one internally agrees with the grapheme to phoneme (GTP) standardization guidelines that are provided by CLTS. To demonstrate our workflow, we chose the IMNCT template to illustrate the tokenization process. As a result, the graphemes that were converted and tokenized according to the IMNCT template may appear odd to some linguists. In particular, the treatments of diphthongs appear to overly modify the raw forms. For example, the diphthong *ua* in our data set is tokenized

as *u/w* and *a*, but the diphthong *au* retains its shape as *au*.⁶

The point to address here is that we do not prohibit users from using customized orthography profiles when working with the CALC workflow. When looking at the code that we provide, please note that there are no strict guidelines stating that users must rely on the CLTS in order to generate an “accurate” orthography profile. We are well aware that there are different levels of analysis for syllable structures; users should use the template that suits the languages’ syllable structures and be consistent throughout the entire experiment.

The dissertation highlights the significance of the CLTS database due to its practicality. The database collects orthography profiles from published linguistic articles and large-scale cross-linguistic databases. Users can begin with one of the orthography profiles included on the database and modify the rules to suit their data set. In addition, many data sets are digitized in the CLDF format. Users curate the data sets according to the CLTS databases, and can increase the compatibility between their data set and the other standardized lexical material. As a result, the possibility that their data set will be re-used by other studies is also increased.

2.6.1 Open Data and the FAIR Principle

Data are the foundation of all quantitative and qualitative research. However, there are usually gate-keeping mechanisms to prevent people from accessing existing data effortlessly. Scientific data are guarded by restricted access rights, the absence of clear authentication or authorization procedures, or data formats that can only be used by specific software. As a result, a huge number of scientific studies are generating first-hand data sets that are “large enough” for their experiments, but which are kept private. Numerous types of data are produced for research projects, but they are rarely re-used or inspected by fellow researchers. Large-scale quantitative studies became a competition to acquire research resources rather than a collaboration involving community efforts.

The open data request was initially a request to access non-sensitive governmental data for non-commercial use. The idea of open data has gained traction in several scientific fields since the 2000s. Both governmental and private sectors have been setting up online storage to permanently store experimental data online as a response to the movement. For example, the NCBI Gene Expression Omnibus (GEO) database provides a space for researchers to upload their micro-array data. The data sets are all free for users to download, validate, and re-use. In recent years, self-archived websites such as *Zenodo*, the *Open Science Framework*, and *Figshare* have been established for users to store their research outputs.

The basic definition of open data addresses re-usability and accessibility (Pollock, 2006). In 2016, the FAIR principle was proposed to further define “open data” in the scientific research domain.

- **Findability.** Data that are used in a study should be accompanied by detailed metadata.

⁶We stated our reasons pertaining to the treatment of diphthongs in the introduction.

Both the data and the metadata should be able to be found by humans and by computers.

- **Accessibility.** Once the data are found, users can access the data either with or without an authentication and authorization request. A procedure for requesting data access rights applies to disciplines that involve the risk of exposing subjects' personal data.
- **Interoperability.** A data set should be inspected and used by different software or platforms. Most linguistic data are printed in book form, which prevents the data from being used by machines.
- **Re-usability.** This is the ultimate goal of FAIR data. Being findable, accessible, and interoperable will greatly increase the likelihood that the data will be re-used by others.

Nevertheless, open data are still an exception rather than the norm in the field of linguistics. We found that linguistic research data were often inaccessible. The data were sometimes presented in books and papers, which cannot be analyzed directly by computer programs. With regard to the data that are presented online, these data are not stored permanently. We encountered the issue of data only being presented and maintained online during the funding phase of a scientific project. The data are taken offline or are no longer maintained once the project is complete. The Kusunda lexical data are an example of this particular case.

This dissertation is based on the second-hand lexical material that we digitized or standardized according to the CLDF formats over the last few years. Our experience of working with lexical data shows that the level of “FAIR-ness” of linguistic data can be improved by putting effort into the standardization process and raising awareness. It is our hope that the workflow can be of use by contributing to the integration of open data in the domain of historical linguistics.

Chapter 3 Morpheme Annotation in Phylogenetic Studies

The Bayesian inference of phylogeny is gradually gaining popularity in the field of historical linguistics. It was developed to model the spread and evolution of biological phenomena, including diseases, bacteria, and viruses. Since then, using the analogy of biology and language, the methodology has been applied to the study of language changes and the relatedness among a set of languages.

A Bayesian phylogenetic analysis allows linguists to express the evolutionary factors via statistical distributions, such as the rate of new languages being born, the rate of existing languages disappearing, and the rate of lexical innovations. The algorithm infers a phylogeny from the given data while referencing the parameters. To date, many studies have shown the feasibility of inferring language differentiation using a Bayesian phylogenetic analysis (Gray et al., 2009; Grollemund et al., 2015; Sagart et al., 2019). Nevertheless, we observed the two significant obstacles when applying the Bayesian framework to historical linguistics. First, the cognate sets and the lexical data cannot be used easily because linguists rarely provide clear guidelines regarding how to use their cognate sets. Second, which is also an important point, the approach models the unknowns using parameters, which makes it susceptible to noise in the input data.

Unfortunately, the existing studies all took the input data for granted without taking the linguistic factors into account. For example, the composition and cognate judgments of the lexical data have not been treated appropriately, despite the problems having been discussed for over a decade (Geisler and List, 2010; Hill and List, 2017; Holm, 2007). Most of the cognate sets are provided based on the word level instead of on the morpheme level. This coding method is confusing and ambiguous when polymorphemic words are involved because, in cognate sets, we often find that not all the parts of the words belong to the same cognate sets. A greater concern is that the coding method overlooks the complexity of language differentiation processes. In particular, the MSEA area has had prolonged language contact with languages within the same language family or across language families. It is not unusual for a compound word to be formed by combining a native morpheme and a loan morpheme. Cases such as this cannot be arbitrarily judged as word cognates. The inadequate cognate coding leads to inappropriate data transformation as input into a Bayesian phylogenetic analysis. Subsequently, we are concerned that the topology of the phylogeny may not reflect the real-world scenario.

In this project, we address the points stated above, which have never been addressed or tested in any of the previous studies. The results provide evidence that the cognate coding in compound words may lead to a different topology. In addition, we developed a morpheme annotation scheme

to overcome the drawbacks in the current cognate annotation practice. Adding morpheme annotation could enhance the clarity of the meaning and the constitution of polymorphemic words. Furthermore, we provide three automatic and one computer-assisted conversion method to convert partial cognate sets into word cognate sets as the input for a Bayesian phylogenetic analysis. We hope that the outcome we present in the paper will encourage people who work within the Bayesian phylolinguistic framework to improve their data preparation processes.

3.1 The Benefits of Morpheme Annotation

In the existing etymology studies, we often find that the cognate sets and the sound correspondences are presented in the form of a table, as in Figure 3.1. The same table format can also be seen in Ratliff (2010), Pan (2007) and other studies. In Chapter 2, we introduced the problem that the actual word forms are not able to be preserved in this type of data presentation. In Chapter 2, we also mentioned that the reason that linguists chose not to report the entire words was due to the definition of cognates: Two **words** are either related or they are not. In this section, we would like to address the problems of data transparency and analysis that are also raised by the data presentation.

59. 繩母	養蒿	臘乙坪	大南山	石門坎	擺托	甲定	紋坨	野鷄坡	楓香	韻類號
↓	lh	lh	↓	↓	↓	lh	l	↓	lh	
橋	—	—	—	la ⁵⁵ ₍₇₎	lo ⁵⁵ ₍₇₎	ha ²⁴ ₍₇₎	la ²² ₍₁₆₎	—	ha ³³ ₍₇₎	(5)
腦髓	he ³³ ₍₇₎	ha ³⁶ ₍₇₎	lu ⁴⁴ ₍₇₎	ly ⁵⁵ ₍₇₎	ou ⁵⁵ ₍₇₎	hur ²⁴ ₍₇₎	ou ²² ₍₁₆₎	lu ³¹ ₍₇₎	lu ³³ ₍₇₎	(8)
燙子	—	—	lar ⁴⁴ ₍₇₎	law ⁵⁵ ₍₇₎	on ⁵⁵ ₍₇₎	hon ²⁴ ₍₇₎	lua ²² ₍₁₆₎	en ³¹ ₍₇₎	—	(24)
月亮	ha ⁴⁴ ₍₆₎	ha ⁵³ ₍₆₎	li ⁴⁴ ₍₆₎	li ⁵⁵ ₍₆₎	le ⁴⁴ ₍₆₎	ha ⁴⁴ ₍₆₎	li ⁵⁵ ₍₆₎	la ²⁴ ₍₇₎	ha ⁵⁵ ₍₆₎	(4)
繩子	ha ⁴⁴ ₍₆₎	ha ⁵⁵ ₍₆₎	ua ⁴⁴ ₍₆₎	la ⁵⁵ ₍₆₎	lo ⁴⁴ ₍₆₎	ha ⁴⁴ ₍₆₎	la ⁵⁵ ₍₆₎	li ²⁴ ₍₇₎	ha ⁵⁵ ₍₆₎	(5)
鐵	ho ⁴⁴ ₍₆₎	ho ⁵³ ₍₆₎	ou ⁴⁴ ₍₆₎	au ⁵⁵ ₍₆₎	lu ⁴⁴ ₍₆₎	ha ⁴⁴ ₍₆₎	lu ⁵⁵ ₍₆₎	o ²⁴ ₍₇₎	hou ⁵⁵ ₍₆₎	(9)
割肉	hei ⁵⁵ ₍₇₎	ha ⁴⁴ ₍₇₎	ai ⁵⁵ ₍₇₎	ai ⁵⁵ ₍₇₎	ai ⁴⁴ ₍₇₎	he ⁴⁴ ₍₇₎	le ⁴⁴ ₍₇₎	—	he ⁵⁵ ₍₇₎	(10)

Figure 3.1: The example extracts from Wang and Mao (1995, p. 24). The table shows the cognates and the sound corresponding pattern across Hmongic languages.

We created the example in Table 3.1. This example model simulates how linguists make cognate judgments: The comparative method compares morphemes inside “words”, and groups the related words into one cognate set. The column “COGID” in the table is the experts’ cognate judgments sourced from Pan (2007) and Wang and Mao (1995). The fourth row 月 *yuè* “moon” entry in Figure 3.1 provided by Wang and Mao (1995) shows the words’ cognacy among various Hmongic languages. Pan (2007) stated that the sound correspondence of the word “moon” was *l-*, *ɬ-*, and *l̥-*. As a result, we assigned the nine words to two different cognate sets; COGID 1 and COGID 2. The word *ne*³³ in the She language is not derived from the same proto-word in the other languages because the *n* violates the sound correspondence *l-*, *ɬ-*, and *l̥-*¹.

We now have Table 3.1, in which we manually added cognate sets to Chen, 2012’s data using external resources. However, this table also raises the issue that the part of the words referring to

¹It is possibly a loanword or a borrowing from the Kam or Tai languages

LANGUAGE ID	VALUE	FORM	COGID
EasternLuobuohe	ʔa ⁰² ɬa ²⁴	ʔa ⁰² ɬa ²⁴	1
NortheastYunnanChuanqiandian	ɬi ³³	ɬi ³³	1
NorthernQiandong	ɬha ⁴⁴	ɬ ^h a ⁴⁴	1
EasternQiandong	la ⁴⁴ la ⁴⁴	la ⁴⁴ la ⁴⁴	1
WesternQiandong	pɔ ¹¹ la ³³	pɔ ¹¹ la ³³	1
WesternBaheng	ʔa ⁰³ ɬa ⁵⁵	ʔa ⁰³ ɬa ⁵⁵	1
Younuo	kwan ¹³ la ⁵⁴	kwan ¹³ la ⁵⁴	1
She	ne ³³	ne ³³	2
Mien	ɬa ²⁴	ɬa ²⁴	1

Table 3.1: The cognacy among nine Hmongic and Mienic words 月亮 “moon”. The data are taken from Chen (2012). The cognate sets in the “COGID” column are annotated according to Pan (2007) in that the sound correspondence of the morpheme “moon” is *l-*, *ɬ-*, and *l-*. GOGID 1 is unclear. The problem is not only that the table does not provide morpheme boundaries, as the parts of the words that denote “moon” are not highlighted. Problematic tables such as the one in our example are found quite frequently in current linguistic studies. The cases are not limited to MSEA linguistic studies, but can also be found for different languages in different geographical areas. We provide an example of Dravidian languages in our paper.

Table 3.1 does not clearly indicate where the morphemes are, or the meanings or functions of each morpheme. It also does not provide sufficient visual cues to indicate which parts are the real cognates. In particular, which part of the Eastern Qiandong (Glottolog: east2370) word la⁴⁴la⁴⁴ has been used to be judged as a cognate with the others is unclear. Hence, it would be beneficial to highlight the parts of the words that belong to a cognate set.

To address the question above, we tokenized the phonetic sequences in the “SEGMENTS” column to show the sound correspondences in each syllable position. We also added a new column, “MORPHEME”, to Table 3.2, which explains the meaning or the function of each part of the words. Adding these two columns to the table enables us to see that the parts that contribute to the meaning “moon” occur mainly in the second morpheme of a multimorphemic word. The word la⁴⁴la⁴⁴ is a reduplicated compound word. Chen (2012, p. 299) explained that a duplicated morpheme was used to emphasize an item being small and delicate. Since the meanings of the morphemes in a reduplicated compound word remain the same, we annotated both morphemes as “moon”.

Linguists who are not studying Hmong-Mien languages can now understand which parts of the words correspond to COGID 1 and the parts that were not used by Pan (2007) and Wang and Mao (1995) to judge words’ cognacy. Table 3.2 shows a simplified version of the proposed morpheme annotation scheme in this section. This simplified version aims to demonstrate how morpheme annotation can assist others to understand lexical data easily without years of research. In our paper, we provide a detailed explanation of our morpheme annotation scheme.

LANGUAGE ID	VALUE	FORM	SEGMENTS	COGID	MORPHEME
EasternLuobuohe	$\text{ʔa}^{024}\text{la}^{24}$	$\text{ʔa}^{024}\text{la}^{24}$	$\text{ʔ a}^{0/2} + \text{ɿ a}^{24}$	1	prefix moon
NortheastYunnanChuanqiandian	ɿi^{33}	ɿi^{33}	ɿ i^{33}	1	moon
NorthernQiandong	ɿha^{44}	$\text{ɿ}^h\text{a}^{44}$	$\text{ɿ}^h\text{a}^{44}$	1	moon
EasternQiandong	$\text{la}^{44}\text{la}^{44}$	$\text{la}^{44}\text{la}^{44}$	$\text{l a}^{44} + \text{l a}^{44}$	1	moon moon
WesternQiandong	$\text{pɔ}^{11}\text{la}^{33}$	$\text{pɔ}^{11}\text{la}^{33}$	$\text{p ɔ}^{11} + \text{l a}^{33}$	1	?/round moon
WesternBaheng	$\text{ʔa}^{034}\text{la}^{55}$	$\text{ʔa}^{034}\text{la}^{55}$	$\text{ʔ a}^{0/3} + \text{ɿ a}^{55}$	1	prefix moon
Younuo	$\text{kwan}^{13}\text{la}^{54}$	$\text{kwan}^{13}\text{la}^{54}$	$\text{k w a n}^{13} + \text{l a}^{54}$	1	light moon
She	ne^{33}	ne^{33}	n e^{33}	2	moon
Mien	ɿa^{24}	ɿa^{24}	ɿ a^{24}	1	moon

Table 3.2: An example of morpheme annotation using nine Hmongic and Mienic words 月亮 “moon”.

3.2 The Data Structure of a Bayesian Phylogenetic Analysis

The input data for a Bayesian phylogenetic analysis are either alphabetical or binary vectors. Alphabetical data were widely used in virology to infer a virus’ evolutionary tree. In the domain of Bayesian philolinguistics, we transform the cognate sets into binary vectors, in which 1 and 0 mark the presence and the absence of a cognate set, to infer the internal structure of a language family. Table 3.3 presents the binary matrix derived from the cognate sets provided in Table 3.2. Each language variety is represented by a binary vector.

The first column in each row is the language variety, also known as taxa in the domain of Bayesian phylogenetic analysis. The other columns use 1 or 0 to mark whether a word’s cognate set is found in the language or not. However, as explained above, the words in Table 3.2 are related on the morpheme level instead of on the word level. Therefore, the COGID column is in fact misleading, and has overly simplified the words’ relatedness.

TAXA	MOON-1	MOON-2
EasternLuobuohe	1	0
NortheastYunnanChuanqiandian	1	0
NorthernQiandong	1	0
EasternQiandong	1	0
WesternQiandong	1	0
WesternBaheng	1	0
Younuo	1	0
She	0	1
Mien	1	0

Table 3.3: Transforming the cognate sets into binary vectors. The column names are presented in gray because the names are not displayed in the actual input data for the Bayesian phylogeny software.

Chapter 2 introduced the importance of using partial cognates in the study of MSEA lexical data sets or languages in which compound words are prevalent. The column “COGIDS” in Table 3.4 represents an attempt to identify the partial cognates among the nine Hmongic and Mienic words. We were able to identify six partial cognates from the data. Although the partial cognates bring the links among words into sharper focus, a method that converts partial cognates into a binary matrix that can then be treated as the input data for a Bayesian phylogenetic analysis is

lacking.

LANGUAGE ID	VALUE	FORM	SEGMENTS	COGID	COGIDS	MORPHEME
EasternLuobuohe	$\text{ʔa}^{02}\text{la}^{24}$	$\text{ʔa}^{02}\text{la}^{24}$	$\text{ʔ a}^{0/2} + \text{ʔ a}^{24}$	1	2 1	prefix moon
NortheastYunnanChuanqiandian	ʔi^{33}	ʔi^{33}	ʔ i^{33}	1	1	moon
NorthernQiandong	ʔha^{44}	ʔha^{44}	ʔ h a^{44}	1	1	moon
EasternQiandong	$\text{la}^{44}\text{la}^{44}$	$\text{la}^{44}\text{la}^{44}$	$\text{l a}^{44} + \text{l a}^{44}$	1	1 3	moon moon
WesternQiandong	$\text{pɔ}^{11}\text{la}^{33}$	$\text{pɔ}^{11}\text{la}^{33}$	$\text{p ɔ}^{11} + \text{l a}^{33}$	1	4 1	?/round moon
WesternBaheng	$\text{ʔa}^{03}\text{la}^{55}$	$\text{ʔa}^{03}\text{la}^{55}$	$\text{ʔ a}^{0/3} + \text{ʔ a}^{55}$	1	2 1	prefix moon
Younuo	$\text{kwan}^{13}\text{la}^{54}$	$\text{kwan}^{13}\text{la}^{54}$	$\text{k w a n}^{13} + \text{l a}^{54}$	1	5 1	?/light moon
She	ne^{33}	ne^{33}	n e^{33}	2	6	moon
Mien	ʔa^{24}	ʔa^{24}	ʔ a^{24}	1	1	moon

Table 3.4: Nine Hmongic and Mienic words 月亮 “moon” are used to illustrate the cognate sets, partial cognate sets, and morpheme annotation.

For the transformation, we attempted to simply list all the partial cognates in columns in a binary matrix (see Table 3.5). In this case, we can see that all the languages are represented by binary vectors that considered all the partial cognates we had identified. This simple transformation can easily be accomplished using a Python script. A binary matrix can be generated automatically with the help of computers. However, we quickly realized that ignoring the linguistic perspective may introduce a significant amount of unwanted noise. For example, the duplicated morphemes do not need to be listed in the matrix because the duplication process does not involve sound changes. The changes only occur on a semantic level to address the fineness of an item.

TAXA	MOON-1	MOON-2	PREFIX	MOON-3	?/round	?/light
EasternLuobuohe	1	0	1	0	0	0
NortheastYunnanChuanqiandian	1	0	0	0	0	0
NorthernQiandong	1	0	0	0	0	0
EasternQiandong	1	0	0	1	0	0
WesternQiandong	1	0	0	0	1	0
WesternBaheng	1	0	1	0	0	0
Younuo	1	0	0	0	0	1
She	0	1	0	0	0	0
Mien	1	0	0	0	0	0

Table 3.5: Transforming the partial cognate sets into binary vectors. The column names are presented in gray because the names are not displayed in the actual input data for the Bayesian phylogeny software.

Another example that would not work in this straightforward approach is the versatility of morphemes in MSEA languages. We took a frequently encountered morpheme 子 $z\check{i}$ “child” in Chinese as an example. The morpheme originally meant “child”. The morpheme then developed multiple extended meanings in ancient times to describe a person’s career; for example, 夫子 $f\bar{u}$ $z\check{i}$ “teacher”. The morpheme can also function like a suffix without a concrete meaning, as in 妻子 $q\bar{i}$ $z\check{i}$ “wife”. The reason that the morpheme 子 $z\check{i}$ became a suffix is that Sinitic languages underwent a series of changes from monomorphemic words to multimorphemic words. Using the most straightforward approach means that all the morphemes 子 $z\check{i}$ will spread into different columns. However, the morphemes 子 $z\check{i}$ are derived from the same proto-form $*tsə?$ (Baxter and Sagart, 2014), and the cognates do not reflect the underlying semantic changes. From a mathematical perspective, these columns are duplicated columns that provide no new information

for the model but increase the number of columns; hence, they are a waste of computing power. Merging all the morphemes 子 $z\check{i}$ into one column to avoid duplicated columns is also not a recommended approach. The underlying language change mechanism is neglected in this overly simplified approach.

We took the basic concept of “gene weighting” from bioinformatics studies to form a “morpheme weighting” concept. The semantic meanings of compound words may be determined by all the parts of a morpheme, or a morpheme may play a more significant role in determining the meaning in other morphemes. In our paper, we further explained that the morpheme that contributed most to a lexical item’s meaning was the salient morpheme. We can avoid the problems stated above by selecting the salient morphemes. We took the concept of determining morpheme saliency a step further by using a computer-assisted method to create a binary matrix for a Bayesian phylogenetic analysis that accounts for both language and mathematical aspects.

Section 3.4 introduces our ideas about morpheme annotation and the methods we developed to convert partial cognates into full cognates that could then be used as input data for a Bayesian phylogenetic analysis.

3.3 Author Contributions

MSW and JML initiated the study. MSW and JML developed and implemented the annotation scheme. MSW and JML wrote the manuscript. All authors agree with the final version of the manuscript.

3.4 Second Paper

The article is freely available as an open access publication under the CC BY-NC 4.0 license. It was published in the *Journal of Language Dynamics and Change* in January 2023 (Wu and List, 2023).



BRILL

LANGUAGE DYNAMICS AND CHANGE (2023) 1–37



brill.com/ldc

Annotating cognates in phylogenetic studies of Southeast Asian languages

Mei-Shin Wu | ORCID: 0000-0001-6544-1163

Department of Linguistic and Cultural Evolution, Max Planck Institute
for Evolutionary Anthropology, Leipzig, Germany
wu@shh.mpg.de

Johann-Mattis List | ORCID: 0000-0003-2133-8919

Department of Linguistic and Cultural Evolution, Max Planck Institute
for Evolutionary Anthropology, Leipzig, Germany; Chair of Multilingual
Computational Linguistics, University of Passau, Passau, Germany
Corresponding author
mattis.list@lingpy.org

Abstract

Compounding and derivation are frequent in many language families. As a consequence, words in different languages are often only partially cognate, sharing some but not all morphemes. While partial cognates do not constitute a problem for the phonological reconstruction of individual morphemes, they are problematic for phylogenetic reconstruction based on comparative word lists. We review current practices of preparing cognate-coded word lists and develop new approaches that make the process of cognate annotation more transparent. Comparing four methods by which partial cognate judgments can be converted to cognate judgments for whole words on a newly annotated data set of 19 Chinese dialect varieties, we find that the choice of conversion method has an impact on the inferred tree topologies that cannot be ignored. We conclude that scholars should take great care with cognate judgments in languages in which compounding and derivation are frequent and recommend always assigning cognates transparently.

Keywords

phylogenetic reconstruction – Chinese dialects – Southeast Asian languages –
cognate annotation – partial cognates

1 Introduction

Computational phylogenetic methods in historical linguistics have been gaining popularity of late, and many studies on a diverse range of language families have been published (Gray et al., 2009; Grollemund et al., 2015; Lee and Hasegawa, 2011; Sagart et al., 2019). While there were quite a few studies criticizing the new quantitative studies in the beginning (Donohue et al., 2012; Geisler and List, 2010; Holm, 2007), the criticisms have not been raised again in recent years, although some of the major problems discussed in the earlier literature have not yet been addressed. Among these is the problem of cognate coding, the representation of cognate words in lexical data sets. Specifically with respect to the coding of partial cognates, not many attempts have been made to address the problem, although there are many language families in which partial cognate relations are frequent due to compounding and derivation.

In order to illustrate this problem, consider the cognate judgments by Koli-pakam et al. (2018) in Table 1. The authors use strings in the column labeled “Cognate” in order to indicate which word forms they assign to the same cognate set. While this procedure of assigning entire words to cognate sets is common in phylogenetic studies and rarely questioned, a closer investigation of the words assigned to the same cognate set shows that—at least for people who are not experts in Dravidian historical linguistics—is not necessarily easy to understand *where* the words in question are actually cognate. Comparing, for example, word forms like Kota [kanʈiko] with Kurukh [kʰajka], it is obvious that the words are not cognate in their entirety, but since the authors did not provide a morphological analysis, it is not possible for us to see *where* the words are cognate after all, or—more importantly—upon which part of the words the authors base their cognate decisions.

While the major issue of this type that arises in the analysis of Dravidian languages results from processes of derivation, and surfaces in cases where words from different languages share similar roots while the derivational suffixes are not necessarily cognate, in other language families, specifically in Southeast Asia and South America, the assignment of words to cognate sets is often made more complex by processes of compounding. Since scholars usually rely on the identification of shared lexical roots in order to assign word forms from differ-

TABLE 1 The word forms of *dry* in a data set of Dravidian etymologies

Variety	Form	Cognate
Tamil	ularnta	dry-A
Telugu	eṇḍu	dry-C
Kota	kaṇṭiko	dry-D
Kurukh	k ^h ajka	dry-D
Tamil	kaindadə	dry-D
Malto	a:ika:	dry-D
Brahui	ba:run	dry-E
Gondi	vaṭṭa	dry-E
Kannada	battida	dry-E
Kannada	oṇagidu	dry-F

KOLIPAKAM ET AL., 2018

ent languages to one and the same cognate set, the specific motivation underlying compounds can make it quite challenging to select one part of a compound over the other. In the Chinese dialects, for example, the concept ‘to swim’ can be expressed by different complex forms, such as Xīān *fú-shuǐ* [fu²⁴-fei⁵³] 浮水 (lit. ‘float-water’), Chángshā *wán-shuǐ* [wan¹³-cʰei⁴¹] 玩水 (lit. ‘play-water’), or Běijīng *yóu-shuǐ* [jou³⁵-ʃwei²¹³] 游水 (lit. ‘wander-water’). While all of these verbs share cognate word forms for ‘water,’ as well as similar motivations, insofar as they express the concept ‘to swim’ by referring to a concrete action that takes place in water, they differ in the word forms that express the action. From one perspective, one could therefore say that none of the three word forms are cognate, since they differ in the main verbs of the phrase, but from another perspective, one might equally argue that the motivation across these varieties is still quite similar, since many languages use a dedicated word form to express the concept ‘to swim’ or make use of different patterns. No matter how one decides, it becomes clear from this example that the cognate judgment is not based on the comparison of cognate relations between entire word forms, but rather depends on assumptions regarding the underlying motivation and a—usually—implicit judgment regarding those parts of a morphologically complex word which scholars consider as representative or salient with respect to the evolutionary process they investigate.

In the concrete practice of phonological reconstruction, scholars often avoid talking about complex words by shifting the object of comparison from the

4

WU AND LIST

TABLE 2 Partial cognate relations among words for ‘head’ in six Tupían languages

Variety	Form	Segments	Morphemes	Partial cognates
Akuntsu	anam	a + n ã m	ROUND ?	1 2
Amanaye	aki	a + k i	ROUND BONE	1 3
Amondawa	akaŋ	a + k a ŋ	ROUND BONE	1 3
Awetí	?aput	? a p + u t	HAIR ?	4 5
Arikem	a	a	ROUND	1
Cinta-Larga	antar	a n t a r	HEAD	6

TAKEN FROM THE TUPÍAN LEXICAL DATABASE, [HTTPS://TULAR.CLLD.ORG/](https://tular.clld.org/parameters/179)
PARAMETERS/179

word to the morpheme. This practice is especially pervasive in the reconstruction of Southeast Asian languages (Mann, 1998; Matisoff, 2003; Ratliff, 2010). In the practice of phylogenetic reconstruction—which typically starts from a list of concepts which are then translated in the target languages before cognate sets inside a given concept slot are identified—complex words cannot be easily ignored. As an example, consider the words for ‘head’ in Tupían languages (South America) in Table 2, taken from the Tupían Lexical Database (version 0.11; Ferraz Gerardi et al., 2021). Here, the authors follow Hill and List (2017) and Schweikhard and List (2020) in annotating cognates on the level of the morpheme accompanied by so-called morpheme glosses, which give hints on the lexical motivation underlying the formation of complex words. As can be seen from the data in the table, there are cases in which ‘head’ is motivated as a compound involving ‘round’ and ‘bone,’ but language varieties differ with respect to the details. There are also a case in which ‘head’ is rather interpreted as a simplex word. While assigning cognates on the level of morphemes can again be done in a mostly straightforward manner, it is far from obvious how cognate judgments pertaining to the whole word forms in this example should be derived. Should one assign all words which show the root glossed as ROUND in the example to the same cognate set, should one rather insist that words should be cognate with respect to all of their parts, or should one decide on a case-by-case basis?

Given the general importance of handling morphologically complex words in phylogenetic studies in historical linguistics, and the particular pervasiveness of morphologically complex words in Southeast Asian languages, we have carried out a detailed case study of the impact which different coding practices can have on phylogenies reconstructed from Chinese dialect data. In

the following, we discuss the problem of handling morphologically complex words when assigning words to cognate sets in more detail, proposing ways to increase the transparency of cognate coding (Section 2). We then present the results of a case study on Chinese dialect evolution in which we carry out a detailed comparison of different coding schemes and present simple but efficient data exploration methods that help scholars to identify those parts of their data where morphologically complex words could cause problems (Section 3). Finally, we discuss our findings (Section 4) and propose some ideas for future work (Section 5).

2 Increasing the transparency of cognate annotation

At the moment, cognate annotation in Southeast Asian languages faces two extremes. One extreme, which is the data model underlying many etymological studies, takes the (unbound) morpheme as a basic unit—ignoring words completely as linguistic units—and assembles cognate sets of morphemes without storing a reference to the words from which these were taken. As an example for this practice, consider the reconstruction of Hmong-Mien proto-forms in Ratliff (2010) and of Proto-Tibeto-Burman proto-forms in Matisoff (2003). In both cases, no full words are reconstructed, but only individual morphemes which may have complex words as reflexes in individual languages; these are, however, often not listed as such. The alternative extreme can be found in phylogenetic approaches, where words are traditionally taken as the basic units of comparison. Here, scholars assemble translational equivalents for a fixed list of basic concepts and then assign these words to cognate sets, without making explicit how partial cognates are handled.

Recent work concentrating on computer-assisted approaches to historical language comparison has shown that the first extreme can be avoided when starting from a careful annotation of partial cognates in comparative word lists (Wu et al., 2020). Instead of picking cognate morphemes from the literature, the new workflow not only allows researchers to maintain the link between the original words in which the morphemes occur and the morphemes themselves, but even offers convenient ways to inspect sound correspondence patterns (List, 2019) and search for partial colexifications (Hill and List, 2017).

What has *not* been sufficiently solved so far, however, is the question of how to deal with the annotation of cognate sets for the purpose of phylogenetic reconstruction. Here, the main problem is how to derive cognate judgments for full words when words are only partially related. In the following, we will discuss some general ideas regarding the annotation of cognate sets in word lists

for the purpose of phylogenetic reconstruction studies and then share some specific recommendations for concrete issues.

2.1 General ideas

When assembling comparative word lists for the purpose of phylogenetic reconstruction, the major problem imposed by language families in which partial cognacy is frequent is that it often becomes very difficult to find clear-cut criteria to assign words to cognate sets. In abstract terms, if one language expresses a concept 'X' with a compound word $a-b$ and another language expresses the same concept with a compound word $a-c$, there are two possibilities: one could either argue that the two words are to be judged as cognate, given that they have one cognate morpheme a in common; or one could argue that they are not cognate, given that they differ due to their respective morphemes b and c , which are not cognate. The complexity increases when more words are brought to the comparison and can easily lead to cases where the decision to assign all words which share at least one common morpheme to the same cognate set yields situations in which our hypothetical word $a-b$ would be cognate with $a-c$ and $a-c$ would be cognate with $d-c$, but $d-c$ would no longer share any common element with $a-b$.

The two most straightforward approaches to assigning words to cognate sets when their partial cognate sets are known have been called "strict" and "loose" cognate coding in previous work (List, 2016; List et al., 2016). In the strict case, only those words which are cognate with respect to all of their morphemes are assigned to the same cognate set. An example for this coding is the study on Chinese dialect evolution by Hamed and Wang (2006). In the loose case, a network of all words is constructed in which words correspond to nodes and links between nodes are drawn whenever two words share at least one cognate morpheme. After the network has been constructed, all words that belong to a connected component in the network are assigned to the same cognate set (Hill and List, 2017). An example for this coding procedure can be found in the study by Satterthwaite-Phillips (2011). Each approach has its advantages and disadvantages. While strict coding may easily increase differences between language varieties, giving the incorrect impression that there is a huge amount of linguistic variation in a given language family, the loose coding practice is unsatisfying as it may easily result in cognate sets consisting of word pairs that do not have a single cognate morpheme in common.

Assuming that partial cognates have been identified, an additional way to code the data in phylogenetic analyses would consist in ignoring the word level and coding the partial cognate sets directly. This technique, however, would contradict the important criterion of character independence, since individ-

ual morpheme cognate sets have not been evolving alone, but together with the words in which they appear. Since character independence is one of the basic criteria upon which phylogenetic models are built, introducing character dependencies may not only impact phylogenetic reconstruction (Felsenstein, 1988: 446), it will also make the results extremely difficult to interpret, since we ultimately want to understand how whole words evolve during language evolution, not how certain morphemes are gained and lost.

In order to avoid counting words which do not share a single cognate morpheme as cognate, Sagart et al. (2019) annotate their cognate sets in such a way that all words assigned to the same cognate set must at least have one morpheme in common. While this coding practice is beyond doubt more principled than the strict or the loose coding practices mentioned before, it has the disadvantage that it cannot be automatically checked. Sagart et al. (2019) make use of alignment analyses in order to make sure that there is a common morpheme in large cognate sets, but since they do not mark partial cognates in their data, it is not trivial to check all of their codings automatically. As a result, it is possible to check the consistency of their cognate annotation, but it is not easy to do so, since one has to go manually through each entry.

It is never trivial to decide whether overall cognacy for a set of words should rely on the presence of one single morpheme shared by all words or the presence of several words. As an example, consider the concept ‘sun,’ which many Austronesian languages lexify as ‘eye of the day,’ with the form for ‘day’ often being equivalent to the original word for ‘sun’ (Starostin, 2013: 121–123). Should we say that in a language which retains the original word for ‘sun’ this is cognate with a word in a language which shows the motivation ‘eye of the sun/day,’ or should we rather say that the latter is an innovation and reflects a clear case of lexical replacement? We think that this question cannot be clearly answered, but depends on the language family in question and our knowledge about it. The problem can therefore not be resolved by a computational approach alone.

While it is not possible to design a straightforward algorithm that would make the cognacy decisions in our place, it is, however, possible to insist on a more explicit *annotation* of lexical cognacy data that would reflect the individual decisions on cognacy taken by individual scholars. The solution we propose for this task is to make use of morpheme glosses, as shown above for the Tupían data in Table 2. Morpheme glosses were first proposed by Hill and List (2017) and further developed by Schweikhard and List (2020). We extend this work by adding a new aspect to the analysis, insofar as we mark the morpheme or the morphemes which we consider as *salient* with respect to the history of the word in question. Under saliency we understand the potential of one or more morphemes to reflect the major evolutionary processes of the words in which they occur.

TABLE 3 Identifying salient morphemes in partial cognates

Variety	Segments	Morphemes	Partial cognates	Analysis 1	Analysis 2
Akuntsu	a + n ã m	ROUND ?	1 2	1	1
Amanaye	a + k i	ROUND BONE	1 3	1	2
Amondawa	a + k a ŋ	ROUND BONE	1 3	1	2
Awetí	? a p + u t	HAIR ?	4 5	2	3
Arikem	a	ROUND	1	1	4
Cinta-Larga	a n t a r	HEAD	6	3	5

Analyses 1 and 2 show two ways to resolve the partial cognate relations to full cognates, the first one taking ROUND to be the sole salient morpheme, while the second one identifies ROUND and BONE as salient morphemes.

As an example, consider the words for ‘head’ in Tupían languages, which can be roughly divided into those words that denote head directly, such as Cinta-Larga [antar], words that involve a morpheme for ‘hair,’ such as Awetí [ʔap-ut], and words that contain a morpheme that means ‘round,’ such as Akuntsu [a-nãm] (with [a] glossed as ‘round’). One potential analysis of these partial cognates would be to take ‘round’ as the salient morpheme and to assume that it reflects an innovation in the language family, which was later diversified, leading to various subtypes that can or should be ignored in a phylogenetic analysis. Another possibility would be to say that the specific combination of ‘round’ and ‘bone’ should be treated as the major innovation. In this case, Amanaye [a-ki] and Amondawa [a-kaŋ] would reflect one common innovation and therefore be treated as one cognate set, while the other words that contain a reflex of ‘round’ but no reflex of ‘bone’ would be kept apart. Table 3 illustrates the consequences of these two decisions regarding the saliency of the morphemes with respect to the evolutionary history of their words.

This idea of marking those morphemes in the morpheme glosses which one identifies as representative for the word history can be seen as a less restricted variant of the aforementioned strict conversion of partial cognates into cognate judgments on whole words. While the strict conversion takes all morphemes in a given word as equally important, our proposal to annotate which morphemes are salient and which are not allows scholars to exclude specific morpheme cognates from the equation. As a result, scholars can, for example, argue that a certain suffix occurs so frequently in a given data set that it does not play a significant role in deciding whether a word that has the suffix should be considered cognate with a word that lacks the suffix.

TABLE 4 Using morpheme glosses to annotate semantic motivation structures for words denoting ‘hatchet’ in six Mienic varieties

Variety	Subgroup	Form	Segments	Morpheme glosses	Cognates
Daping	Zao Min	hɔŋ ⁵³ dziu ²²	h ɔ ŋ ⁵³ + dz j u ²²	firewood knife	1 2
Dongshan	Biao Mon	tsaŋ ³¹ du ⁴²	ts a ŋ ³¹ + d̥ u ⁴²	firewood knife	1 2
Jiangdi	Iu Mien	dzu ¹² ŋau ³³	dz u ¹² + ŋ au ³³	knife bent	2 3
Liangzi	Kim Mun	d̥u ²² ŋau ³³	d̥ u ²² + ŋ au ³³	knife bent	2 3
Luoxiang	Iu Mien	d̥u ²² ŋau ³⁵	d̥ u ²² + ŋ au ³⁵	knife bent	2 3
Miaoziyuan	Iu Mien	dzəu ²¹ ŋau ³³	dz əu ²¹ + ŋ au ³³	knife bent	2 3

ORIGINAL DATA FROM MÁO, 2004

Morpheme glosses are a free annotation form that serves to describe the semantic motivation structure of a given word. The term “motivation” is based on Koch (2001) and is used by Hill and List (2017) and Schweikhard and List (2020) to denote the semantics underlying word formation processes. As an example, consider Mandarin Chinese *shù-pí* 树皮 ‘bark (of tree)’, which consists of the two morphemes *shù* 树 ‘tree’ and *pí* 皮 ‘skin.’ The semantic motivation underlying the compound is thus the metaphorical use of ‘skin’ to denote the cover of trees. Hill and List (2017) indicate these motivation structures in their tabular word list data with the help of an extra column in which individual morphemes of multimorphemic words are glossed.

As an example for this annotation practice, consider the example of words denoting ‘hatchet’ in six Mienic varieties (original data taken from Máo, 2004) given in Table 4. In this table, we can observe three distinct morphemes from which all six words are built. All words share one morpheme that means ‘knife’ in isolation (colored in red in the table), but in Daping and Dongshan, the reflexes *dziu*²² and *du*⁴² appear at the end of the words, while they appear at the beginning in the other four varieties. The first morphemes in Daping and Dongshan are reflexes of Proto-Hmong-Mien *dzaŋ^A ‘firewood’ in the reconstruction of Ratliff (2010: 254), and the semantic motivation of the words in the two varieties is ‘firewood-knife,’ indicating that a hatchet is a specific kind of knife predominantly used for the preparation of firewood. In the remaining four varieties, where the morpheme for ‘knife’ appears at the beginning of the word, the second morpheme can be translated as ‘bent, crooked’ in isolation. Since most Mienic languages place the modifier after the modified, the semantic motivation for ‘hatchet’ is ‘bent knife,’ that is, a knife that has a bent form.

TABLE 5 An illustration of using morpheme glosses to derive cognate sets for whole words from partial cognate sets

Variety	Segments	Morpheme glosses	Partial	Strict	Loose	Salient
Western Xiangxi	q o ³⁵ + tɕ ^h i ³⁵	_prefix/Q belly/A	1 2	1	1	1
Eastern Xiangxi	k i ⁰³ + t ^h i ⁵³	_prefix/K belly/A	3 2	2	1	1
Western Baheng	? a ⁰³ + ŋ ŋ ³¹	_prefix/A belly/B	4 5	3	1	2
Numao	ŋ u ŋ ¹³	belly/B	5	4	1	2
Chuanqiandian (NEY)	? a ⁵⁵ + tɕ ^h au ⁵⁵	_prefix/A belly/A	4 2	5	1	1

By marking non-salient morphemes with a preceding underscore , we can explicitly select only those partial cognate sets relevant for the assignment of word cognates, arriving at a transparent procedure for the annotation of cognate judgments for full words. The data shows the words for ‘belly’ in five Hmongic languages.

DATA TAKEN FROM CHÉN, 2012: 599

Once morpheme glosses have been added to a data set, the annotation of salient morphemes, that is, morphemes one deems representative for the whole history of the words, can be done in a very straightforward way by simply indicating the saliency along with the morpheme glosses. In our concrete annotation, this means that we add an underscore in front of each morpheme gloss which we consider as *not* salient. When later converting partial cognates to “full” cognates, we only extract those cognate sets whose morpheme glosses have been annotated as salient and then use the strict conversion procedure on these selected cognate sets.

As an example for this procedure, consider the words for ‘belly’ in five Hmongic languages in Table 5 (Chén, 2012: 599). All words show the same basic structure of being composed of a prefix with synchronically untransparent semantics and a main morpheme with the core meaning ‘belly.’ As can be seen from our partial cognate annotation (provided in the column “Partial”), we identify three distinct prefixes and two distinct morphemes for ‘belly,’ one going back to Proto-Hmong-Mien *ch̥ɛi^A in the reconstruction of Ratliff (2010), the other of an origin unknown to us. When computing strict cognate sets from the partial cognates, all words will be placed into distinct cognate sets, since none of the words coincide in all their morphemes. When using the procedure of loose cognate annotation, all words would be placed into the same cognate set, since they all form one big connected component, in which words containing a reflex of Proto-Hmong-Mien *ch̥ɛi^A, labeled belly/A in our morpheme glosses, are connected to the words with the reflex labeled belly/B via the prefix prefix/A, shared between Western Baheng and Chuanqiandian. Our procedure of salient cognate coding, on the other hand, deliberately ignores

the prefixes—given that their presence or absence provides little evidence for the historical development of the words on which they occur, but rather points to largely language-specific processes of productive prefixation that are not well understood—and thus divides the five words neatly into two cognate sets, depending on the basic morpheme used to express the meaning of ‘belly.’

2.2 *Specific ideas*

The schema as presented in the previous section relies entirely on human judgment, and it is difficult—at least for the time being—to think of an automated approach to approximate human judgments. The reason is not the impossibility of finding alternatives to the strict and the loose practice of converting partial to full word cognate sets. As we will show in the following sections, we can easily implement a method that accounts for the cognate coding practiced by Sagart et al. (2019). The problem is that it is often not clear what should count as the best solution and that there is no real way to tell based on the data alone. In the following, we will nevertheless try to provide some general criteria that may help scholars in arriving at decisions in particularly difficult situations.

There are three major caveats when deciding about full word cognacy in multilingual word lists. First, when annotating cognates, scholars should try to avoid coding as cognates those cases that are highly likely to have evolved as a result of parallel independent evolution (i.e., avoid homoplasy). Second, one should try to make sure that the characters, that is, the cognate sets, are maximally independent (i.e., minimize character dependency). Third, one should make sure to identify cases of free or pragmatically conditioned synchronic variation and control for them systematically (i.e., control variation).

As an example for the first problem, that of parallel independent evolution or homoplasy, consider cases of lexical motivation in compounding (Koch, 2001). Words for ‘tears’ in Hmong-Mien languages are a good example, since as in many Southeast Asian languages, ‘tears’ tends to be expressed through a compound, of which one part in isolation is related to a word that means or originally meant ‘water’ (consider Mandarin Chinese *lèi-shuǐ* 泪水 ‘tears,’ which can be glossed as ‘tears-water’). In the Hmong-Mien languages, the other part of the compound is typically the same as the word for ‘eye,’ and the lexical motivation of ‘tears’ can thus be described as the ‘water’ of the ‘eye’ (Chén, 2012: 609). Unlike most Chinese dialect varieties, which tend to place the modifier before the modified in compounds, Hmong-Mien languages typically use the opposite order (‘water-eye’ instead of ‘eye-water’). In Sinitic, there are some exceptions of this rule in the south, which scholars tend to attribute to influence from the Hmong-Mien languages (Vittrant and Watkins, 2019), but we can find the opposite influence in some Hmong-Mien varieties as well. As a

result, some Hmong-Mien languages lexify ‘tears’ as ‘eye-water,’ such as Zao Min *mai*⁵³.*m*²⁴ (*mai*⁵³ means ‘eye’ in isolation, going back to Proto-Hmong-Mien *mɤejH; and *m*²⁴ means ‘water,’ going back to Proto-Hmong-Mien *ɣəm; see Chén, 2012; Ratliff, 2010), while the majority have a compound ‘water-eye,’ such as Western Qiangdong *ɣeu*⁴⁴ *me*²² (*ɣeu*⁴⁴ is ‘water’ and *me*²² is ‘eye’; Chén, 2012). Note that the morphemes in the words in Zao Min and Western Qiangdong both go back to the same proto-forms, even if it is quite likely that the word for ‘eye’ was borrowed from Chinese. While it is trivial (despite the complex sound correspondences) to identify the morphemes in both words as cognate, it is far from trivial to decide on the cognacy of both words. One could assume that Proto-Hmong-Mien once had a compound ‘water-eye’ and that this compound was inherited by both Zao Min and Western Qiangdong, and that the lexical motivation of the compound did not lose its transparency until Zao Min began to reverse the order of compound constituents from modified-modifier to modifier-modified, possibly under the influence of Chinese dialect varieties. The reversed word for ‘tears’ thus reflects some global innovation in the language which affected a large part of its lexicon. Another possibility, however, is to assume that the motivation underlying words for ‘tears’ in the Hmong-Mien languages is so obvious and general that we can easily assume that it could recur independently throughout the history of many languages. As a result, it would be wrong to say that the words as such are cognate, since one would assume that they were coined independently and therefore do not reflect shared innovations in the language family. With the knowledge we have at our disposal, we consider this case as undecidable. As a result, it seems best to ignore items like ‘tears’ when applying phylogenetic reconstruction methods to the Hmong-Mien language family in order to make sure that the phylogenetic signal is not contaminated by instances of parallel evolution.

As an example for the problem of character dependence, consider the analytical derivation of plural forms for personal pronouns in many Southeast Asian languages. While plural forms for personal pronouns tend to have an independent (suppletive) form in most Indo-European languages (compare German *ich* ‘I’ vs. *wir* ‘we,’ *du* ‘thou’ vs. *ihr* ‘you [pl.]’), many Southeast Asian languages derive plural forms from the singular forms by means of suffixation (Mandarin *wǒ* 我 ‘I’ vs. *wǒ-men* 我们 ‘we,’ *nǐ* 你 ‘thou’ vs. *nǐ-men* 你们 ‘you [pl.]’). As a result, the plural form can be regularly predicted from the singular form for most languages in which the plural is built analytically. However, many questionnaires for phylogenetic reconstruction in linguistics contain concepts for singular and plural personal pronouns, and so in these languages the corresponding characters for ‘I,’ ‘thou,’ ‘we,’ and ‘you (pl.)’ can no longer be considered to have evolved independently, since singular pronouns are reused to form

the plural pronouns and all plural pronouns tend to share the same affix that derives the plural meaning.

When encountering these processes across all languages in a given data set, the only consequent way to deal with the cognate assignments is to code each morpheme only *once*, which would mean that one needs to modify the underlying questionnaire in such a way that only singular forms are used as the base forms, while plural forms of personal pronouns are collapsed into one single ‘plural’ category. If, however, not all plural forms are constructed analytically—as is the case for the Hmong-Mien languages, where some varieties have a regular plural suffix, similar to Mandarin Chinese (e.g., Jiongnai, a Hmongic language, has *wa*³¹ ‘I’ vs. *wa*³¹ *klun*⁵³ ‘we’; Iu Mien, a Mienic language, has *ze*³³ ‘I’ vs. *ze*³³ *wo*³³ ‘we’), but some also have suppletive forms (Eastern Xiangxi, Hmongic, *m*³¹ ‘thou’ vs. *ma*⁵³ ‘you [pl.]’)—we recommend excluding plural forms directly from the analysis, since the independency of the characters cannot be guaranteed.

As an example for the problem of controlling variation, consider the phenomenon of affixation in the Hmong-Mien language family. In many Hmong-Mien languages, one finds a certain number of productive prefixes or suffixes which are typically used to derive nouns from a base form. Some of these derivations are mandatory, while some can be omitted, depending on the context. Thus, the word for ‘star’ in Xia’ao (Western Xiangxi, Hmongic branch of Hmong-Mien) will typically be elicited as *qa*⁰²-*sin*⁴⁴ (Chén, 2012: 145, 282), consisting of the prefix *qa*⁰²-, which derives inanimate nouns, and the noun *sin*⁴⁴, an early borrowing from Chinese *xīng* 星, which was pronounced as *seŋ* in the sixth century AD (Baxter, 1992). The use of the prefix, however, is not obligatory: it can be omitted, depending on the context (Chén, 2012: 145). When deriving cognate judgments for cases of this sort where free variation can be observed, we recommend first checking to ensure that the variation can be observed in all or most of the languages in a given sample, and if this is the case, excluding the longer forms from the data.

As we have tried to illustrate throughout this section: it is by no means trivial to deal with these questions, and we expect that the impact on phylogenies when adopting arbitrary solutions for cognate coding could be rather substantial. In order to address the problems in a straightforward manner, we suggest that scholars working with languages in which partial cognacy is a frequently recurring problem, resulting from abundant compounding and rich derivational processes, carry out a very close analysis of language-internal cognacy. Using morpheme glosses, it is possible to rigorously mark prefixes, suffixes, and the lexical motivation structures underlying compounds. Once this analysis has been carried out and partial cognates have been identified across languages

as well as language-internally, thus taking both words with the same meaning and words with different meanings into account, scholars can carefully check individual semantic slots and try to identify whether any of the three problems discussed in this section applies. If this turns out to be the case, one should: (a) ignore the concepts that are expressed by words that are suspicious of parallel evolution due to frequently recurring patterns of lexical motivation (avoid homoplasy); (b) try to identify the phylogenetically important alternations when dealing with problems of character dependency and re-code the data accordingly (minimize character dependency); and (c) carefully study how words vary when being used in different contexts in order to handle problems resulting from language-internal variation (control variation).

3 A case study on Chinese dialect history

In order to illustrate the problems resulting from cognate coding when working with language families in which compounding and derivation are frequent, we have prepared a case study on Chinese dialect history, based on a data set which we have coded, following the principles discussed in the previous section. In this section, we will first present how the original data set was lifted from its raw tabular version without cognate judgments to a standardized version in which partial cognates have been identified both across and inside language varieties, and how morpheme glosses were used to characterize the semantics of morphemes (Section 3.1). We will then show how the standardized version of the data allows us to automatically infer those cases which constitute a problem for phylogenetic analysis (Section 3.2) and finally report the results of this analysis, accompanied by individual examples from the data (Section 3.3). The annotated data set and a small collection of Python scripts used for the analysis are available as supplementary materials; scholars can use the scripts to investigate their own data sets.

3.1 Materials

The data set was originally published by Liú et al. (2007) and later digitized for this study by manually entering the data into text files. The data consists of 201 concepts translated into 19 Chinese dialect varieties (see Table 6) which provide at least one variety as a representative for each of the seven major subgroups proposed by Norman (1988: 181)—Mandarin (*Guānhuà*) 官话, Wú 吴语, Xiāng 湘语, Mǐn 闽语, Yuè 粤语, Gàn 赣语, and Hakka (*Kèjiā*) 客家—as well as one variety for each of the three subgroups which are often additionally proposed—Jīn 晋语, Píng huà 平话, and Huī 徽语 (Yan, 2006). In order to

TABLE 6 List of Chinese dialect varieties in our sample along with the subgroups they can be assigned to

Variety	Subgroup	Chinese name
Běijīng	Mandarin	北京
Chángshā	Xiāng	长沙
Chéngdū	Mandarin	成都
Fúzhōu	Mǐn	福州
Guìlín	Píngguà	桂林
Guǎngzhōu	Yuè	广州
Hāěrbīn	Mandarin	哈尔滨
Jìxī	Huī	绩溪
Jǐnán	Mandarin	济南
Lóudī	Xiāng	娄底
Méixiàn	Hakka	梅县
Nánchāng	Gàn	南昌
Nánjīng	Mandarin	南京
Róngchéng	Mandarin	荣成
Sūzhōu	Wú	苏州
Tàiyuán	Jìn	太原
Wēnzhōu	Wú	温州
Xī'ān	Mandarin	西安
Xiàmén	Mǐn	厦门

guarantee the comparability of our data set with other data sets, we linked the concept list to the Concepticon reference catalog (<https://concepticon.clld.org>; List, Tjuka et al., 2022) and the language varieties to Glottolog (<https://glottolog.org>; Hammarström et al., 2021); see the supplementary material.

In the raw data, the translations for each concept in each variety are given in phonetic transcription and in Chinese characters (Liú et al., 2007). The latter are frequently used by Chinese dialectologists in order to mark etymologically related morphemes across different dialects (*běnzì* 本字, literally ‘original characters’; see Mei, 1995). Although the Chinese character information on cognacy needs to be treated with some care, it is a good starting point for the annotation of cognate sets both across dialects and inside one and the same dialect.

Phonetic transcriptions in the original data set were standardized by converting the original transcriptions—which follow specific peculiarities as they are typically found in Sinitic varieties descriptions—to the transcriptions pro-

posed by the Cross-Linguistic Transcription Systems (CLTS, <https://clts.cld.org>; List et al., 2021; see Anderson et al., 2018, for details on the CLTS system). This reference catalog is one of the core components of the Cross-Linguistic Data Formats (CLDF, <https://cldf.cld.org>; Forkel et al., 2018). The CLTS system can be seen as a narrower version of the International Phonetic Alphabet insofar as it resolves several of its ambiguities. For the conversion and segmentation of the transcriptions, orthography profiles (Moran and Cysouw, 2018) were used and all individual transcriptions were later manually checked.

Partial cognate sets were first automatically added to the data by employing the Chinese character readings, and later systematically refined using the interactive web-based *EDICTOR* tool for the creation of etymological data sets (<https://digling.org/edictor>; List, 2017, 2021). Morpheme glosses, following Hill and List (2017) and Schweikhard and List (2020), were manually added for all morphemes, based on the previously inferred partial cognate sets. In order to facilitate the reuse of the data, we used the *CLDFBench* software package (Forkel and List, 2020) with the *Lexibank* plugin (List, Greenhill et al., 2022) to convert the data to the tabular standards proposed by the CLDF initiative. The entire data set contains a total of 4,302 words, with 65.6% of these being monosyllabic words and 34.4% polysyllabic words.

The original data set of Liú et al. (2007) often contains multiple translations for the same concept in the same variety, and this can easily influence the results of phylogenetic reconstruction approaches. We therefore carefully excluded some of the translations which reflect specific colloquial registers. Following standard practice in phylogenetic reconstruction in historical linguistics, we also made sure to mark known borrowings in the data, relying on our own knowledge of Chinese dialect history as well as cases of borrowings annotated in similar data sets (Sagart et al., 2019). All decisions about which items were excluded or marked as borrowings are transparently reflected in the data and can be inspected, criticized, and improved in future research.

3.2 *Methods*

In the following, we present a range of techniques that can be used to detect problems resulting from partial cognacy in phylogenetic reconstruction. Once these problems have been detected, they can be addressed by refining annotations or excluding concepts with high amounts of variation from an analysis.

3.2.1 Deriving full cognates from partial cognates

We have discussed different techniques of converting partial to full cognates in Section 2.1. While the strict and the loose conversion method are straightfor-

ward to implement and have been available as part of the LingPy software package (<https://lingpy.org>; List and Forkel, 2021) since 2016, the method employed by Sagart et al. (2019) has so far only been manually applied. Notwithstanding certain problems resulting from the proper handling of recurring suffixes, this method can be approximated by a greedy algorithm.

The algorithm we propose proceeds in two stages. In a first stage, we construct “fuzzy clusters” from all words in a given meaning slot by creating one cluster for each distinct morpheme (as indicated by the partial cognate identifier) in the selection. In a second stage, we order the clusters by size, starting from the largest cluster, and mark all words which contain the morpheme represented by this cluster as salient. We then iterate over the remaining clusters and remove all words which occurred in our first cluster from the remaining clusters.

As an example, consider four languages A, B, C, and D which express one word with two morphemes each: *a-b*, *a-c*, *a-d*, *d-c*. In our first stage, we assign the words to four clusters *a* (A, B, C), *b* (A), *c* (B, D), and *d* (C, D). Ordering them by size yields the order $a \rightarrow c \rightarrow d \rightarrow b$ or $a \rightarrow d \rightarrow c \rightarrow b$. Which order is the best cannot be determined automatically, so either can be used, but we use the first order for our illustration here. When iterating over the clusters, we start from cluster *a*, mark all words as salient (*a-b*, *a-c*, *a-d*), and remove the words with morpheme *a* from the remaining cluster. As a result, cluster *b* is empty, as it contains only one word with *a*, while *c* loses the word from language B and *d* loses the word from language C. The next cluster in our ordered list is *c*, which now contains only one member, the word from language D. Once the morpheme *c* is marked as salient, the word from language D is also removed from cluster *d*, leaving all words assigned exactly one salient morpheme. The method has been implemented as part of the LingRex Python library (version 1.3.0; List and Forkel, 2022).

The procedure should be undertaken with some care, since its greediness can easily lead to an overcounting of affixes. However, it has proven useful to us as we are able to preprocess a data set first and later correctly annotate it manually.

3.2.2 Identifying potential cases of homoplasy and character dependencies

It is challenging if not impossible for the time being to design algorithms that directly distinguish homoplasy from character dependence. However, we provide two evaluation methods to “flag” the concepts which may lead to different word cognate sets between different conversion methods and further influence the subsequent phylogenetic analysis.

The first method is based on the automated comparison of different methods for the conversion of partial to full cognate sets. This method works for all data sets in which partial cognate sets have been identified, regardless of whether partial cognates have been identified within meaning slots or cross-semantically. The approach is extremely straightforward. We first automatically compute strict cognates from the partial cognates in our data set and then compute loose cognates from the same data. In a second step, strict and loose cognate sets are systematically compared with the help of B-Cubed scores (Amigó et al., 2009), which are typically used to compare how well an automated cognate detection method performs in comparison to a gold standard (Hauer and Kondrak, 2011; List et al., 2017). B-Cubed scores come in the form of “precision,” “recall,” and their harmonic mean, the “F-score,” which ranges from 0 (completely different clusters) to 1 (identical clusters). List (2014) details the B-Cubed algorithm and the calculation is implemented in the LingPy Python library (List and Forkel, 2021). By ranking the concepts in a given data set according to the differences in the F-scores computed for strict and loose cognates, we can identify the extreme cases in which the conversion of partial to full cognates causes trouble. Using strict and loose cognate conversion is specifically useful in this context, since the approaches represent two extremes.

Our second evaluation method requires partial cognates to be consistently identified across meaning slots in a given data set. In contrast to the method based on cluster comparison, it systematically takes language-internal information into account. The method proceeds in two stages. In a first stage, we iterate over the word list and count for each distinct morpheme and each language in our data in how many concepts it recurs. In a second stage, we summarize the cross-semantic partial cognate statistics on the word level for each concept by first averaging the number of cross-semantic partial cognates for each individual word and then averaging the individual word scores for an entire meaning slot. The score for individual words starts from 1 (a cognate set occurs once in the data set for the given language) and has a theoretical maximum of the size of the concept list (a cognate set occurs in all words for a given language). We subtract 1 from this score in order to make sure that the score starts from zero. The resulting score thus ranges between 0 and the length of the concept list minus 1 and allows us to identify those concepts in which most cross-semantic partial cognates occur. Since the identification of cross-semantic partial cognates can be tedious, the method may not be available in the early stages of data curation. Once cross-semantic partial cognates have been identified, however, the method can be very helpful, since it accounts for cases in variation that might not be spotted by the method based on cluster comparison. Both methods have been implemented as part of the LingRex Python library (version 1.3.0; List and Forkel, 2022).

3.2.3 Annotating salient morphemes

Our methodology is oriented towards a computer-assisted as opposed to a pure computer-based workflow because we acknowledge the difficulty of identifying full cognates in comparative word lists automatically. This requires—in addition to providing code that may help to detect inconsistencies in the data—that we also discuss and test options to manually refine a data set that was computationally preprocessed. We have presented our main idea for the annotation of salient morphemes in partial cognate sets in Section 2.1. While this annotation can theoretically be done in a simple text file or with the help of a spreadsheet editor, we have used the web-based *EDICTOR* tool for the creation and curation of etymological data sets (<https://digling.org/edictor>, List, 2017; List, 2021); this tool has recently added a function that allows for an improved handling of morpheme glosses. Once partial cognates and morpheme glosses have been annotated, scholars can quickly mark whether individual morphemes are considered as “salient” with respect to the history of the languages in question, or not. To classify individual morphemes as salient or not, users simply have to right-click the morpheme gloss with the mouse in the *EDICTOR* interface. This will add or remove an initial underscore (which we use as a marker of non-salient morphemes in our code) to the respective morpheme gloss and also change its visual appearance by increasing the transparency.

Once a data set has been annotated in the form described here, the conversion of partial to full cognates can be done in a rather straightforward way. Our algorithm proceeds in two steps. In a first step, it iterates over all cognate sets and removes all those cognate sets which have been annotated as non-salient. In a second step, we use the remaining cognate sets to compute strict cognate sets, as discussed above. The *LingRex* package (List and Forkel, 2022) offers an automatic solution for the conversion into full cognates of partial cognates with salient morphemes indicated in morpheme glosses.

3.3 Results

We applied the methods described above to the newly compiled data set for Chinese dialect varieties in order to investigate to what degree an extensive number of partial cognates could have an impact on phylogenetic reconstruction analyses. In the following, we will discuss our experiments in detail. We start from our heuristics for the identification of concepts susceptible to high variation due to partial cognacy (Section 3.3.1) and discuss some examples where cognate codings differ, depending on the approach used to make cognacy judgments for entire words from partial cognates. We then carry out a systematic comparison of dialect distances resulting from different coding

practices (Section 3.3.2) and conclude by investigating how the coding practice influences the results of phylogenetic reconstruction analyses (Section 3.3.3).

3.3.1 Identifying concepts susceptible to high variation

The upper part of Table 7 shows the 10 concepts with the lowest B-Cubed F-scores, derived from the comparison of strict and loose partial cognates in the data set (the full table is provided in our supplementary material). As can be seen from the table, concepts with high variation mostly comprise certain nouns which tend to have a complex motivation structure in the Chinese dialect varieties ('knee,' 'neck,' 'wing,' etc.) a few complex verbs ('live,' 'swim'), as well as demonstrative pronouns ('here'), which tend to vary greatly among Chinese dialects. The lower part of the table shows 10 of the 100 examples in which F-scores reach 1.0, indicating that there is no difference between strictly and loosely converted cognate sets. Here, we find mostly those concepts which are expressed by monosyllabic words in the Chinese dialects, including specifically most adjectives ('yellow,' 'wet'), most basic verbs ('wash,' 'walk'), and some very basic nouns ('wind,' 'water'). All in all, these results are not surprising, but they prove the usefulness of our very simple approach to identify those cognate sets which could cause problems in later phylogenetic analyses.

The results of our test on cross-semantic partial cognates are given in Table 8, again showing the 10 concepts which showed the highest average number of colexifications per word and per concept slot in the upper part of the table and 10 concepts for which no colexifications could be identified throughout all words in the lower part. As can be seen from this table, the highest scoring concept is 'person,' typically expressed as *rén* 人 in Chinese. The word recurs in many words denoting specific kinds of persons, such as 'woman,' typically expressed as *nǚ-rén* 女人, or 'man,' typically expressed as *nán-rén* 男人. Additional concepts with high potential of being expressed by morphemes that are reused to express other concepts are 'water' 水, which often recurs in words for 'fruit' (*shuǐ-guǒ*, lit. 'water-fruit' 水果), and 'bark' whose lexical motivation is 'tree-skin' (*shù-pí* 树皮) in almost all Chinese dialect varieties. Looking at the cases with no cross-semantic partial cognates, it is difficult to find a clear pattern, apart from a tendency for these to be monosyllabic words, which will naturally decrease the chance of a word of showing at least one part which colexifies across the data under consideration.

All in all the results are not identical with the ones reported in Table 7, but they show some similar tendencies with respect to monosyllabicity. This similarity in the rankings of concepts can also be computed. Using the Kendall's τ correlation coefficient test, we find a weak negative association between the results of the two rankings (Kendall's τ coefficient = -0.25 , $p < 0.001$). The fact

TABLE 7 Upper and lower parts of the comparison of B-Cubed F-scores between loosely and strictly derived cognate sets

Concept	Chinese	Pinyin	F-score
breasts	奶子 乳房	<i>nǎi-zi rǔ-fáng</i>	0.35
live (alive)	活着 活的	<i>huó-zhe huó-de</i>	0.37
knee	膝盖 膝头	<i>xī-gài xī-tóu</i>	0.37
here	这里 这	<i>zhè-lǐ zhè</i>	0.39
woman	女人 女的	<i>nǚ-rén nǚ-de</i>	0.47
child	孩子 孩	<i>hái-zi hái</i>	0.49
nose	鼻子 鼻	<i>bí-zi bí</i>	0.49
rope	绳子 绳	<i>shéng-zi shéng</i>	0.5
sky	天空 天上	<i>tiān-kōng tiān-shàng</i>	0.5
claw	爪子 爪	<i>zhǎo-zi zhǎo</i>	0.51
...
turn	转	<i>zhuǎn</i>	1.00
two	二 兩	<i>èr liǎng</i>	1.00
walk	走 行	<i>zǒu xíng</i>	1.00
wash	洗	<i>xǐ</i>	1.00
water	水	<i>shuǐ</i>	1.00
wet	湿 潮	<i>shī cháo</i>	1.00
white	白	<i>bái</i>	1.00
wide	宽 阔	<i>kuān kuò</i>	1.00
wind	风	<i>fēng</i>	1.00
yellow	黄	<i>huáng</i>	1.00

The 10 concepts with the lowest B-Cubed F-scores are shown in the upper part of the table, and 10 of the concepts with the highest F-scores of 1.0 are shown in the lower part of the table. The column labeled “Chinese” shows the up to three of the most frequent exemplary reflexes in Chinese for the given concept slot; that labeled “Pinyin” shows the pronunciation in Mandarin Chinese using pinyin transliteration.

that the two tests only correlate weakly emphasizes how important it is to use both of them when investigating the potential impact of partial cognates on lexical phylogenies.

One can be tempted to assume that our concept of “morpheme saliency” might be replaced by some independent principle, such as, for example, the underlying dependency structure of compound words expressing a given concept. Following this line of argumentation, one could, for example, argue that

22

WU AND LIST

TABLE 8 Top 10 concepts with highest scores and 10 of the concepts with the lowest scores in the test on cross-semantic partial cognate statistics (overall ranking)

Concept	Chinese	Pinyin	Score
person	人	<i>rén</i>	2.47
hit	打 拍	<i>dǎ pāi</i>	1.95
old	老	<i>lǎo</i>	1.6
tree	树 树儿	<i>shù shù-ér</i>	1.53
water	水	<i>shuǐ</i>	1.32
bark	树皮	<i>shù-pí</i>	1.29
woman	女人 女的	<i>nǚ-rén nǚ-de</i>	1.17
man	男人 男的	<i>nán-rén nán-de</i>	1.16
fight	打架 相拍	<i>dǎ-jia xiàng-pāi</i>	1.08
we	我们 我竹固哩	<i>wǒ-men wǒ-zhú-gù-lǐ</i>	1.08
...
back	背 背脊	<i>bèi bèi-jǐ</i>	0
bad	坏 否	<i>huài fǒu</i>	0
because	因为 庸乎	<i>yīn-wéi yōng-hū</i>	0
bird	鸟 雀	<i>niǎo què</i>	0
bite	咬	<i>yǎo</i>	0
blood	血	<i>xuè</i>	0
blow	吹	<i>chuī</i>	0
burn	烧	<i>shāo</i>	0
cloud	云 云彩	<i>yún yún-cǎi</i>	0
count [noun]	数	<i>shù</i>	0

only heads should be considered as the salient morphemes in a word, or only modifiers. However, due to complexity of lexification processes, head-modifier structures of compounds barely reflect the pathways of lexical motivation. As an example, consider Table 9, where we show how concepts such as ‘moon’ and ‘woman’ are expressed in four Chinese dialect varieties in our sample along with the motivation structure underlying the words. The concept ‘moon’ is expressed as *yuè-liàng* 月亮, literally ‘moon-shine,’ in Mandarin Chinese, with 月 ‘moon’ being the modifier and 亮 ‘shine’ being the head. The concept ‘woman’ is expressed as *nǚ-rén* 女人, literally ‘woman-person,’ in Mandarin Chinese, with 女 ‘woman’ being the modifier and 人 ‘person’ being the head. When comparing how the concepts are reflected across the other varieties, we

TABLE 9 The concepts ‘moon’ and ‘woman’ and their inherent motivation structure in four Chinese dialects

Variety	Concept	Segments	Chinese	Morphemes
Běijīng	moon	ɥ ɛ ⁵¹ + l j a ŋ ⁰	月亮	moon <i>shine</i>
Jínán	moon	ɥ ɣ ²¹ + l j a ŋ ^{31 0}	月亮	moon <i>shine</i>
Wēnzhōu	moon	ɲ y ²¹ + k w ɔ ⁴⁴	月光	moon <i>ray</i>
Méixiàn	moon	ŋ j a t ⁵ + k w o ŋ ³³	月光	moon <i>ray</i>
Běijīng	woman	n y ²¹⁴ + z ɛ n ³⁵	女人	<i>female</i> person
Jínán	woman	ɲ y ⁴⁵ + z ẽ ⁵³	女人	<i>female</i> person
Wēnzhōu	woman	l ə ²⁴ + ɲ j a ŋ ³⁴¹ + k ^h a ⁴¹	老娘客	old <i>woman</i> guest
Méixiàn	woman	m oi ⁵³ + j e ⁰ + ŋ i n ¹¹	妹兒人	<i>sister</i> suffix person

The morphemes which we judge as salient in this context are marked with italic font.

can quickly see that the archaic varieties in the south of China (Wēnzhōu and Méixiàn) tend to express the concept for ‘moon’ as *yuè-guāng* 月光 ‘moon-ray,’ while more innovative Mandarin varieties (Běijīng and Jínán) show the Mandarin form 月亮 ‘moon-shine.’ In terms of the motivation underlying this process of lexical change, we therefore find 月, the modifier, as the stable part, while the head of the compound has changed and would therefore be treated as the salient morpheme in our annotation. Contrasting these cases with the expressions for ‘woman,’ we find another situation, with the Mandarin dialects showing the same form, and some southern dialects showing diverging motivations, like Méixiàn 妹兒人 *mèi-ér-rén*, ‘sister-suffix-person’ or Wēnzhōu 老娘客 *lǎo-niáng-kè*, ‘old-woman-guest.’ While the head stays stable in Méixiàn, we find an innovation with respect to the modifier in both southern varieties and would therefore annotate the modifier as the salient morpheme. This example shows that the saliency of a morpheme with respect to the history of the word in which the morpheme occurs cannot be determined from the dependency structure alone, although the dependency structure is of crucial importance when it comes to identifying the underlying motivation that led to the creation of a compound.

3.3.2 Cognate coding and language distances

Having shown that we can identify quite a few concepts in the Sinitic data in which compounding patterns are so complex that they make the conversion of partial into full cognate sets difficult, we wanted to analyze to what degree this may influence the computation of lexical distances between languages.

We therefore computed distance matrices, following classical lexicostatistical methodology (counting shared cognates per meaning slot) for both strictly and loosely converted cognate sets as well as for the two new approaches we introduced in Section 3.2, conversion by common morphemes and conversion by salient morphemes. In order to get a better impression on the theoretical impact which partial cognates can have on lexical distance computation, and the differences between the individual partial cognate conversion schemes, we prepared two distance matrices. In one matrix, only those 59 concepts for which the B-Cubed F-scores would be 0.8 or less were used, and in one matrix all data were used.

In order to compare the two sets of four distance matrices which were the output of this procedure, we used the traditional Mantel test (Mantel, 1967), which calculates the correlation between distance matrices by means of a permutation method, using 999 permutations per run and the Pearson correlation coefficient as our correlation measure. The correlation scores of the Mantel test fall between -1 and 1 , with -1 indicating high negative correlation, 1 indicating high positive correlation, and 0 indicating no correlation.

Table 10 shows the result of this comparison. While the correlations are extremely high when taking the full data sets (all 201 concepts) into account, we find more fine-grained differences when inspecting only the subsets. The loose and strict conversion schemes show the highest difference, with a (still high) correlation of 0.71. Our salient morpheme conversion (which is based on the hand-curated assignment of salient as opposed to non-salient morphemes in the data) comes second with respect to its difference from the loose coding scheme and a score of 0.76. The highest correlation between distance matrices can be observed for the salient morpheme scheme and the strict conversion scheme, with a score of 0.96.

Although the correlations between the different coding schemes are all high, even for our worst-case subset, the matrix comparison offers us some clearer insights into the specifics of the different conversion schemes. With the strict and the loose conversion schemes representing two extremes, our two new approaches, automated conversion by common morphemes and hand-curated conversion by salient morphemes, fall between the two extremes, with the salient morpheme conversion—in the way in which it was practiced by us—coming closer to the strict conversion than the common morpheme conversion does.

In order to explore the differences between strictly and loosely converted partial cognates, we visualized the results with the help of heat maps, shown in Fig. 1, where we compare pairwise similarities between the dialects (measured by counting shared cognates) for the strictly and loosely converted par-

TABLE 10 Mantel tests of distance matrices derived from a subset of highly divergent concepts (“Subset”) and from considering the full set of data (“Full data set”)

	Subset	Full data set
Loose vs. strict	0.71	0.95
Loose vs. common morpheme	0.85	0.99
Loose vs. salient morpheme	0.76	0.97
Strict vs. common morpheme	0.87	0.96
Strict vs. salient morpheme	0.96	0.98
Common morpheme vs. salient morpheme	0.94	0.99

Mantel tests were calculated from 999 permutations, using the Pearson correlation coefficient as the correlation measure. Significance scores are not provided here, since all permutation tests showed a p -value of less than 0.001, but they are available in the supplementary materials.

tial cognates, using the classification of the seven standard dialect groups by Sagart (2011), later adjusted for subgroups and additional dialect groups by List (2015), as our reference tree. As can be seen from Fig. 1, we have to deal with a lot of reticulation (borrowings or parallel changes due to language contact) in this data set, as reflected in the fact that certain dialects, such as Guǐlín (assigned to the Píngguà group in the source of Liú et al., 2007) or Wēnzhōu (a traditional Wú dialect), show high similarities with the northern dialects (Mandarin and Jin) in the sample. We also observe considerably low similarity scores between dialects which are traditionally assigned to the same dialect groups, such as Lóudī and Chángshā (Xiāng group). Determining the detailed reasons for these skewed similarities requires a thorough comparison of the individual cognate sets, which would go beyond the scope of this paper. However, that the history of the Chinese dialects is intertwined and contains many reticulate events has been observed in many previous studies (List et al., 2014; Norman, 2003) and should not surprise us too much in this context.

The differences between the two matrices in Fig. 1 are striking, but difficult to assess from a direct comparison. All in all, and also due to the specific conversion scheme, the loose conversion yields much higher similarity scores than the strict conversion. In Fig. 2, we have tried to visualize these by plotting the differences in the observed distances for strict and loose cognate conversion. We can see that specifically the southern dialects (Mǐn and Yuè), show the largest differences compared to the other dialects in both conversion schemes. The

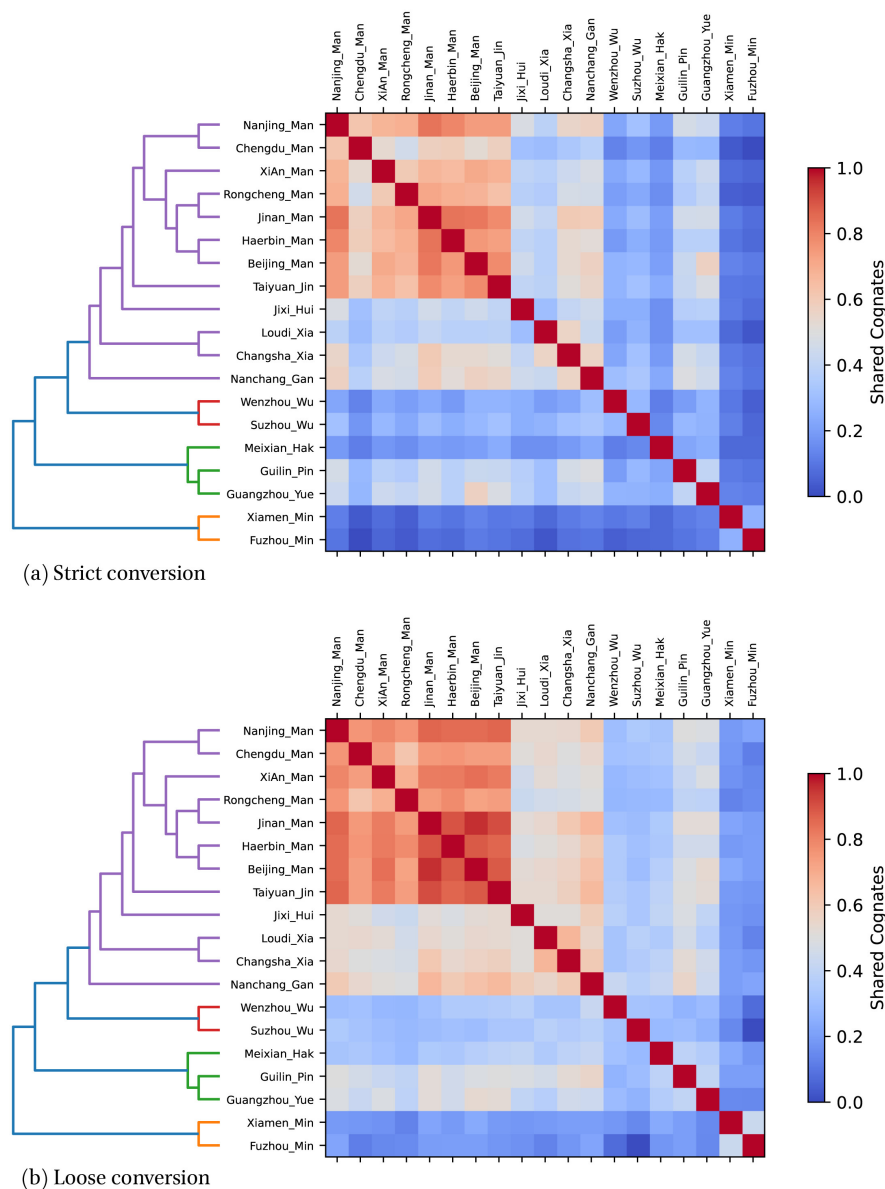


FIGURE 1 Comparing the pairwise similarities in strictly (*top*) and loosely (*bottom*) converted partial cognate sets for the dialects in our sample

Note: The reference phylogeny is based on the classification by Sagart (2011) for the seven major dialect groups, further extended to include all 10 dialect groups and subgrouping inside the groups by List (2015). The same reference phylogeny is used for both matrices. The colors range from red (languages share many cognates) to blue (languages share few cognates).

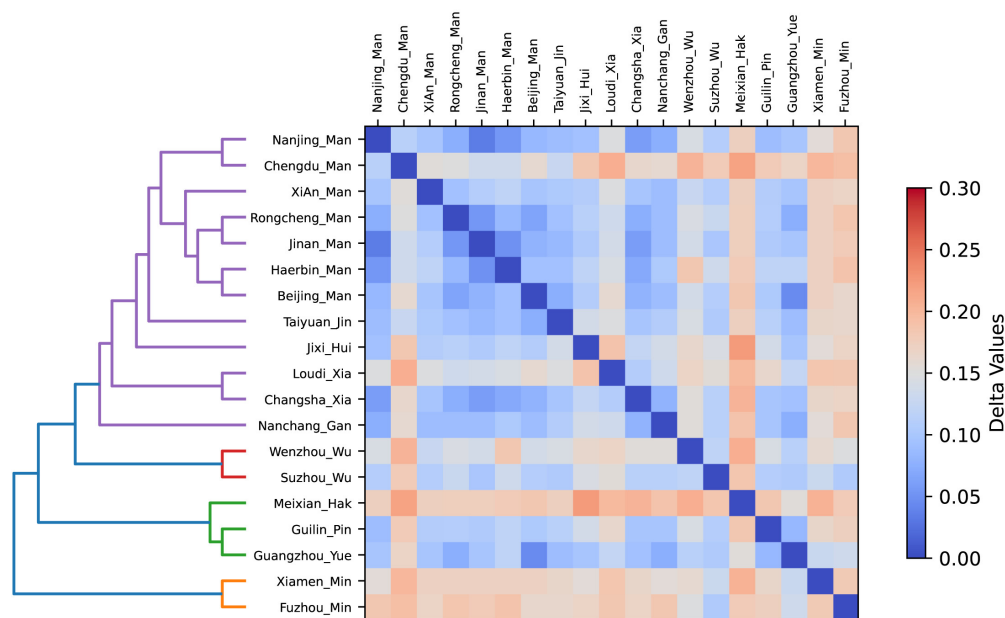


FIGURE 2 Differences in shared cognate sets between loosely and strictly converted cognate sets

reason for these huge differences, which can reach 20 % in some extreme cases, can be found in the difference between the word structures in northern and southern Chinese dialects. While northern dialects tend to have more multi-syllabic words with a complex motivation structure, we find considerably more monosyllabic items in the southern dialects. Since the dialects still employ the same inherited word material, but differ with respect to the compositionality of their words, the strict conversion scheme will increase their divergence, while the loose conversion scheme will increase their similarity.

3.3.3 Partial cognates and language phylogenies

Having analyzed the differences between the distance matrix retrieved from cognate sets derived from partial cognates using different conversion methods, we find that there is a high correlation between all distance matrices when looking at the data set as a whole, while these correlations drop when taking into account only those concepts which we automatically identified as diverse. What remains to be investigated is whether these differences in the distance matrices have a direct impact on the computation of phylogenetic trees. In order to explore this, we took the cognate sets from the 59 highly diverse concepts and generated four Bayesian phylogenies, one for each of the four conversion schemes, following the standard practice of converting cognate sets to

binary presence-absence matrices in which language evolution is modeled as a process of cognate gain and cognate loss (Greenhill et al., 2021).

Bayesian phylogenies have become a standard way of inferring phylogenies from lexical data coded for cognate sets. For our analysis, we used the MrBayes software (Ronquist and Huelsenbeck, 2003) and analyzed the data for the four conversion schemes with the help of a fossilized birth-death model (Stadler, 2010), commonly used in Bayesian phylogenetic studies applied to linguistic data (Chang et al., 2015; Sagart et al., 2019). In order to make sure we received comparable results for root ages (also with respect to alternative analyses that have been done on different data sets in the past), we placed the root age between 1,500 to 2,500 years BP, following a uniform distribution. We had the software generate 20,000,000 different trees in two independent runs from which we sampled every 10,000th tree. Low differences between the trees generated in the independent samples indicated that all four analyses reached convergence. Discarding 10% of the initially generated trees (so-called burn-in), we then reconstructed consensus trees from the remaining 1,800 trees sampled from each of the two runs.

Figure 3 displays the consensus phylogenies reconstructed from the different tree samples. As can be seen from the figure, the tree topologies reconstructed from our four conversion schemes vary quite substantially. Thus, while we find that Hakka (Méixiàn) and Mǐn (Xiàmén and Fúzhōu) form a clade in the strict and the common morpheme conversion, they appear in separate groups in the remaining conversion schemes. While the strict conversion phylogeny provides a scenario in which the more archaic dialect groups of Mǐn, Wú, and Hakka—with the exception of Yuè (Guǎngzhōu), which causes problems in all approaches, probably due to the heavy recent contact with Mandarin—split off first, while more innovative groups are established later, this scenario is less supported by the remaining approaches. With the exception of the loose conversion scheme, in which Chéngdū, a Mandarin dialect, is surprisingly clustered with Xiāng and Wú dialects, all schemes basically recover the traditionally proposed dialect subgroups. The only exception is the Jīn group, represented by Tàiyuán, which is heavily disputed among traditional scholars of Chinese dialectology and classified as a Mandarin dialect in alternative proposals; it appears inside the Mandarin group in all four scenarios.

The scenarios also differ quite substantially with respect to the degree to which the trees are resolved. While we find a clear binary split at the top of the tree only for the strict conversion scheme, we find star-like top-level branchings to different degrees in all other approaches. Here, the loose conversion shows the lowest degree of resolution, failing to resolve eight branches at the

COGNATES IN PHYLOGENETIC STUDIES OF SOUTHEAST ASIAN LANGUAGES 29

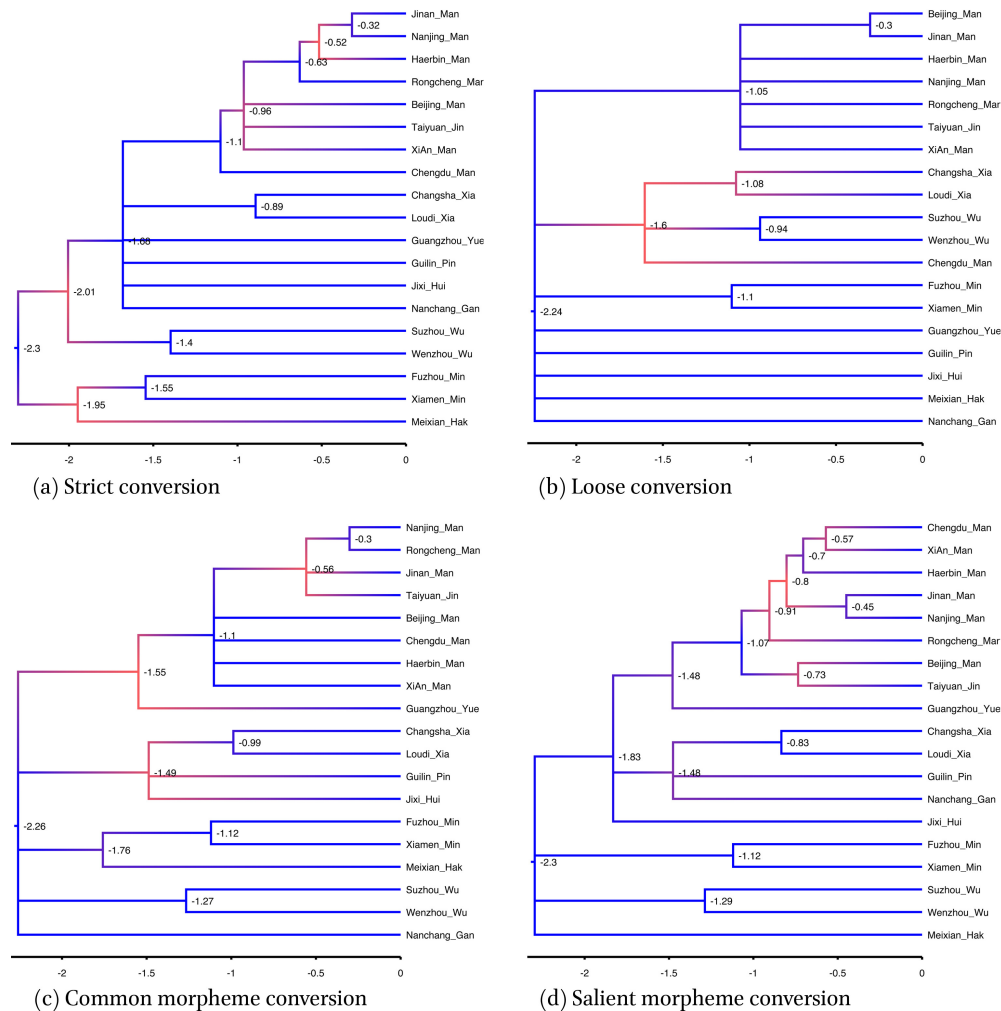


FIGURE 3 Comparing Bayesian phylogenies (consensus trees) based on our four different conversion schemes: strict conversion (a), loose conversion (b), common morpheme conversion (c), and salient morpheme conversion (d)

Note: Nodes are annotated with the age of the branching events; branches are colored according to the probabilities, with blue indicating high probabilities and red indicating low probabilities.

top level, followed by the common morpheme conversion with five branches, and the salient morpheme conversion with four branches.

Given that we fixed the age of the tree, providing divergence dates conforming to traditional assumptions of Chinese dialect diversification, and given that we did not use any internal calibration points, we cannot learn much from the overall tree ages, which are largely the same in all four approaches. However,

internal age estimates show some remarkable differences, specifically for the Wú dialect group, where estimates differ by more than 400 years when comparing the loose conversion estimate of 940 years with the strict conversion estimate of 1,400 years. Similarly, the split of the Mǐn varieties of Fúzhōu and Xiàmén is dated at 1,550 years in the strict conversion, while the three other conversion methods provide estimates of around 1,100 years.

In traditional Chinese historical linguistics, there are different accounts of the overall pattern of Chinese dialect evolution. Norman (2003) assumes that there was a split into three groups, consisting of a southern group comprising Hakka, Mǐn, and Yuè, a northern group consisting of the Mandarin dialects (including Jīn), and an intermediate group consisting of Wú, Xiāng, and Gàn dialects. An alternative scenario, specifically propagated by Karlgren (1954), assumes that the Mǐn dialects split off first, and that the other dialects evolved from a koine that formed around AD 600. Sagart (2011) follows Karlgren (and most Chinese dialectologists) in assuming that the Mǐn dialects split off first, but proposes a more complex diversification scenario, in which the other branches split off step by step, starting from Yuè and Hakka, followed by Wú, Gàn, and Xiāng (see List, 2015, for details on this scenario).

When comparing these scenarios with the phylogenies based on the four conversion schemes, we can see that all four of them diverge from traditional accounts, most likely due to problems in dealing with the impact of undetected borrowings, large-scale convergence in some of the dialect groups, and because the phylogenies were only reconstructed from a small number of concepts susceptible to high variation resulting from lexical compositionality. However, we can also see that the conversion schemes differ regarding the degree to which they diverge from the traditional scenarios. Thus, while the strict conversion scheme conforms in part to the idea of Sagart that Chinese dialect groups split off step by step, the loose conversion scheme proposes a largely star-like diversification of Chinese dialects, in which multiple branches originate from the root at the same time. While the salient morpheme conversion scheme likewise reflects parts of Sagart's nested scenario in proposing a clade comprising Mandarin, Xiāng, and Gàn (and the highly mixed Pínghuà), the common morpheme comparison only uncovers Mandarin (with Jīn) as a distinct clade, with Gàn as a top-level clade.

4 Discussion

Lexical compositionality creates a considerable problem for the identification of cognate sets in lexicostatistical word lists. Since processes of derivation and

compounding are frequent in the languages of the world and often also include the realm of basic vocabulary, which is predominantly used to reconstruct language phylogenies, we think that it cannot be simply neglected but must be actively taken into account and dealt with if we want to improve current approaches to phylogenetic reconstruction. Given that the problem of lexical compositionality resulting from compounding and derivation is particularly prominent in Southeast Asian languages, we conducted an experiment on Chinese dialect evolution by creating a new data set of Chinese dialects in which partial cognates are annotated in great detail. Assuming that different coding techniques by which cognate judgments for entire words are derived from cognate judgments from cognates annotated for individual morphemes might have a direct impact on phylogenetic reconstruction, we conducted an experiment in which we compared four different coding schemes. Three of these four coding schemes can be automatically derived from data annotated for partial cognates, while one additional coding scheme, which we label “salient morpheme conversion,” requires human assessment. In order to provide guidance in conducting these different forms of data annotation, we developed some basic techniques by which scholars can explore their data in order to identify potential difficulties. Applying the methods to a newly compiled data set of 19 Chinese dialect varieties, originally collected by Liú et al. (2007), we find that although the distance matrices derived from the different conversion methods strongly correlate, they yield quite different tree topologies when analyzed with Bayesian methods for phylogenetic reconstruction.

All in all, the differences in the phylogenies allow us to provide a rough ranking of the different approaches to cognate set conversion. We find that the loose conversion scheme performs worst, leading to mostly star-like phylogenies without much resolution, accompanied by clearly wrong groupings of individual varieties, and probably also largely inconsistent age estimates. The reason for these problems lies in the fact that loose conversion artificially increases similarities between varieties by assigning words to the same cognate sets even though they do not share a single cognate morpheme (Hill and List, 2017). While the common morpheme conversion scheme deals to some degree with the problem of low resolution, we find that it yields inconsistent groupings in comparison with traditional accounts. The reason for these problems can be found in the greediness of the approach, which does not further differentiate morphemes with respect to their potential to reflect overall word histories. The strict and salient morpheme conversion schemes perform best in our opinion, with the strict conversion scheme leading to a higher resolution of the phylogeny, but also to larger divergence estimates for individual subgroups. Specifically in data sets of larger time depths in which diverse language vari-

eties are investigated, the strict conversion scheme might artificially increase the distance among the individual language varieties. As a result, it may be recommendable to code for salient morphemes.

All in all, we believe that our study clearly shows that all analyses in which partial cognates recur frequently (and this includes quite a few language families) should be done with great care. Initial cognate annotation should always be done at the morpheme level, ideally including detailed phonetic alignments. Assigning cognate sets to full words should always be based on clear annotation principles. While we know that the conversion of partial cognates to full word cognates is difficult, we think that the techniques for data exploration we provide in this study can help scholars in their concrete annotation practice. Furthermore, by providing a coding technique that tries to closely reflect how scholars conducted implicit cognate judgments in the past, we hope to contribute to the growing work on computer-assisted as opposed to computer-based language comparison.

5 Outlook

In this study we have tried to show that the problem of cognate coding in languages in which we find a rich inventory of word formation processes cannot be easily ignored. We illustrated this with the help of a case study of Chinese dialect varieties which shows that tree topologies can differ drastically, depending on the approaches used to convert partial cognates, annotated on the morpheme level, into full cognates, annotated at the word level.

While we hesitate to recommend one particular conversion scheme as the only one to be used in the future, we are convinced that our study shows that certain conversion practices should be undertaken with great care. Particular practices, like conversion based on a loose assignment of cognacy (loose cognate conversion) or the greedy assignment of words to the same cognate set even though they may share only one common morpheme (common morpheme conversion), need to be considered carefully before they are used. We hope that our case study helps to increase awareness among colleagues working in the field of phylogenetic reconstruction that the way in which one derives cognate judgments from comparative data has an immediate impact on the results.

Supplementary material

The data set compiled by Liú et al. (2007) has been converted to Cross-Linguistic Data Formats and is curated on GitHub (<https://github.com/lexibank/liusinitic>, version 1.3) and has been archived with Zenodo (<https://doi.org/10.5281/zenodo.6637640>). The new methods for the conversion of partial cognates into full cognates using the greedy algorithm described in this study, as well as the checks for partial cognates which recur across different concepts and the difference between strict and loose cognates measured by calculating B-Cubed F-scores, have been included in the LingRex library (<https://pypi.org/project/lingrex>, version 1.3; List and Forkel, 2022). Detailed instructions on how to run the experiments reported here (including detailed analyses for the Bayesian phylogenies) and a Makefile that allows for the quick replication of all studies are available on GitHub (<https://github.com/lingpy/evaluation-paper>, version 1.0) and have been archived on Zenodo (<https://doi.org/10.5281/zenodo.6726637>).

Acknowledgments

MSW and JML were funded by the ERC Starting Grant 715618 “Computer Assisted Language Comparison” (cf. <https://digling.org/calc/>). The study was furthermore funded by the Max Planck Society, through the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology. We thank Russell D. Gray for generously funding APCs to make this article available in Open Access.

References

- Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12(4): 461–486. <https://doi.org/10.1007/s10791-008-9066-8>.
- Anderson, Cormac, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting* 4(1): 21–53. <https://doi.org/10.2478/yplm-2018-0002>.
- Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin: de Gruyter.
- Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1): 194–244. <https://doi.org/10.1353/lan.2015.0005>.

- Chén, Qíguāng. 2012. 苗瑶语文 *Miáoyáo yǔwén* [Miao and Yao language]. Beijing: 中央民族大学 Zhōngyāng Mínzú Dàxué [Central Institute of Minorities].
- Donohue, Mark, Tim Denham, and Stephen Oppenheimer. 2012. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* 29(4): 505–522. <https://doi.org/10.1075/dia.29.4.04don>.
- Felsenstein, Joseph. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* 19(1): 445–471. <https://doi.org/10.1146/annurev.es.19.110188.002305>.
- Gerardi, Fabrício Ferraz, Stanislav Reichert, Carolina Aragon, Johann-Mattis List, Robert Forkel, and Tim Wientzek. 2021. TuLeD: Tupian Lexical Database. Version 0.11. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://doi.org/10.5281/zenodo.4629306>. URL: <https://tular.clld.org/contributions/tuled>.
- Forkel, Robert and Johann-Mattis List. 2020. CLDFBench: Give your cross-linguistic data a lift. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 6997–7004. Luxembourg: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf> (accessed October 14, 2022).
- Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(180205): 1–10. <https://doi.org/10.1038/sdata.2018.205>.
- Geisler, Hans and Johann-Mattis List. 2010. Beautiful trees on unstable ground: Notes on the data problem in lexicostatistics. Unpublished manuscript. (To appear in H. Hettrich (ed.), *Die Ausbreitung des Indogermanischen: Thesen aus Sprachwissenschaft, Archäologie Und Genetik*. Wiesbaden: Reichert.)
- Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913): 479–483. <https://doi.org/10.1126/science.1166858>.
- Greenhill, Simon J., Paul Heggarty, and Russell D. Gray. 2021. Bayesian phylolinguistics. In Richard D. Janda, Brian D. Joseph, and Barbara S. Vance (eds.), *The Handbook of Historical Linguistics, Volume 2*, 226–253. West Sussex: Blackwell.
- Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* 112(43): 13296–13301. <https://doi.org/10.1073/pnas.1503793112>.
- Hamed, Mahé Ben and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23(1): 29–60. <https://doi.org/10.1075/dia.23.1.04ham>.
- Hammarström, Harald, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2021.

- Glottolog. Version 4.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://glottolog.org>.
- Hauer, Bradley and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In Haifeng Wang and David Yarowsky (eds.), *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 865–873. Chiang Mai: Asian Federation of Natural Language Processing. <https://aclanthology.org/I11-1000> (accessed October 14, 2022).
- Hill, Nathan W. and Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting* 3(1): 47–76. <https://doi.org/10.1515/yplm-2017-0003>.
- Holm, Hans J. 2007. The new arboretum of Indo-European “trees”: Can new algorithms reveal the phylogeny and even prehistory of Indo-European? *Journal of Quantitative Linguistics* 14(2–3): 167–214. <https://doi.org/10.1080/09296170701378916>.
- Karlgren, Bernhard. 1954. Compendium of phonetics in ancient and archaic Chinese. *Bulletin of the Museum of Far Eastern Antiquities* 26: 211–367.
- Koch, Peter. 2001. Lexical typology from a cognitive and linguistic point of view. In Gerold Ungeheuer et al. (eds.), 2. *Halbband, Linguistic Typology and Language Universals*, 1142–1178. Berlin: de Gruyter. <https://doi.org/10.1515/9783110194265-022>.
- Kolipakam, Vishnupriya, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(171504): 1–17. <https://doi.org/10.1098/rsos.171504>.
- Lee, Sean and Toshikazu Hasegawa. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences* 278(1725): 3662–3669. <https://doi.org/10.1098/rspb.2011.0518>.
- List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.
- List, Johann-Mattis. 2015. Network perspectives on Chinese dialect history. *Bulletin of Chinese Linguistics* 8: 42–67.
- List, Johann-Mattis. 2016. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution* 1(2): 119–136. <https://doi.org/10.1093/jole/lzw006>.
- List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In André Martins and Anselmo Peñas (eds.), *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 9–12. Valencia: Association for Computational Linguistics. <https://aclanthology.org/E17-3003.pdf> (accessed October 14, 2022).
- List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns

- across multiple languages. *Computational Linguistics* 1(45): 137–161. https://doi.org/10.1162/coli_a_00344.
- List, Johann-Mattis. 2021. EDICTOR: A web-based tool for creating, maintaining, and publishing etymological data. Version 2.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://digling.org/edictor/> (accessed October 14, 2022).
- List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. CLTS. Cross-Linguistic Transcription Systems. Version 2.1.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://10.5281/zenodo.4705149>.
- List, Johann-Mattis and Robert Forkel. 2021. LingPy: A python library for quantitative tasks in historical linguistics. Version 2.6.9. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://pypi.org/project/lingpy/> (accessed October 14, 2022).
- List, Johann-Mattis and Robert Forkel. 2022. LingRex: Linguistic reconstruction with LingPy. Version 1.3.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://pypi.org/project/lingrex/> (accessed October 14, 2022).
- List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymiski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9(316): 1–16. <https://doi.org/10.1038/s41597-022-01432-0>.
- List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1): 1–18. <https://doi.org/10.1371/journal.pone.0170046>.
- List, Johann-Mattis, Philippe Lopez, and Eric Baptiste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, 599–605. Berlin. <https://doi.org/10.18653/v1/P16-2097>.
- List, Johann-Mattis, Shijulal Nelson-Sathi, William Martin, and Hans Geisler. 2014. Using phylogenetic networks to model Chinese dialect history. *Language Dynamics and Change* 4(2): 222–252. <https://doi.org/10.1163/22105832-00402008>.
- List, Johann-Mattis, Annika Tjuka, Christoph Rzymiski, Simon J. Greenhill, Nathanael Schweikhard, and Robert Forkel. 2022. Concepticon: A resource for the linking of concept lists. Version 2.6.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://doi.org/10.5281/zenodo.4911605>.
- Liú, Lǐlǐ, Hóngzhōng Wáng, and Yíng Bái. 2007. 现代汉语方言核心词·特征词集 Xiàndài hànyǔ fāngyán héxīncí, tèzhēng cíjí [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nanjing: 凤凰 Fènghuáng.
- Mann, Noel Walter. 1998. *A Phonological Reconstruction of Proto Northern Burmic*. PhD dissertation, University of Texas, Arlington.
- Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27(2): 209–220.
- Máo, Zōngwǔ. 2004. 瑶族勉语方言研究 Yáo zú miǎnyǔ fāngyán yánjiù [Research on the Mien dialect of the Yao people]. Beijing: 民族出版社 Mínzú Chūbǎnshè.

- Matisoff, James A., ed. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University Presses of California, Columbia; Princeton.
- Mei, Tsu-lin. 1995. 方言本字研究的两种方法 *Fāngyán běnzì yánjiū de liǎngzhǒng fāngfǎ*. 吴语和闽语的比较研究 *Wúyǔ hé mǐnyǔ de bǐjiào yánjiū* 1.
- Moran, Steven and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Language Science Press.
- Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.
- Norman, Jerry. 2003. The Sino-Tibetan languages. In Graham Thurgood and Randy J. LaPolla (eds.), *The Sino-Tibetan Languages*, 72–83. London: Routledge.
- Ratliff, Martha. 2010. *Hmong-Mien Language History*. Canberra: Pacific Linguistics.
- Ronquist, Frederik and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12): 1572–1574. <https://doi.org/10.1093/bioinformatics/btg180>.
- Sagart, Laurent. 2011. Classifying Chinese dialects/Sinitic languages on shared innovations. Séminaire Sino-Tibétain du CRLAO, 28 March. https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations (accessed October 14, 2022).
- Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America* 116: 10317–10322. <https://doi.org/10.1073/pnas.1817972116>.
- Satterthwaite-Phillips, Damian. 2011. *Phylogenetic Inference of the Tibeto-Burman Languages or on the Usefulness of Lexicostatistics (and Megalo-Comparison) for the Subgrouping of Tibeto-Burman*. PhD dissertation, Stanford University.
- Schweikhard, Nathanael E. and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics* 17(1): 2–26. http://www.skase.sk/Volumes/JTL43/pdf_doc/01.pdf (accessed October 14, 2022).
- Stadler, Tanja. 2010. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* 267(3): 396–404. <https://doi.org/10.1016/j.jtbi.2010.09.010>.
- Starostin, George S. 2013. *Metodologija: Kojzanskie jazyki, vol. 1*. Moscow: Jazyki Russkoj Kult'ury.
- Vittrant, Alice and Justin Watkins. 2019. *The Mainland Southeast Asia Linguistic Area*. Berlin: De Gruyter Mouton. <https://doi.org/10.1515/9783110401981>.
- Wu, Mei-Shin, Nathanael E. Schweikhard, Tim A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison. State of the art. *Journal of Open Humanities Data* 6(2): 1–14. <https://doi.org/10.5334/johd.12>.
- Yan, Margaret Mian. 2006. *Introduction to Chinese Dialectology*. Munich: LINCOM Europa.

3.5 Retrospective

In the past, scholars have made efforts to tune the model parameters to the “finest”, or to develop sophisticated models that include evolutionary factors. However, our experiment revealed that different cognate coding logic would result in different tree topologies despite the model parameters remaining the same. This result should compel scholars to pay more attention to the most basic level of the Bayesian phylogenetic analysis, namely the input data.

3.5.1 Complex Algorithms Require Careful Data Treatment

In this retrospective, we would like to mention that the input data for complicated algorithms require delicate data handling. Since this statement would take up too much space in the paper, we could not include it in Wu and List (2022). We experimented with a variety of phylogenetic reconstruction methods while working on the project. Simple NJ methods (Figure 3.2), NJ with bootstraps and maximum likelihood, and a Bayesian phylogenetic algorithm were all tested at various levels of complexity.

The images were all derived from the same set of cognate sets. However, we could see that the topologies made tremendous differences when using algorithms such as the maximum likelihood or Bayesian phylogeny.

3.5.2 Salient Morpheme Annotation Requires Expert Knowledge

Identifying the salient part of a compound word requires specialist knowledge and cannot be automated at this time, as stated in our manuscript. At first, we considered the association between morpheme saliency and compound type. However, there is no standard for the compound type. Another factor that influences a compound word’s salient morpheme’s position is diachronic linguistic changes. We took Sinitic languages as an example. Baxter and Sagart (2014) pointed out that there were both monosyllabic and disyllabic words in their reconstruction of Old Chinese. The disyllabic words were then reduced to monosyllabic words in Middle Chinese. For example, **kə.rʰak* 落 *luò* “fall” in Old Chinese became the Middle Chinese *lak*, and **mə.rʰək* 來 *lái* “come” in Old Chinese became *loj* in Middle Chinese (ibid., p. 53). At the same time, another tendency that became significant was that lexical items were expressed via multiple monosyllabic morphemes; that is, disyllabic words during the Middle Chinese period were created via word compounding and affixation. Modern Sinitic languages have an even stronger preference for multisyllabic words. For example, the concept of “fall” is often expressed as 落下 *luòxià* “fall down” in Standard Mandarin despite the fact that the morpheme 落 *luò* already has the meaning of “fall down”. The diachronic development of the fondness for disyllabic words resulted in the situation in which the salient parts, the morphemes that contribute most to the compound words’ semantics, are not always at the head of the words. In view of the unpredictable locations of salient morphemes, no computer program can be used to detect the salient part of a word, which is why we claim that expert knowledge is required to determine a morpheme’s saliency in this manuscript.

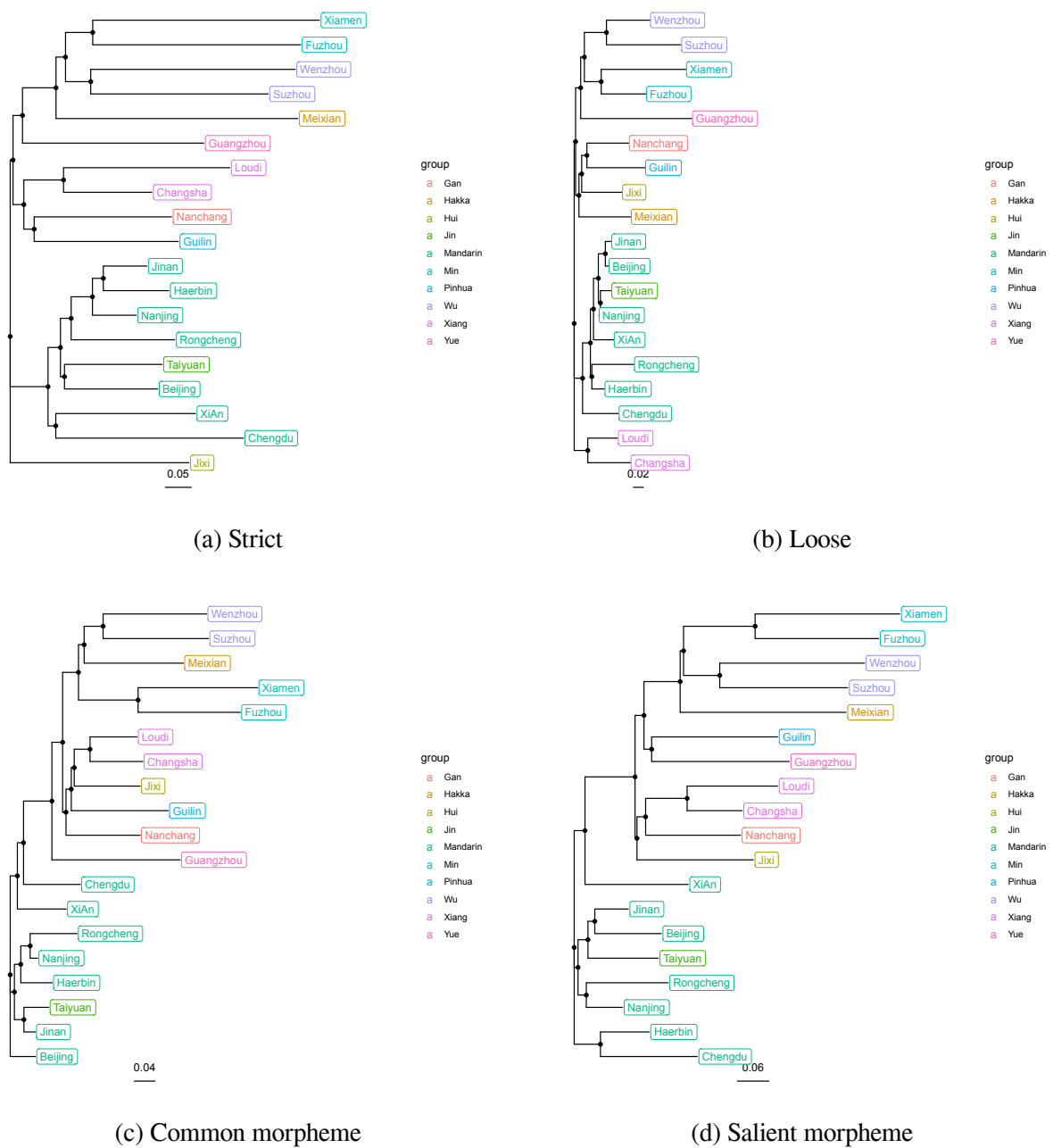


Figure 3.2: Neighbor-joining trees

3.5.3 Morpheme Annotation Is a Solution for Partial Loanwords

During the project, we realized that the salient morpheme conversion method had the benefit of filtering out partial loanwords. Partial loanwords means that a word is a combination of native morphemes and loan morphemes. As stated several times in this dissertation, language contact in the MSEA area has occurred not only within a language family but also across various language families due to long-term migration patterns, the mixed habitat of speaker populations, and frequent trading activities. Due to the long-term migration and intensive language contact, compound words that are composed using native and borrowed morphemes appear in many MSEA languages. Chen (2012, p. 354) pointed out that the Younuo language had compound words that contained a native morpheme and a Sinitic morpheme; for example, 樹林 *fo*⁴⁴*lin*²¹ “forest” is the combination of a native morpheme *fo*⁴⁴ “tree” and a loan morpheme *lin*²¹ “woods, forest”. Partial loanwords can also be found in Sino-Tibetan languages. For example, 奶茶 *ludʒá* “milk tea” in the Wobzi Khroskyabs language variant. The first morpheme *lú* is a native morpheme “milk” and the second morpheme *dʒá* was borrowed from the Tibetan language’s “tea”. Our morpheme annotation is able to extract the native and borrowed components from the compound words without having to discard the complete lexical entry and result in having too few data points to analyze in cases such as these.

3.6 Future Work

A short-term goal is to improve the explicitness of lexical data by using such an annotation scheme. Researchers who are not experts in a given language can quickly comprehend the data set with the assistance of morpheme annotation. Inter-disciplinary studies are facilitated by having lexical data that are annotated well. In addition, this encourages the expression of linguists’ reasoning underlying their cognate decisions.

A long-term goal might be to automatically annotate the morphemes’ meanings and to identify the salient morphemes. At present, we hope that linguists will accept our method, and will allow us to collect data sets with well-annotated morphemes and highlighted salient morphemes for use as training data. However, the proposed annotation scheme is in its initial stages at this point. It is thus necessary to work with languages from other language families in order to identify new questions and to implement new features.

Chapter 4 Bayesian Phylogenetic Analysis

Sagart et al. (2019) proposed a Sino-Tibetan language phylogeny using a Bayesian phylogenetic approach. Their research included a survey of 50 Sino-Tibetan languages that were spoken across a wide geographical area; their findings supported the northern China origin hypothesis, and proposed a time depth for the proto-Sino-Tibetan language of around 7200 B.P. Our study builds on their lexical data and cognate annotations, as well as their calibrations, and extends the language collection to 84 language variants.

Two main factors motivated us to conduct an extended study. First, some lesser-known Sino-Tibetan languages were not included in the aforementioned authors' language samples, particularly the Kho-Bwa and Hrusish languages. These languages are mainly spoken in Arunachal Pradesh, an area bordering China and other nations. We were curious about the relationships among our selected languages, as well as between our selected languages and the languages in the neighboring area. Second, we attempted to avoid the sampling bias by using a two-stage Bayesian phylogeny analysis. The Sino-Tibetan language family (Klaproth, 1823; Matisoff, 2015; Post and Burling, 2017; Sagart, 2011a; Shafer, 1955; Thurgood and LaPolla, 2003) is a large group that is composed of more than 400 languages¹, and many of the languages only have sporadic lexical data available. Therefore, there is unlikely to be a language phylogeny that includes all the Sino-Tibetan languages. Although Sagart et al. (2019) sampled languages in several subgroups, the language subgroups were sampled unevenly. For example, the Sinitic subgroup was represented by six language variants, but some subgroups were only represented by one variety. Our study was also unable to extend the language collection to all 400 languages. Therefore, we utilized a two-stage Bayesian phylogeny approach to adjust the sampling bias. The sampling issue may have affected the topology of the Bayesian phylogeny; however, this has rarely been discussed among scholars. Therefore, our project also attempted to extend the language subgroups in Sagart et al. (ibid.) from a single language to being represented by two or more languages. In addition to the reasons stated above, we are aware that Bayesian phylolinguistic research is a more recent approach than is the lexicostatistic method, or any other qualitative approach. Therefore, we have included a brief section in the article to introduce the method and to discuss its viability.

4.1 The Sino-Tibetan Language Family

The Sino-Tibetan language family is one of the largest language families, both regionally and worldwide, in terms of the number of native speakers. In the nineteenth century, Leyden (1808) classified the languages spoken by “the inhabitants of the regions which lie between India and China and the greater part of the islanders in the eastern sea” as the *Indo-Chinese* language family (ibid.). The definition indicates that his hypothesis was based on ethnicity and geography

¹Glottolog v 4.4 indicates 497 varieties, including languages and dialects.

rather than on linguistics. Leyden's Indo-Chinese language family also included the Austroasiatic, Tai-Kadai, and Hmong-Mien language families. At present, most linguists agree that the similarities among these language families are the result of long-term language contact.

Since the first definition of this language family by Leyden (1808), various labels for this language family have been proposed. In addition to *Sino-Tibetan*, *Tibeto-Burman* (Klaproth, 1823) and *Trans-Himalayan* (van Driem, 2014; van Driem, 2015) have also been suggested.

4.1.1 The Sino-Tibetan paradigm

In 1931, the label *Indo-Chinese* was replaced by the label *Sino-Tibetan* by Przyluski and Luce (1931). Przyluski and Luce compared the word “hundred” in Tai varieties and in Chinese varieties to the same word in Tibeto-Burman varieties, and reconstructed the proto-Sino-Tibetan word **pargya*. The authors also noted that the Tai varieties were not descended from the same ancestor as the Chinese and Tibeto-Burman varieties.

The languages included in the Sino-Tibetan language family have been constantly revised over decades of research. At present, the internal structure of the Sino-Tibetan language family is still highly disputed, except for the relatively isolated position of the Sinitic subgroup in terms of phylogeny. Shafer categorized 300 Sino-Tibetan languages into six groups, namely Sinitic, Daic, Bodic, Burmic, Baric, and Karenic (see Figure 4.1) (Shafer, 1955). His comparison showed an asymmetrical relationship between the Daic and Sinitic varieties. Daic is closer to Sinitic than it is to the other Sino-Tibetan languages. However, Sinitic is actually closer to Bodic than it is to Daic. The author also expressed doubt about Daic's relationship to the Sino-Tibetan languages, and stated that the relationship must be extremely distant.

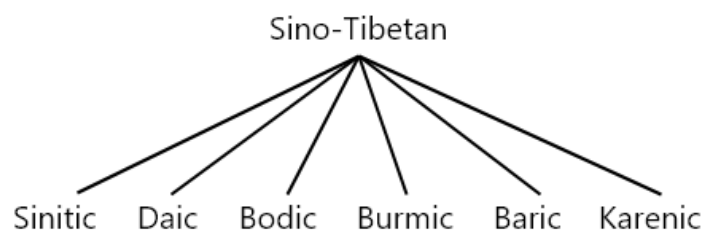


Figure 4.1: The Sino-Tibetan phylogeny proposed by Shafer (1955)

Matisoff (2003) (also see Matisoff (2015)) proposed that the entire language family should be divided into Sinitic and Tibeto-Burman branches, and placed seven subgroups under the Tibeto-Burman branch. He saw the entire Sino-Tibetan language family as being divided into a Sinosphere (Sinitic) and an Indosphere (Tibeto-Burman) based on the political influence of China and India on the Sinitic and the Tibeto-Burman languages, respectively, throughout history. The Sinosphere part tends to develop tones, while the Indosphere part is inclined to develop relative pronouns and correlative structures, as well as retroflex initial consonants (Matisoff, 2003).

Thurgood (2003) provided a different model from Matisoff (2003) (see Figure 4.2b). He stated that the structure of Sino-Tibetan phylogeny was not binary (see the figure below), and that the undefined subgroups indicated that the languages were not related to either the Sinitic branch or to the Tibeto-Burman branch.

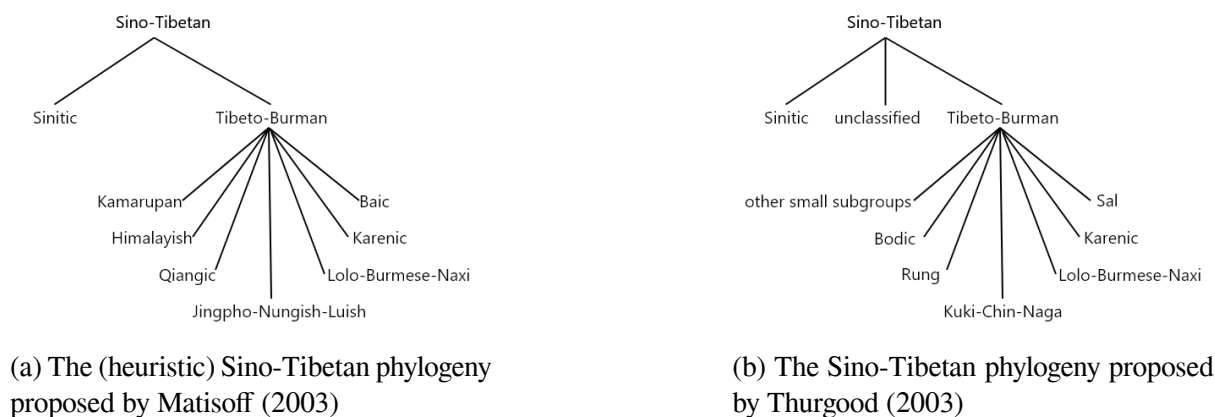


Figure 4.2: Two models of the Sino-Tibetan phylogeny

There is also disagreement regarding the actual internal structure of the Sinitic branch. Some linguists agree with having six subgroups, namely Mandarin, Xiang, Wu, Gan, Yue, and Min, while others have suggested four additional subgroups: Hakka, Jin, Hui, and Pinghua. Despite linguists not yet having agreed on the final classifications, Mandarin has the largest number of language varieties. Furthermore, Min varieties are thought to be an outgroup that is marginally attached to the Sinitic subgroup.

4.1.2 The Tibeto-Burman Paradigm

Tibeto-Burman, as a label to describe the entire language family, was proposed by Klaproth (1823); the author defined the language family as being composed of Chinese, Tibetan, Burmese, and languages related to them (van Driem, 2011). This was a competing hypothesis to the Indo-Chinese view, but it was not well received prior to the past two decades. Three developments (see the items listed below) converged to yield insights heralding a resumption of interest in the Tibeto-Burman language family (ibid.):

1. A better understanding of Old Chinese.
2. Improved insights into the genetic position of Sinitic and an appreciation of its Tibeto-Burman character.
3. The exhaustive identification of all the Tibeto-Burman subgroups.

Van Driem later suggested a newer version, the *Trans-Himalayan* paradigm (van Driem, 2007), to represent the language family for two reasons. First, the languages that are spoken

in the Himalayan mountainous regions are largely undocumented and understudied. Second, the degree of language diversity reaches its peak along the Himalayan mountain range, and the diversity is even better exemplified by the profusion of Sino-Tibetan/Trans-Himalayan language groups that are spoken on both sides of the range.

We referred to the language family as the Trans-Himalayan language family in our accepted article. Compared to the Sino-Tibetan or Tibeto-Burman paradigm, “Trans-Himalayan”, which was proposed by van Driem (2007), is mentioned relatively less frequently by scholars. We use “Trans-Himalayan” to acknowledge the tremendous diversity in the languages spoken on the two sides of the Himalayan mountainous region. Our reason for using the label “Trans-Himalayan” is explained in the accepted article. However, we use “Sino-Tibetan” language family throughout the entire dissertation, except in the accepted article, to avoid readers being confused by our terminology.

4.2 Author Contributions

MSW and TBA initiated the study. MSW and TBA compiled the lexical materials. TBA provided the cognate sets. MSW and TT prepared the Bayesian phylogenetic models. MSW, TBA and TT wrote the manuscript. All authors agree with the final version of the manuscript.

The content below is the paper in the form of the authors’ copy. The paper will appear in *Linguistic of Tibeto-Burman Area* in 2022 (Wu et al., 2022).

4.3 Third Paper

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 Wu¹, Timotheus A. Bodt², and Tiago Tresoldi³

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

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

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

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 decades¹, 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 & Matisoff 1980; Bradley 2002; Thurgood & LaPolla 2003; Sagart 2011; Blench & 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

¹ In this paper, we call the language family also known as Sino-Tibetan or Tibeto-Burman “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 level. 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.

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 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, 2007²) 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

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

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 the forest – were determined by the topography and climate of the mountain ranges they lived in, and were 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

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.”

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 & Lieberherr 2015 for Hrusish; Lieberherr 2015, Lieberherr & 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 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, 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), based 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 much 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, 2014b; Lieberherr 2015; Lieberherr & 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, 2014b) 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 to 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 Gémàn⁴) in addition to the Yidu (hereinafter Yìdū) and Darang Mishmi (hereinafter Dáràng) varieties already present in 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

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

Hrusish languages (Blench 2017: 3, 14). However, besides some perfunctory remarks, we will not pay more attention to these languages.

3 Material and Methods

3.1 Lexical data

The lexical data in our study comprise 86 linguistic varieties, i.e., “doculects” (Good & 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 & 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

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.

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 (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 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-smad⁷ 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.

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 [unlan] 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.

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.

3.1.5 *Tani and Mishmic*

The large Tani group was only represented by Bokar Luoba (Bógǎēr 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 Proto-Tani 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 Yidū 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).

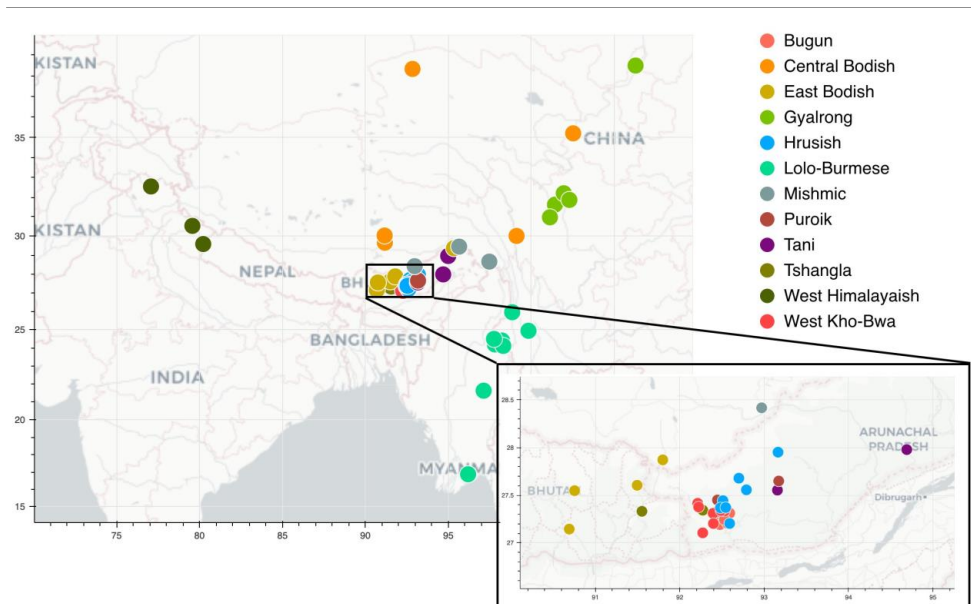


Figure 1: Languages in Arunachal and Tibet in our selection.

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 & 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 detection of 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 & 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 researchers 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

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.

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 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-

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.

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 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 & Moulton 2004) with *SplitsTree 4* (Huson & 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 & 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), an 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, .xlsx, 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.

¹² The project repository is archived on Open Science Framework (<https://osf.io/7u8cw/> and <https://osf.io/9x4s8/>).

¹³ DOI: 10.5281/zenodo.5554780

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 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.

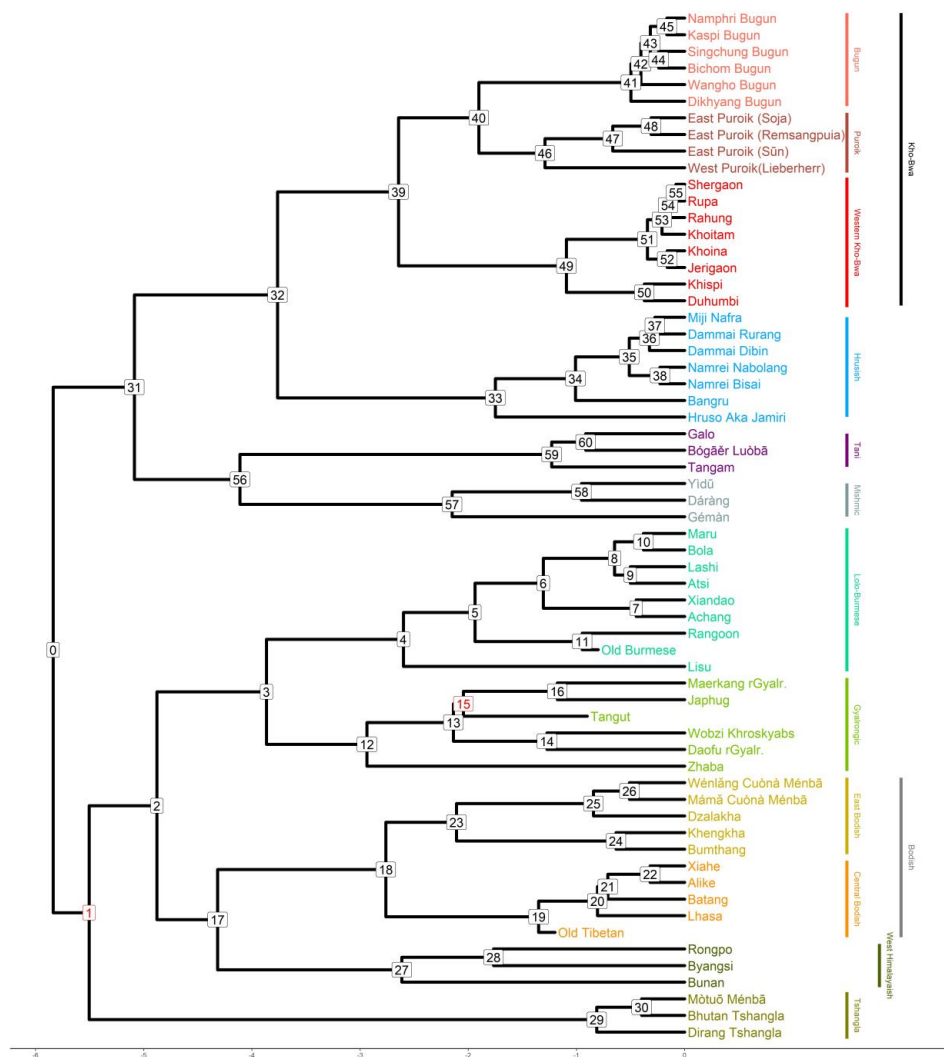


Figure 2: Phylogeny of Trans-Himalayan languages of Tibet-Arunachal

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 were 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 Yidū and Dàràng 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, are 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 & Lieberherr 2015, Lieberherr & 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 Mishmic emerged, the Proto-Hrusish, Proto-Kho-Bwa, Proto-Tani and Proto-Mishmic 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 & 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 & 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 counteract 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, which allows us to apply 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 Wénlǎng Cūonà Ménbā form for ‘eye’ is *mek*⁵⁵, while the cognate is actually the Wénlǎng Cūonà Ménbā form *me*⁵⁵.*loŋ*⁵⁵ ‘eyebrow’. There was a semantic change between ‘eyebrow’ in Wénlǎng 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. We are also 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 Rzymiski 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 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.

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.

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 of 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.

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, Singapore, July 16–18, 2014.
- 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 & Nathan Hill (eds.), *Trans-Himalayan linguistics*, 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* (2nd edition). 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 & 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*, 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* [Pacific Linguistics Series A-86], 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], 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* (1st edition). [Routledge Language Family Series], 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). doi: 10.2307/413505.
- Currie, Thomas E., Andrew Meade, Myrtille 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*, 41–70. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110310832.41.
- van Driem, George. 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.
- van Driem, George. 2007. The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese. *Bulletin of Chinese Linguistics* 1(2).211–270.
- van Driem, George. 2014. Trans-Himalayan. In Thomas Owen-Smith & Nathan Hill (eds.), *Trans-Himalayan linguistics*, 11–40. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110310832.11.
- van Driem, George. 2015. Synoptic grammar of the Bumthang language. *Himalayan Linguistics Archive* 6.1-77.

- van Driem, George. 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). 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*. Arunachal Pradesh: Department of History, Rajiv Gandhi University, Arunachal Pradesh PhD dissertation.
- 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*, 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).80205. doi: 10.1038/sdata.2018.205.

-
- Fürer-Haimendorf, Christoph 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 edition). [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. In Hans Henrich Hock & Elena Bashir (eds.), *The Languages and Linguistics of South Asia: A Comprehensive Guide* [The World of Linguistics, Volume 7.], 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. 331–359.
- 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*. New Jersey: John Wiley Sons, Ltd. doi: 10.1002/9781118732168.
- Grewal, Dalvinder Singh. 1992. *The Aka Miji and their kindred in Arunachal Pradesh: An enquiry into determinants of their identity*. Siliguri: University of North Bengal, Siliguri, India, PhD thesis.
- 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 Qíngxià Dài (eds.). 1992. *Zàngmiǎn yǔzú yǔyán cihui* 《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 & Stephen Morey (eds.), *North East Indian Linguistics*, Vol. 5, 91–110. India: Foundation Books. doi: 10.1017/9789382993285.005.
- Hyslop, Gwendolyn. 2014. A preliminary reconstruction of East Bodish. In Nathan Hill & Thomas Owen-Smith (eds.), *Trans-Himalayan Linguistics*, 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], 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*. 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 & Amos Teo (eds.), *North East Indian Linguistics (NEIL)* 7, 235–286. Canberra: Asia-Pacific Linguistics Open Access.
- Lieberherr, Ismael. 2017. *A grammar of Bulu Puroik*. Bern, Switzerland: Universität Bern PhD thesis.
- Lieberherr, Ismael & Timotheus A. Bodt. 2017. Sub-grouping Kho-Bwa based on shared core vocabulary. *Himalayan Linguistics* 16(2).26-63 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 Rzymiski, 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 Rzymiski, 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 yánjiū* 《门巴语方言研究》 [*A study of Menba*]. Běijīng: Mínzú chūbǎn shè [民族出版社].
- Lǐ, Dàqín. 2004. *Sūlóngyǔ yánjiū* 《苏龙语研究》 [*A study of Sulong*] (1st edition). Běijīng: Mínzú chūbǎn shè [民族出版社].
- Macario, Florens Jean-Jacques. 2015. The genetic position of Apatani within Tibeto-Burman. In Linda Konnerth, Stephen Morey, Prizankoo Sarmah & Amos Teo (eds.), *North East Indian Linguistics (NEIL) 7*, 213–233. Canberra: Australian National University.
- 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*], 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 & 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*, 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*, 73–88. London: Routledge.
- Modi, Yankee & Mark W. Post. 2009. The sociolinguistic context and genetic position of Holon (Milang) in Tibeto-Burman. Paper presented at the 42th International Conference on Sino-Tibetan Languages and Linguistics, Chiang Mai, November 2-4, 2009.
- 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*. Melbourne, Australia: La Trobe University PhD thesis.
- 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. In Graham Thurgood & Randy J. LaPolla (eds), *The Sino-Tibetan Languages*, 213–242. London: Routledge.
- 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*. Itanagar , India: Rajiv Gandhi University PhD thesis.
- Rutgers, Leopold Roland. 1999. Puroik or Sulung of Arunachal Pradesh. Paper presented at the 5th Himalayan Languages Symposium, Kathmandu, 13–15 September, 1999.
- Rzyski, 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).143–147/e21. 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*. [Yale agrarian studies series.] New Haven: 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*. Berkeley, USA: Univerisity of California at Berkeley PhD thesis.
- 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 & Randy J. LaPolla (eds.), *The Sino-Tibetan languages*, 3–21. London: Routledge.

- Thurgood, Graham & 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 Manuel Widmer (ed.), *A descriptive grammar of Bunan*, 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.

Authors' contact information:

Mei-Shin Wu
 1Department of Linguistic and Cultural Evolution
 Max Planck Institute for Evolutionary Anthropology
 Leipzig
 Germany
 Email (for correspondence): mei_shin_wu@eva.mpg.de

Timotheus A. Bodt

4.4 The Supplementary Material of the Third Paper

The following content is the supplementary material accompanying to Wu et al. (2022). The PDF file and the datasets are also archived on the *Open Science Framework* (OSF) website. The online archive link is <https://osf.io/7u8cw/>.

Supplementary of “Bayesian phylogenetics illuminate shallower
relationships among Trans-Himalayan languages in the
Tibet-Arunachal area”

Mei-Shin Wu¹, Timotheus A. Bodt², and Tiago Tresoldi³

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

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

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

July 29, 2022

Contents

S1 Dataset and Codes	1
S2 Additional languages	1
S3 Concept coverage	3
S4 Trans-Himalayan phylogeny	7
S5 Tibet-Arunachal in the DensiTree visualization	9
S6 Neighbor-net network	10
S7 Heatmaps	10

S1 Dataset and Codes

Our raw data in .tsv, .ods, and .xlsx formats can be found on Open Science Framework (<https://osf.io/7u8cw>) and Zenodo (DOI: 10.5281/zenodo.5554780, this repository will be publicly accessible when the manuscript is published). At this moment, the Open Science Framework provides a reviewer only link so that reviewers can download and inspect our data in various formats. Both repositories will be publicly accessible when the manuscript is published.

We archive our data in both the Wordlist format and in the cross-linguistic data format, Python scripts and models of Bayesian phylogenetic analysis in another repository on Open Science Framework (<https://osf.io/9x4s8>).

S2 Additional languages

In this table, we present only the additional datasets we prepared for the study. Please refer to the archive on Zenodo (10.5281/zenodo.2598440) for the languages which are included in Sagart et al. (2019).

Table 1: Additional Languages

ID	Doculect	Glottocode	Longitude	Latitude	Source
Bodic (Tshangla)					
DirangTshangla	Dirang Tshangla	dira1243	92.273057	27.343674	primary source
BhutanTshangla	Bhutan Tshangla	tsha1245	91.551505	27.332081	primary source
Tani					
Galo	Galo	galo1242	94.69	27.98	Post (2007)
Tangam	Tangam	tang1377	94.990737	28.956354	Post (2017)
Kho-Bwa (Puroik)					
KBWestPuroikLieberherr	Western Puroik	bulu1255	92.44627	27.451029	Lieberherr (2017, 2015)
KBEastPuroikSoja	Eastern Puroik	sulu1241	93.172368	27.650542	Soja (2009); Tayeng (1990)
KBEastPuroikRemsangpuia	Eastern Puroik	sulu1241	93.172368	27.650542	Remsangpuia (2008)
KBEastPuroikSun	Eastern Puroik	sulu1241	92.97226	28.41558	Sun et al. (1991); Li (2004)
Kho-Bwa (Bugun)					
DikhyangBugun	Bugun	dikh1234	92.454924	27.320448	Bodt (2017) and primary source
WanghoBugun	Bugun	wang1301	92.420806	27.24169	Abraham et al.(2018[2005])
BichomBugun	Bugun	bich1234	92.592007	27.31177	Abraham et al.(2018[2005])
SingchungBugun	Bugun	sing1271	92.474176	27.191743	Abraham et al.(2018[2005])
KaspiBugun	Bugun	kasp1234	92.564228	27.204536	Abraham et al.(2018[2005])
NamphriBugun	Bugun	namp1239	92.52836	27.24403	Abraham et al.(2018[2005])
Kho-Bwa (Western Kho-Bwa)					
Duhumbi	Duhumbi	chug1252	92.212034	27.416112	Bodt and List (2019); Bodt (2019, 2021, 2022)

Khispi	Khispi	lish1235	92.223221	27.378042	Bodt and List (2019); Bodt (2019, 2021, 2022)
Khoina	Sartang	khoi1253	92.52994	27.334981	Bodt and List (2019); Bodt (2019, 2021, 2022)
Khoitam	Sartang	khoi1252	92.439455	27.327487	Bodt and List (2019); Bodt (2019, 2021, 2022)
Rahung	Sartang	rahu1234	92.395028	27.310778	Bodt and List (2019); Bodt (2019, 2021, 2022)
Rupa	Sherdukpen	rupa1234	92.398757	27.203065	Bodt and List (2019); Bodt (2019, 2021, 2022)
Shergaon	Sherdukpen	sher1261	92.272133	27.105018	Bodt and List (2019); Bodt (2019, 2021, 2022)
Jerigaon	Sartang	jeri1243	92.486595	27.340629	Bodt and List (2019); Bodt (2019, 2021, 2022)
<hr/> East Bodish <hr/>					
Khengkha	Khengkha	khen1241	90.689999	27.144878	Yangzom and Arkesteijn (1996); primary source
Dzalakha	Dzalakha	dzal1238	91.494864	27.605989	Dzongkha Develop- ment Commission (2017); primary source
Bumthang	Bumthang	bumt1240	90.753738	27.549248	van Driem (2015); Dzongkha Develop- ment Commission (2018); primary source
MamaCuonaMenba	Tshona Dakpa (Mámă Cuònă Ménbā)	dakp1242	91.798576	27.872413	Lù (2002, 1986)
WenlangCuonaMenba	Tawang Dakpa (Wénláng Cuònă Ménbā)	dakp1242	95.339063	29.367765	Lù (2002, 1986)
<hr/> Hrusish <hr/>					
HrusoAkaJamiri	Hruso Aka	hrus1242	92.590615	27.203617	Abraham et al.(2018[2005]), Simon(1993[1970])
Bangru	Bangru	bang1369	93.164169	27.952786	Bodt and Lieberherr (2015); primary source
NamreiBisai	Namrei	east2847	92.699998	27.680112	Abraham et al.(2018[2005])
NamreiNabolang	Namrei	east2847	92.789317	27.558051	Abraham et al.(2018[2005])
DammaiDibin	Miji	west2937	92.511881	27.444541	Abraham et al.(2018[2005])
DammaiRurang	Miji	west2937	92.48215	27.368891	Abraham et al.(2018[2005])

MijiNafra	Miji	west2937	92.544458	27.371831	Abraham et al.(2018[2005]), Simon (1979)
Mishmic					
MishmiKaman	Kaman (Gémàn)	miju1243	92.970278	28.416389	Sun (1993)

S3 Concept coverage

Table 2: The mutual coverage rate of concepts. All percentages are calculated based on 86 languages.

concepts	Entries (percentage)	Cognate sets (percentage)	Singleton
above	81 (90.7%)	81 (40.74%)	22
all	82 (84.88%)	82 (54.88%)	32
bad	92 (90.7%)	92 (54.35%)	31
below, under	87 (90.7%)	87 (35.63%)	15
big	90 (100.0%)	90 (38.89%)	25
black	91 (100.0%)	91 (29.67%)	16
cold (of temperature)	100 (98.84%)	100 (41.0%)	26
correct (right)	62 (67.44%)	62 (56.45%)	22
dark	66 (69.77%)	66 (40.91%)	16
dirty	76 (83.72%)	76 (60.53%)	30
dry	91 (98.84%)	91 (35.16%)	17
early	62 (70.93%)	62 (46.77%)	21
eight	81 (93.02%)	81 (8.64%)	3
far	86 (100.0%)	86 (33.72%)	16
firewood	60 (69.77%)	60 (16.67%)	6
five	83 (96.51%)	83 (4.82%)	2
four	85 (98.84%)	85 (1.18%)	0
full	84 (93.02%)	84 (25.0%)	14
good	101 (100.0%)	101 (42.57%)	26
green	87 (97.67%)	87 (32.18%)	16
hard	81 (88.37%)	81 (39.51%)	21
he or she [third person singular]	88 (96.51%)	88 (42.05%)	24
heavy	86 (97.67%)	86 (18.6%)	13
here	67 (74.42%)	67 (53.73%)	26
high / tall	72 (76.74%)	72 (31.94%)	12
horizontal	46 (53.49%)	46 (52.17%)	18
hot	96 (100.0%)	96 (40.62%)	24
hundred	80 (90.7%)	80 (13.75%)	4
I [first person singular]	88 (98.84%)	88 (19.32%)	10
inside	77 (84.88%)	77 (35.06%)	16
knife	69 (76.74%)	69 (36.23%)	14
late	75 (80.23%)	75 (50.67%)	22
left	88 (96.51%)	88 (27.27%)	13
light (of weight)	85 (97.67%)	85 (18.82%)	6
long	90 (100.0%)	90 (32.22%)	22
many	90 (97.67%)	90 (45.56%)	29
middle	73 (80.23%)	73 (35.62%)	18
morning	88 (97.67%)	88 (38.64%)	15
narrow	79 (89.53%)	79 (51.9%)	27
near	89 (98.84%)	89 (39.33%)	25

new	83 (96.51%)	83 (21.69%)	11
nine	80 (93.02%)	80 (7.5%)	2
noon	78 (87.21%)	78 (50.0%)	31
old (of person)	86 (97.67%)	86 (33.72%)	18
one	88 (98.84%)	88 (19.32%)	9
outside	73 (81.4%)	73 (47.95%)	26
red	90 (98.84%)	90 (27.78%)	12
right	87 (98.84%)	87 (29.89%)	16
round	86 (89.53%)	86 (39.53%)	21
salty	55 (61.63%)	55 (32.73%)	11
seven	81 (93.02%)	81 (13.58%)	4
sharp	75 (83.72%)	75 (37.33%)	14
short	82 (94.19%)	82 (34.15%)	16
shy	53 (61.63%)	53 (37.74%)	14
six	80 (93.02%)	80 (6.25%)	2
small	93 (100.0%)	93 (44.09%)	25
smooth	71 (79.07%)	71 (49.3%)	27
soft	82 (89.53%)	82 (47.56%)	28
straight	74 (83.72%)	74 (51.35%)	27
ten	81 (94.19%)	81 (16.05%)	6
that	76 (83.72%)	76 (39.47%)	15
the ant	88 (100.0%)	88 (30.68%)	19
the armpit	54 (61.63%)	54 (61.11%)	25
the bamboo	89 (98.84%)	89 (29.21%)	14
the barley (tibetan or highland)	41 (46.51%)	41 (46.34%)	13
the belly	86 (98.84%)	86 (32.56%)	13
the bird	88 (100.0%)	88 (21.59%)	6
the blood	87 (100.0%)	87 (8.05%)	3
the body hair (hair or fur)	75 (86.05%)	75 (12.0%)	5
the bone	84 (94.19%)	84 (20.24%)	10
the branch	82 (94.19%)	82 (29.27%)	13
the breast (female)	80 (91.86%)	80 (18.75%)	7
the child (young human)	67 (74.42%)	67 (52.24%)	21
the cloud	86 (100.0%)	86 (24.42%)	12
the daughter	89 (98.84%)	89 (34.83%)	17
the dew	69 (76.74%)	69 (39.13%)	18
the dog	88 (100.0%)	88 (18.18%)	7
the dream	85 (97.67%)	85 (4.71%)	1
the dust	81 (89.53%)	81 (29.63%)	15
the ear	85 (97.67%)	85 (12.94%)	6
the earth (soil)	85 (96.51%)	85 (29.41%)	13
the earthworm	66 (76.74%)	66 (60.61%)	27
the egg	89 (100.0%)	89 (24.72%)	11
the eye	85 (98.84%)	85 (7.06%)	2
the father	89 (98.84%)	89 (10.11%)	7
the feather	73 (81.4%)	73 (20.55%)	8
the fire	87 (98.84%)	87 (10.34%)	5
the fish	86 (100.0%)	86 (9.3%)	4
the flea	59 (68.6%)	59 (23.73%)	5
the flower	86 (100.0%)	86 (16.28%)	3
the fog	67 (76.74%)	67 (35.82%)	17
the foot	87 (98.84%)	87 (28.74%)	15
the forest	74 (82.56%)	74 (35.14%)	16
the fox	47 (54.65%)	47 (36.17%)	7

the frog	86 (98.84%)	86 (23.26%)	11
the front (front side)	69 (76.74%)	69 (31.88%)	9
the frost	62 (69.77%)	62 (29.03%)	11
the fruit	89 (100.0%)	89 (20.22%)	8
the goat	86 (96.51%)	86 (26.74%)	12
the grass	78 (87.21%)	78 (42.31%)	20
the hail	68 (77.91%)	68 (38.24%)	15
the hair (of the head)	86 (100.0%)	86 (23.26%)	11
the hand	85 (97.67%)	85 (12.94%)	4
the head	95 (100.0%)	95 (24.21%)	10
the heart	88 (100.0%)	88 (22.73%)	10
the hoof	60 (66.28%)	60 (41.67%)	13
the horn (keratinized skin)	84 (96.51%)	84 (11.9%)	5
the horse	62 (69.77%)	62 (20.97%)	8
the house	80 (89.53%)	80 (18.75%)	4
the husband	78 (87.21%)	78 (46.15%)	26
the ice	64 (70.93%)	64 (23.44%)	7
the knee	84 (93.02%)	84 (29.76%)	12
the lake	64 (72.09%)	64 (34.38%)	13
the leaf	88 (100.0%)	88 (22.73%)	7
the lip (the lips)	66 (75.58%)	66 (34.85%)	18
the liver	79 (89.53%)	79 (15.19%)	7
the louse	78 (89.53%)	78 (10.26%)	5
the lung	76 (87.21%)	76 (27.63%)	10
the man (male human)	88 (97.67%)	88 (43.18%)	27
the meat	88 (100.0%)	88 (15.91%)	8
the moon	86 (98.84%)	86 (8.14%)	5
the mosquito	77 (88.37%)	77 (41.56%)	24
the mother	91 (98.84%)	91 (23.08%)	9
the mountain	89 (100.0%)	89 (31.46%)	15
the mouse or rat	88 (100.0%)	88 (23.86%)	12
the mouth	92 (98.84%)	92 (35.87%)	20
the mud	80 (93.02%)	80 (41.25%)	20
the nail (fingernail or claw)	89 (100.0%)	89 (21.35%)	12
the name	87 (100.0%)	87 (2.3%)	1
the neck	93 (100.0%)	93 (27.96%)	13
the needle (for sewing)	80 (93.02%)	80 (17.5%)	5
the nit	40 (46.51%)	40 (32.5%)	7
the nose	87 (100.0%)	87 (11.49%)	4
the otter	56 (65.12%)	56 (12.5%)	5
the pig	87 (98.84%)	87 (9.2%)	3
the rain	89 (98.84%)	89 (22.47%)	9
the rainbow	77 (88.37%)	77 (44.16%)	20
the rice plant	61 (69.77%)	61 (22.95%)	6
the river	91 (96.51%)	91 (50.55%)	24
the road	87 (98.84%)	87 (21.84%)	9
the root	87 (100.0%)	87 (34.48%)	21
the rope	78 (87.21%)	78 (32.05%)	14
the salt	87 (98.84%)	87 (12.64%)	6
the sand	83 (96.51%)	83 (20.48%)	7
the sea	54 (62.79%)	54 (42.59%)	14
the seed	86 (96.51%)	86 (24.42%)	11
the sheep	69 (75.58%)	69 (28.99%)	10
the shit	74 (84.88%)	74 (17.57%)	5

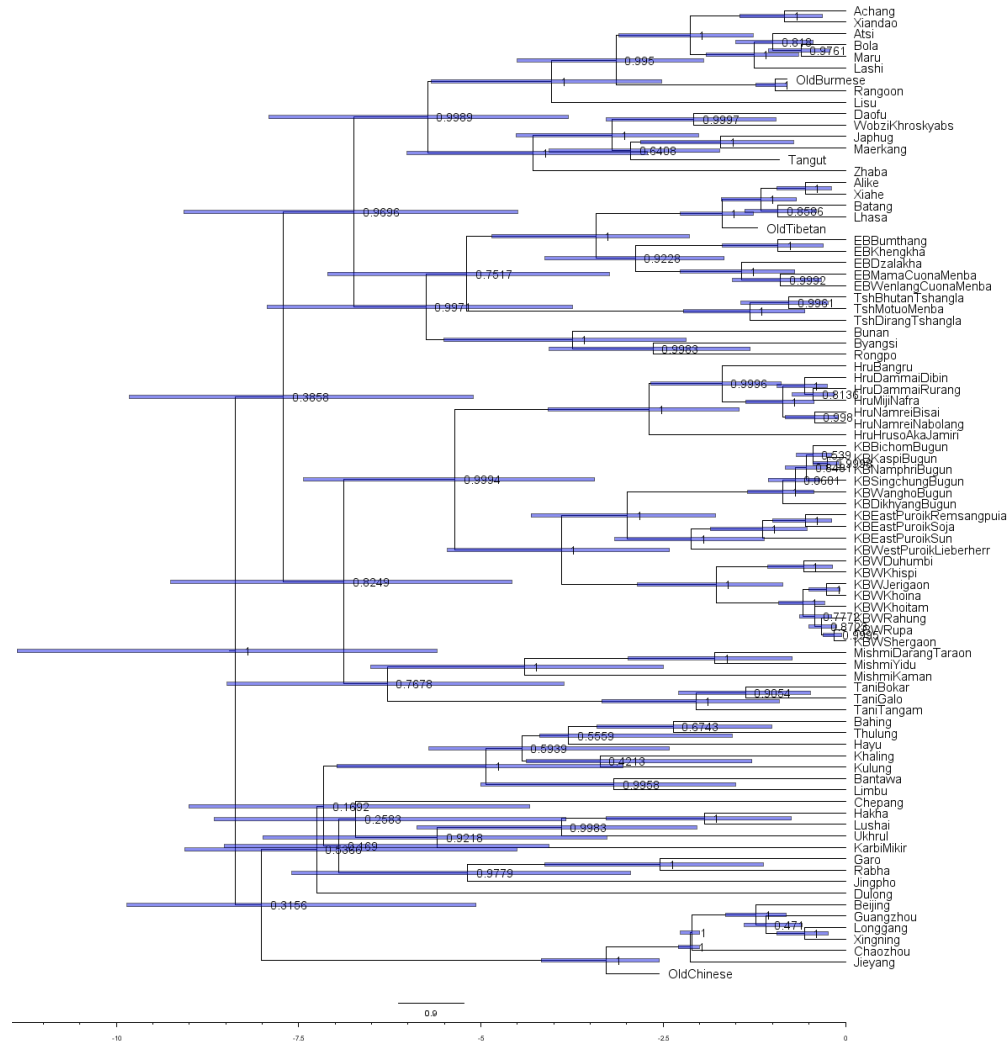
the shoulder	78 (84.88%)	78 (24.36%)	12
the sickle	51 (58.14%)	51 (47.06%)	15
the skin	88 (97.67%)	88 (25.0%)	14
the sky	86 (97.67%)	86 (22.09%)	11
the smoke	86 (100.0%)	86 (10.47%)	4
the snake	88 (100.0%)	88 (12.5%)	6
the snow	74 (83.72%)	74 (36.49%)	18
the son	88 (100.0%)	88 (26.14%)	13
the sparrow	51 (59.3%)	51 (58.82%)	19
the spider	89 (97.67%)	89 (50.56%)	29
the star	87 (100.0%)	87 (26.44%)	12
the stick	75 (83.72%)	75 (58.67%)	29
the stone (a piece of)	86 (100.0%)	86 (19.77%)	7
the sun	88 (95.35%)	88 (20.45%)	6
the tail	90 (100.0%)	90 (15.56%)	5
the thigh	67 (77.91%)	67 (50.75%)	25
the thunder	86 (97.67%)	86 (37.21%)	20
the tiger	87 (100.0%)	87 (29.89%)	15
the tongue	87 (100.0%)	87 (17.24%)	8
the tooth (front)	88 (100.0%)	88 (21.59%)	9
the tree	88 (97.67%)	88 (13.64%)	5
the water	88 (100.0%)	88 (18.18%)	9
the wheat	67 (77.91%)	67 (46.27%)	24
the wife	83 (86.05%)	83 (53.01%)	35
the wind	89 (100.0%)	89 (26.97%)	10
the wing	73 (83.72%)	73 (36.99%)	15
the wolf	39 (44.19%)	39 (48.72%)	13
the woman	92 (98.84%)	92 (44.57%)	28
the wood (material)	71 (81.4%)	71 (19.72%)	10
the year	88 (98.84%)	88 (17.05%)	8
there	68 (73.26%)	68 (70.59%)	37
thick	76 (83.72%)	76 (38.16%)	18
thin (object)	86 (96.51%)	86 (34.88%)	20
this	78 (86.05%)	78 (46.15%)	19
thou [second person singular]	86 (96.51%)	86 (18.6%)	6
three	85 (98.84%)	85 (4.71%)	3
to be alive	72 (83.72%)	72 (34.72%)	14
to bite	88 (98.84%)	88 (39.77%)	18
to blow (of wind)	62 (70.93%)	62 (77.42%)	41
to burn [intransitive]	80 (90.7%)	80 (51.25%)	25
to buy	87 (100.0%)	87 (32.18%)	19
to chew	69 (77.91%)	69 (42.03%)	12
to come	95 (100.0%)	95 (30.53%)	16
to count	77 (86.05%)	77 (48.05%)	27
to cry (weep)	86 (96.51%)	86 (20.93%)	11
to die	91 (100.0%)	91 (9.89%)	7
to dig	76 (83.72%)	76 (38.16%)	19
to drink	89 (98.84%)	89 (24.72%)	13
to eat	88 (98.84%)	88 (14.77%)	5
to fight	73 (81.4%)	73 (63.01%)	34
to float	65 (69.77%)	65 (58.46%)	27
to flow	62 (72.09%)	62 (51.61%)	21
to fly (move through air)	88 (100.0%)	88 (36.36%)	24
to forget	87 (97.67%)	87 (35.63%)	16

to give	89 (96.51%)	89 (24.72%)	14
to hear	86 (98.84%)	86 (36.05%)	17
to hide (conceal)	78 (86.05%)	78 (67.95%)	42
to hold	81 (84.88%)	81 (55.56%)	32
to hunt	76 (83.72%)	76 (61.84%)	34
to kill	92 (98.84%)	92 (17.39%)	12
to knead	43 (48.84%)	43 (60.47%)	23
to know (something)	84 (90.7%)	84 (39.29%)	24
to laugh	88 (100.0%)	88 (25.0%)	10
to learn	55 (60.47%)	55 (47.27%)	21
to lick	82 (87.21%)	82 (20.73%)	10
to lie down	72 (79.07%)	72 (56.94%)	30
to marry (a man marries a woman)	46 (51.16%)	46 (76.09%)	29
to plant (vegetals, rice)	66 (74.42%)	66 (40.91%)	21
to play	88 (98.84%)	88 (55.68%)	36
to pull	85 (91.86%)	85 (54.12%)	33
to push	90 (97.67%)	90 (45.56%)	28
to reside (live)	80 (90.7%)	80 (45.0%)	26
to run	71 (77.91%)	71 (46.48%)	20
to scratch	76 (83.72%)	76 (52.63%)	29
to see	90 (98.84%)	90 (40.0%)	21
to shoot (an arrow)	75 (86.05%)	75 (28.0%)	12
to sing	88 (95.35%)	88 (52.27%)	29
to sleep	88 (97.67%)	88 (28.41%)	18
to smell (perceive odor) [transitive]	68 (77.91%)	68 (17.65%)	8
to sow (broadcast, scatter seeds)	48 (54.65%)	48 (47.92%)	15
to spit	74 (84.88%)	74 (56.76%)	29
to stand	88 (100.0%)	88 (21.59%)	11
to steal	86 (97.67%)	86 (17.44%)	9
to think (reflect)	79 (87.21%)	79 (37.97%)	19
to throw	92 (98.84%)	92 (55.43%)	40
to vomit	74 (84.88%)	74 (20.27%)	5
to walk	91 (98.84%)	91 (40.66%)	25
to wipe	88 (94.19%)	88 (42.05%)	25
today	87 (98.84%)	87 (44.83%)	23
tomorrow	88 (100.0%)	88 (47.73%)	26
twenty	71 (81.4%)	71 (30.99%)	12
two	86 (98.84%)	86 (12.79%)	6
we [first person plural inclusive]	76 (84.88%)	76 (34.21%)	11
wet	85 (97.67%)	85 (48.24%)	26
what	77 (84.88%)	77 (53.25%)	26
where	73 (84.88%)	73 (56.16%)	29
white	87 (100.0%)	87 (27.59%)	9
who	88 (100.0%)	88 (32.95%)	21
yellow	84 (96.51%)	84 (32.14%)	13
yesterday	88 (100.0%)	88 (46.59%)	28
you [second person plural]	71 (82.56%)	71 (29.58%)	11
young	60 (67.44%)	60 (61.67%)	27

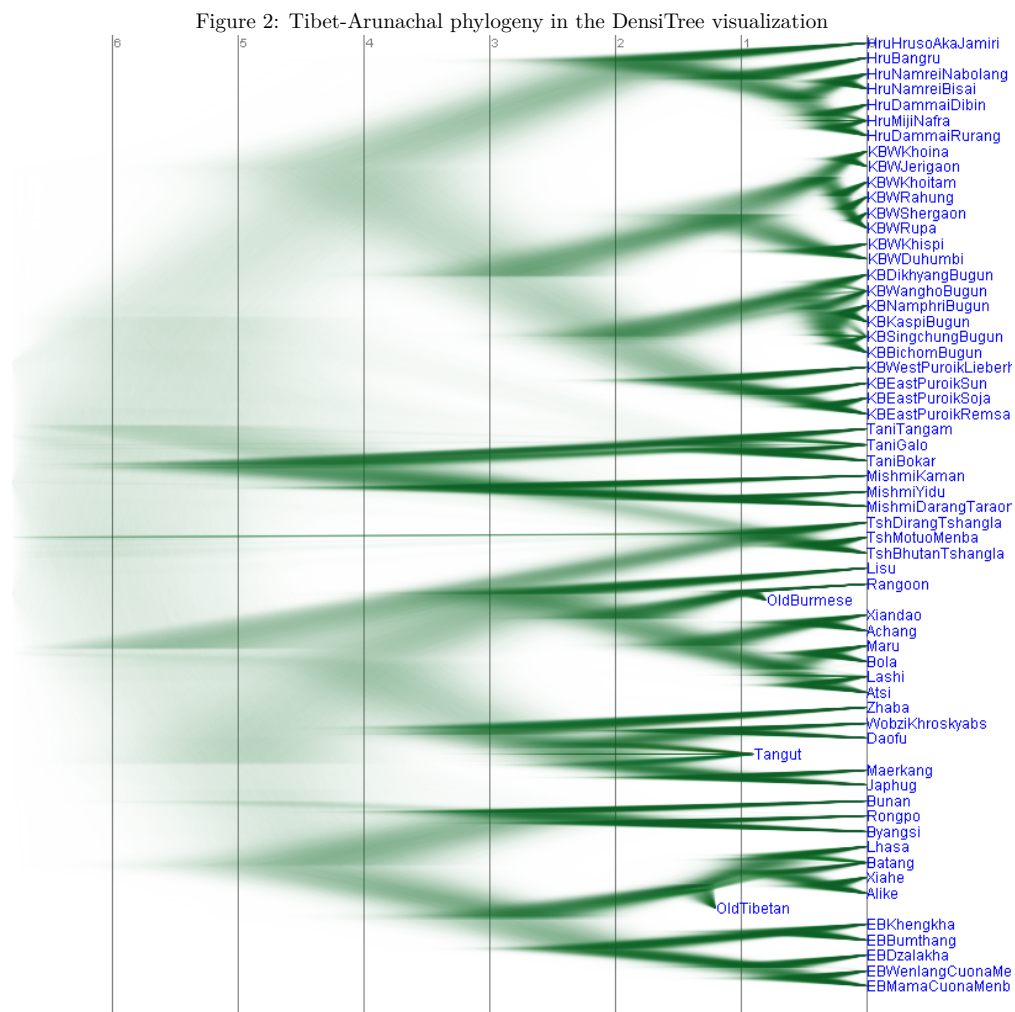
S4 Trans-Himalayan phylogeny

We use Birth-Death Skyline Serial prior model to reconstruct the Trans-Himalayan phylogeny. Blue bars show the 95 percent HPD time estimation. And the posteriors are shown next to the internal nodes.

Figure 1: Trans-Himalayan phylogeny



S5 Tibet-Arunachal in the DensiTree visualization



We generated a neighbor-net network to inspect the potential language contacts between different languages in Tibet-Arunachal area.

S6 Neighbor-net network

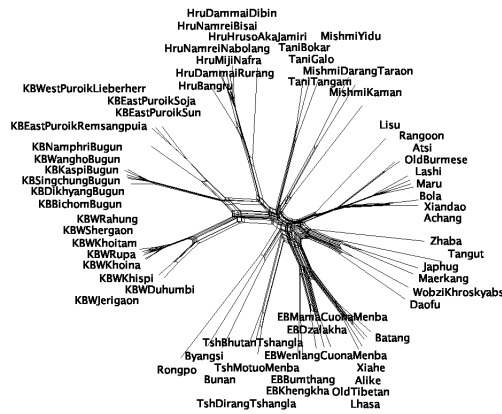


Figure 3: Neighbor-net visualization of selected languages in Tibet-Arunachal area

S7 Heatmaps

We calculate the amount of shared cognates pairwise across the languages in our dataset (figure 5). The Python code to generate the heatmap is uploaded in Open Science Framework.

Heatmap showing the pairwise genetic differentiation (F_{ST}) among 100 populations. The color scale ranges from 0 (dark purple) to 0.200 (dark red). The populations are listed on both the x and y axes, ordered by their geographic location from north to south. The diagonal is white, indicating $F_{ST} = 0$ for self-comparisons. The heatmap shows a clear pattern of increasing genetic differentiation with geographic distance, with the highest F_{ST} values (darkest red) observed between populations at the extreme north and south ends of the distribution.

References

- 11

- 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. 2022. Reconstruction of Proto-Western Kho-Bwa. forthcoming.
- Bodt, Timotheus A. and 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. and 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.
- Driem, George van 2015. Synoptic grammar of the Bumthang language. *Himalayan Linguistics*.
- Dzongkha Development Commission 2017. *Dzongkha-English-dZalakha lexicon*. Thimphu: Dzongkha Development Commission.
- Dzongkha Development Commission 2018. *Bumthangkha-Dzongkha-English lexicon*. Thimphu: Dzongkha Development Commission.
- Lieberherr, Ismael 2015. A progress report on the historical phonology and affiliation of Puroik. In Konnerth, Linda, Stephen Morey, Priyankoo Sarmah, and Amos Teo (eds.), *North East Indian Linguistics (NEIL)* 7, 235–286. Canberra: Asia-Pacific Linguistics Open Access.
- Lieberherr, Ismael 2017. *A Grammar of Bulu Puroik*. PhD thesis Universität Bern, Bern, Switzerland.
- 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 yánjiū* 《门巴语方言研究》 [*A study of Menba*]. Běijīng Mínzú chūbǎn shè [民族出版社].
- Lǐ, Dàqín 2004. *Sūlóngyǔ yánjiū* 《苏龙语研究》 [*A study of Sulong*]. Běijīng: Mínzú chūbǎn shè [民族出版社] 1 edition.
- Post, Mark 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.
- Remsangpuia 2008. *Puroik phonology*. Shillong: Don Bosco Centre for Indigenous Cultures.
- Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and 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.
- Simon, Ivan M. 1979. *Miji language guide*. Shillong: Philological Section, Directorate of Research, Govt. of Arunachal Pradesh.
- Simon, Ivan M. 1993. *Aka language guide*. Shillong: Research Department, North Eastern Frontier Agency.
- Soja, Rai 2009. *English-Puroik dictionary*. Shillong: Living Word Communicators.
- 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, Tianshin 1993. *A historical-comparative study of the Tani (Mirish) branch in Tibeto-Burman*. PhD thesis University of California at Berkeley, Berkeley, USA.
- Tayeng, Aduk 1990. *Sulung language guide*. Shillong: The Director of Information and Public Relations, Arunachal Pradesh.
- Yangzom, Deki and Marten Arkesteijn 1996. *Khengkha lessonbook*. Thimphu: SNV Bhutan.

4.5 Retrospective

Various scales of population movements within the MSEA language area occurred constantly during prehistorical times. Moreover, many great ancient civilizations originated in MSEA, and were influential in the languages spoken in the area throughout history. As a result, Sino-Tibetan languages are mutually influenced within the language family and have long-term and intimate language contact externally. The contact-induced language changes may be incorporated well in the native languages; thus, these features may be considered to be native features. Therefore, identifying the internal structure of the Sino-Tibetan languages, particularly given the limited available data, is clearly challenging.

This was a thought-provoking project for all the authors who participated in the study. We recompiled the lexical data several times, including the collection of new data and the revision of the cognate annotation. To avoid sampling bias, we also tested several methods. Each revision suggested a new way of interpreting our findings, and we eventually arrived at a version on which all the authors, from linguistics and data analysis perspectives, could agree. The following subsections are brief summaries of the challenges and our reflections.

4.5.1 Reasons for Data Recompile

The initial goal of the study was to investigate two specific language subgroups, Kho-Bwa and Hrusish, and their relationships with the languages in Sagart et al. (2019). Therefore, the first round of data collection only included Kho-Bwa and Hrusish. We then realized that many Sino-Tibetan languages, which we suspected were related to Kho-Bwa and Hrusish, were not included in the sampled languages. Therefore, we conducted a second round of data collection in order to include more languages from the Himalayan mountain region. The second round of data collection was challenging because we rarely found a sufficient amount of lexical data or well-organized documents for language subgroups in the Himalayan mountain region in the online data archives. Finally, we were only able to add 35 Sino-Tibetan languages to our study.

Due to political or geographical factors, the 35 Sino-Tibetan languages are spoken in areas to which there is limited access. Because the languages are spoken in such a confined and isolated region, only a few scholars were able to visit the villages. The prehistorical languages and their migration history are both understudied. These impediments prevent these languages from being included in large-scale linguistic studies. We acknowledge the difficulty of conducting research with limited resources; thus, our lexical data are stored in formats other than CLDF, such as Excel and tab-delimited plain text, to make other scholars' work more efficient. We hope that our research will encourage scholars to include the languages that are spoken in the Himalayan mountain region in large-scale surveys in the future.

4.5.2 Challenges of the Sampling Bias

We put effort into data preparation and statistical modeling to decrease the effect of sampling bias. First, we expanded the subgroups in Sagart et al. (2019) that were represented by only one language to two or more languages. Despite the fact that we could classify languages into groups, this does not imply that we could arbitrarily choose one language to represent an entire group because each language has its own pattern of language change as a result of factors such as historical contact or the migration history of the speakers. Even if one were to use one language to represent an entire subgroup, one would need to discuss the characteristics of a language subgroup and then choose a representative language that was the best match for the characteristics instead of choosing one language arbitrarily. To bypass the typological linguistic discussion, we expanded the subgroups to be represented by more than one language. Second, we experimented with the methods suggested by Sagart et al. (*ibid.*), and used one language from each subgroup to infer the Bayesian phylogeny. However, this also falls under the purview of the same issue of representation. The bootstrapping method, which involves repeatedly sampling one language from each subgroup to infer a Bayesian phylogeny tree, was considered. This approach was also not appropriate. We quickly realized that the bootstrapping concept was far removed from the goal of the study, which was to investigate the relationships among a set of given languages, as the bootstrapping method was designed to seek out modern languages that were most similar to ancient languages such as Old Chinese, Old Tibetan, and Old Burmese. We finally established the two-stage Bayesian analysis to overcome the sampling bias.

4.5.3 Challenges in Data Interpretation

To date, human activities along the two sides of the Himalayan ridge during prehistorical times remain unclear. We attempted to infer the movement of speaker populations in prehistorical times, but had very limited information to assist us to provide a complete image of the expansion route from northern China to present-day Arunachal Pradesh and beyond. The paragraphs below present the research results of genetic and archaeological studies, as well as our opinion regarding the link we are missing to infer the expansion of Sino-Tibetan languages in this area.

4.5.3.1 Anatomically Modern Human Diaspora

According to archaeological, genetic, and linguistic studies, the most plausible hypothesis is that the language families in EA and MSEA began to differentiate and expand due to the Neolithic agricultural revolution, during which lifestyles changed from hunter-gathering to farming. Diamond (2003) provided a review of the apparent strengths and weaknesses of the Neolithic revolution, and applied the analysis to nine cases of language dispersion across the world. Practicing agriculture apparently had three main benefits for societies. First, food production yielded much larger quantities than did hunting and gathering, thus enabling the population size to increase steadily. Second, establishing settlements and accumulating food to cope with the fallow season drove anatomically modern humans (AMHs) to develop complex technologies and social

stratification. Third, crowded settlements, despite increasing the risk of diseases spreading, promoted enhanced herd immunity for farming populations (Diamond, 2003). Thus, the homelands of language families are often associated with the origins of agriculture.

AMHs originated in Africa, and entered EA and SEA via multiple waves of dispersal. This model is largely supported by population genetics and by cranial phenotype data (Reyes-Centeno et al., 2014). The question is how and when they expanded throughout the MSEA. The most prominent hypothesis was associated with the findings of a craniometric study that suggested a two-layer hypothesis for the population of the MSEA (Matsumura et al., 2019). The “layers” represent two distinct ancestral groups expanding in two distinct periods. Scholars believe that the first group (“first layer”) of AMHs entered EA via the SEA landmass prior to 65–50 thousand years ago (kya) (Reyes-Centeno et al., 2014). Around 43.5 kya, the hunter-gatherer societies in SEA established the Hòabinhians (Hòa Bình, Vietnam) culture, which could be identified via stone tools and pottery, and buried their dead without offerings (ibid.). The Hòabinhian culture was originally thought to have been restricted to northern Vietnam. In 2006, Hòabinhian rock shelters in Yunnan, Southwest China, were found, which proved that the pre-Neolithic hunter-gatherers had also occupied southwest China (Ji et al., 2016). Hòabinhians are thought to be related to ancestral Andaman, Australian, Papuan, and Jomon groups. Later, the “second-layer” AMHs possibly came from northeast Asia and swept across central China during the early Neolithic period, which has been ascribed to their agricultural lifestyle. The descendant groups of the second layer expanded into SEA after 4 kya (Matsumura et al., 2019). The population of the second-layer AMH partially replaced the pre-Neolithic hunter-gatherers of the first layer. Furthermore, the population of the second layer is thought to be the ancestor of the present Sino-Tibetan people.

4.5.3.2 Language Family Differentiation During the Neolithic Period

The Neolithic agricultural revolution had multiple origins in Asia, including in central China between the Yellow River (黃河) and the Yangtze River (長江) basins, as well as in Man Bac in Vietnam (Bellwood, 2005; Diamond, 2003). The crops cultivated along the Yellow River and the Yangtze River were millet and rice, respectively. According to comparative linguistics, the Sino-Tibetan and Hmong-Mien language families first began to differentiate along the Yellow River and the Yangtze River, respectively.

In the north, the archaeological findings indicate that foxtail millet, pig, sheep, rice plant, cattle, and horse domestication/cultivation occurred at around 7–9 kya in the Yellow River region in association with the Yangshao (仰韶) culture (about 5–7 kya). Based on Sagart (2011b) (also see Baxter and Sagart (2014)) reconstructed words for agricultural and animal domestication, such as 稷 **[ts]ək* “millet (*Setaria italica*)”, 田 **lʲiŋ* “field”, and 豚 **lʲu[n]* “young pig”, he hypothesized that the AMH group lived in the Yellow River region associated with proto-Sino-Tibetan speakers. He further stated that the Yellow River region was where the first split into the Sino-Tibetan language family occurred. The southward migration expanded to the coastal

area. Some linguists believe that the group that lived along the coastal area might have spread to Taiwan, and have become the ancestor of Austronesian. The evidence is supported by material culture (Sagart et al., 2019).

The group of farmers expanded not only into southern China, but also into the west, which is currently the north of Tibet. This westward expansion corresponds with archaeologists' findings that the territory shows signs of millet domestication. Farmers also established territories on the Tibetan plateau, and then migrated into present-day Nepal, Bhutan, and Myanmar.

4.5.3.3 Debates About the Sino-Tibetan Language Family's Homeland

The version of the Neolithic agricultural expansion that we presented previously cannot fully explain the Sino-Tibetan languages that are spoken in the Himalayan mountain region. There is evidence indicating that, 9 kya, nomadic hunter-gather societies were based along the southern flank of the Himalayan mountains. They migrated seasonally between the valley and the Tibetan plateau, and signs indicate that a human population had been based on the Tibetan plateau since 6.9 kya. Some populations in this area, including in northeast India, still practice a semi-hunter-gatherer lifestyle.

The millet farmers only adapted genetically to the high altitude environment (low oxygen levels) in the high mountains in approximately 4 kya. This research outcome raised several questions. Were these hunter-gatherers also Sino-Tibetan speakers? If not, did the Sino-Tibetan speakers completely replace the local population, or were the Sino-Tibetan languages expanded via cultural transmission rather than via demographic expansion?

Four different hypotheses can be suggested here: We produced three of them, and one was provided by Blench and Post (2013). The first is that the millet farmers replaced the local population and continuously expanded into northeast India. The second is that the millet farmers introduced the languages to the local population and caused a language shift. The third is that the northeast Indian Sino-Tibetan speakers came from elsewhere, probably southern China, but the Sino-Tibetan languages spoken on the Tibetan plateau, Nepal, and other Himalayan valleys are associated with the northern Sino-Tibetan speakers. Fourth, the hunter-gatherer societies around 9 kya were the ancestors of Sino-Tibetan speakers.

The first and second assumptions follow the northern origin hypothesis. We cannot confirm which scenario is the true scenario due to the absence of population genetic studies.

The third hypothesis originates from archaeological findings that indicate that the Sino-Tibetan speakers in northeast India may have been associated with the Hòabình culture. There were also suggestions in the existing research that Tani speakers, a group of Sino-Tibetan language speakers who live in what is currently northeast India, came from somewhere in southern China (their folklore states "a place where the sun rises").

Our clues, which were derived from the linguistic and archaeological research, ended at the

Tibetan plateau and the Himalayan valley. We had no further resources to infer how the languages came to be spoken in what is currently northeast India. There may be other factors that triggered population migration after the Neolithic period; for example, warfare or religion. Arunachal Pradesh is located in the eastern Himalayas, and is bordered by the Himalayan ranges and by the Tibetan plateau in the north, and by the alluvial plains of Assam in the south. Linguistic diversity in this area might be partially explained by the area having served as a mountain refuge for diverse and successive populations that had migrated from both the Tibetan plateau and from the Assam plains for millennia, whereas other populations moved into and settled across the more easily accessible, inhabitable, and arable stretches of land. The idea sourced from the “Zomia” geographical area extends from SEA into the Himalayas. Zomia was first proposed by Schendel (2002), and was further elaborated on by Scott (2009). However, we agree with the criticism of the Zomian theory provided by authors such as Michaud (2010), Lieberman (2010), and Brass (2012). According to these authors, the people currently inhabiting SEA’s mountain ranges may not have always “chosen” to migrate from their original homelands to avoid being “enslaved” by “nation states”. Instead, they may have been driven out by more technologically advanced and numerous migrant populations, and may have encountered linguistic, cultural, and ethnic assimilation, or worse. Overall, the reasons for migration to modern-day northeast India, particularly Arunachal Pradesh, remain unknown. However, we have a language phylogeny as a basis for study (see Chapter 4, and we await more archaeogenetic or population genetic findings to assist us to move forward.

The fourth hypothesis stems from Blench and Post (2013). Other homelands are being proposed by linguists (see Chapter 4). However, the northern origin hypothesis has received the most support from existing Bayesian phylogenetic analyses (Sagart et al., 2019; Zhang et al., 2020; Zhang et al., 2019).

We would like to state two caveats here. In our third project (Chapter 4), we did not survey many archaeological and genetic findings. We followed the existing theory, the northern China origin, to interpret our findings. We naturally linked the population of the Tibetan plateau with the population in Arunachal Pradesh based on the geographical area (as per our hypotheses one and two). We only realized that the link between Tibetic languages and Arunachal Pradesh Sino-Tibetan speakers was much weaker than we assumed when we began to survey a wider range of articles. We are not yet close to understanding the prehistory of this area. We are optimistic that more archaeological findings will be presented in the near future, when we will be one step closer to understanding the population movement in prehistoric times in this area.

The second caveat is as stated in Bellwood (2005). Human prehistory and the language family expansions cannot be fully explained by the transition from hunter-gatherer to farming lifestyles. The most common assumption does not consider the pros and cons of practicing foraging. Moreover, hunter-gatherer societies may also expand into new locations. We hope that more studies that consider a different perspective will be presented in the future.

Overall, we look forward to further fine-grained studies that can contribute to the complete picture of how languages differentiated and expanded in this vast geographic area.

4.6 Future Work

Due to data collection challenges, our lexical data set consists of 84 languages, covering more than 20% of the entire Sino-Tibetan language family. We were unable to include all of the subgroups spoken in the area of our research focus. As a result, our Bayesian phylogenetic tree is only the beginning. More research is needed to broaden the phylogeny to include more understudied Sino-Tibetan languages.

The project was conducted at the same time that we were developing a lexical data annotation method. It is a pity that we could not annotate partial cognates because the study by Sagart et al. (2019) did not annotate partial cognates. We were also unable to use morpheme annotation due to time constraints. Since the article's acceptance, a linguist has expressed interest in applying our method to a different data set. We took advantage of the opportunity to discuss our proposal for adding morpheme annotation to improve the transparency of the lexical data. Further linguistic research involving clear morpheme annotation that can be used to improve the Sino-Tibetan phylogeny is anticipated.

Finally, interpreting the results requires the input of experts in various disciplines. Furthermore, revealing human activities in the Himalayan region during the prehistoric period is dependent on archaeological and genetic research. We have shared the information in the article, as well as this retrospective, in the hope that it would assist such researchers to interpret their results and to exchange their findings with Bayesian phylolinguistic researchers.

Chapter 5 Discussion and Conclusion

Historical linguistics studies seek to improve our understanding of human (pre)history by providing evidence from a linguistic perspective. The questions to be answered are frequently associated with the global and regional routes of human diaspora, the relationships among population movements, and the spread of cultures, languages, and lifestyles.

We presented an advanced workflow for generating and analyzing large-scale lexical MSEA data that combined the classical comparative method and computational modules (Wu et al., 2020). We addressed the inadequacy of identifying cognates at the word level, and proposed a new approach for annotating morpheme cognacy, which we trialled using a Bayesian phylogenetic analysis (Wu and List, 2022). We also highlighted the significance of surveying and incorporating knowledge from other research domains into historical linguistic research (Wu et al., 2022). Overall, the methods proposed in this dissertation addressed six perspectives.

1. Standardization (Chapter 2)
2. Aggregation (Chapter 4)
3. Annotation (Chapter 3)
4. Transformation (Chapter 3)
5. Application (Chapter 4)
6. Interpretation (Chapter 4)

Chapters 2, 3, and 4 describe three projects demonstrating the stages involved in using lexical data belonging to two language families. We drew examples from data sets that we digitized in the CLDF format to show the connections among the methods introduced in the three projects (including Chen 2012, Clark 2008, Hsiu 2015, Mão 2004). Furthermore, we provided our reflections regarding some of the solutions we provided, and suggested caveats.

5.1 Standardization

Standardization refers to homogenizing the diversity of data sets as the same format and at the same level of resolution. Moreover, the common links among multiple data sets were created because standardization should be based on the same set of principles or reference catalogs. Standardization is a critical aspect in the success of any data-driven study; however, it is also the most tedious process.

Most of the linguistic data were not digitized, or were digitized in different formats. The majority of the linguistic data we used in this project needed to undergo a certain degree of

standardization before they could be used for computational, computer-assisted, or even manual analyses. In our workflow, we began by either digitizing the data sets (for example, the 116 words from Máo (2004)) or by unifying the data formats in various other formats (such as Chen (2012), Ratliff 2010, Clark 2008, Hsiu 2015, and many other data sets).

The standardization process was described in Chapter 2; we followed the principles for standardizing the formats according to the CLDF style. The task of standardization also included adjusting all the studies to the same resolution. In linguistic terms, this means that all the lexical data should be presented on an equal level, either as words or as morphemes. However, fixing a truncated word at the stage of standardizing each individual data set is almost impossible. Consider the two data sets provided by Chen (2012) and Ratliff (2010) as examples of the issue of “truncated words”. Chen (2012) listed 888 concepts in 23 Hmong-Mien languages in their full word forms. Ratliff (2010) presented the morphemes in eleven Hmong-Mien languages; her work followed the classical comparative method to reconstruct the proto-Hmong-Mien languages. This is a limitation of applying the unmodified classical approach to the analysis of MSEA language data. The best way to restore the full word forms is to have a linguistic expert present the full words, or to cross-reference different data sets. Since this cannot be done via a computational approach, we will not discuss this issue further.

A few points about changing data into a standard format did not receive further elaboration in Chapter 2. Therefore, the following subsections explore three aspects, namely GTP conversion, the standardization of tonal markers, and tokenization. In addition, our workflow used the orthography profiles for phonetic sequence standardization. We also discussed the pros and cons of using an orthography profile as a standardization tool.

5.1.1 Graphemes to Phonemes

Early linguistic fieldwork in the MSEA area often had the aim of spreading a particular religion. Therefore, linguists and missionaries frequently used a foreign orthographic system (such as the Roman alphabet) to document languages or to translate the Christian Bible if the local languages lacked their own writing systems. These systems have been used continuously by the local communities up to the present day. For example, the Ntawv Hmoob writing system (Romanized Popular Alphabet, or so-called RPA) was established by William Smalley, Linwood Barney, and Yves Bertrais in the 1950s (Smalley et al., 1990) to document Hmongic languages, and White Hmong communities worldwide are still using the RPA system. Such orthography is practical when translating the Bible for preaching purposes. Even though the words may not be “loyal” to the original pronunciations, the local communities are able to learn and write, and to possibly adapt the system to their languages (for example, pop songs in Formosan languages in Taiwan are currently being written in the Roman alphabet). The Bible is available in 2,299 languages,¹ which means that the text forms a valuable cross-linguistic corpus for computational linguistic studies.

¹The number comes from the website <https://www.wycliffe.org.uk/>.

Another common practice among linguists when documenting a language or presenting a cross-linguistic data set is to use a combination of the IPA and customized phonetic symbols. For example, the word “hair” in the Northern Qiandong language (Glottolog: nort2747) is presented as $\underset{\cdot}{l}u^1$ in Ratliff (2010, p. 46). The $\underset{\cdot}{l}$ is not the standard IPA symbol, but is a language-dependent IPA extension. Linguists tend to present words in a way that is as close to the original pronunciation as possible; however, the IPA system does not cover all the possible sounds in the world’s languages. Therefore, customized phonetic symbols are developed to represent the sounds that are lacking in basic IPA.

Depending on the purpose of individual studies, the same words in different data sets may be presented in different combinations of graphemes ². For example, the word 頭髮 *tóu fā* “hair (head)” is represented by $qa^{33}\underset{\cdot}{t}ju^{33}qho^{33}$ or 毛 *máo* “hair (body)” $qa^{33}\underset{\cdot}{t}ju^{33}$ in Chen (2012). The words in Chen (ibid.) are presented in their complete forms, including prefixes. The onset position of the morpheme that represents “hair” is analyzed differently. Chen (ibid.) presented the onset as $\underset{\cdot}{t}j$ and Ratliff (2010) analyzed it as $\underset{\cdot}{l}$.

Documenting lexical items using phonetic symbols can be seen as a type of “abstraction”; missionaries and linguists are trained to represent the surface pronunciation of a word via a combination of phonetic symbols based on what they hear and on the demands of both the language phonology and the research purpose. In our definition, these various symbols are *graphemes*, which refers to the word forms in the original data set. Data sets with different graphemes cannot be compared by computer programs because graphemes such as *ng* and η are not the same for a computer program, even though *ng* is often used to represent η . Therefore, the word forms in different data sets should be standardized before being merged in order to increase comparability.

There are two ways of standardizing individual data sets before merging them. The first is to treat one data set as a reference and to convert other data sets to match the reference, while the other is to convert all the data sets to match a third-party orthography; for example, standard IPA or customized sound classes (cf. Holman et al., 2008). The first method is straightforward as long as the expert knows the sound inventory well, or the language’s sound inventory has already been agreed upon by linguists. The second method is commonly used when merging a large number of languages or cross-language families. For example, the Automated Similarity Judgment Program (ASJP) database summarizes the phonetic diversity of more than 5,000 language variants across the world as a system of seven vowels and 34 consonants, not counting modifiers that act as diacritics, which means that many IPA symbols are linked to one sound class. For example, the E sound class includes *a*, *æ*, *ɛ*, *œ*, *ø*, and other vowels ibid. For wide-ranging cross-linguistic studies, such data sets provide a bird’s-eye view of the relationships across various languages (cf. Jäger, 2018). The issue of converting data into sound classes is that, once the data are reduced to sound classes, if the information about the raw source is not preserved, it becomes difficult to recover the original graphemes, particularly for under-researched languages. Furthermore, such

²The research purpose would also influence the different levels of “completeness”.

large-scale databases cannot be used in projects dealing with semantic evolution or phonological change, even though this was possible when using the raw data.

We adopted the second approach. Our standardization guidelines were the CLTS sound class and the IMNCT template. The graphemes in the data sets Chen (2012), Ratliff (2010), Clark (2008), Hsiu (2015) and Mão (2004) were all converted from their row form to the CLTS sound classes. Morpheme boundaries were also added during the GTP process with the assistance of the IMNCT template. The benefit of using the CLTS transcription system is not only that the comparability among data sets is increased, as the transcriptions in various sound class systems can also be converted. During standardization, *CLDFbench* (Forkel and List, 2020) will report if there are graphemes in the data set that were not converted to the CLTS sound classes. This evaluation ensures the compatibility among data sets on the phoneme level.

5.1.2 Tokenization

Tokenization in natural language processing is a frequent task, and involves dividing a text into a collection of words. Tokenization in the study of language comparison means dividing words into a list of phonemes. For example, for the word 父 *fù* “father” in Mandarin Chinese *fu*⁵³, we can tokenize the phonetic string into *f*, *u*, and the high-falling tone⁵³. This is an essential step prior to phonetic sequence alignment in computational or computer-assisted analyses. The outcome of phonetic sequence tokenization impacts significantly on the alignment, and subsequently controls the accuracy of computational cognate judgments and sound correspondences.

Some linguists would argue that it seems unnatural to tokenize a string of phonetic sequence into phonemes because humans would not pronounce the phonemes one after another when pronouncing a word. For example, a Mandarin speaker would not pronounce the phonetic strings and then pronounce the tone. It is true that tokenizing words according to a sequence of phonemes does not reflect how humans articulate a word. It is an analysis that is based on the syllable structure in order to work with computer programs.

Tokenization involves understanding the syllable structure of a language. Chapter 1 presented the syllabic structure in MSEA languages. Even though the underlying template tends to be fixed, there are still various ways of analyzing a syllable. Chen (2012) presented the sound inventory for each language in two large tables, onsets and rimes. In the same work, the author presented a large table in which he compared vocabularies between Northern Qiandong (Glottolog: nort2747) and Chuanqiandian (Glottolog: chua1248) with different templates, which do not only include the categories of onset and rime. Each of the onsets and rimes in the analysis is further defined as the 頭 *head*, the 身 *body*, or the 尾 *tail*; he then merged the onset-tail and the rime-head with the medial (ibid., p. 18).³ The boundary between onset and rime in this analysis is ambiguous.

The onset-rime model is used widely across various language families, and the fine-grained

³The rime head is the on-glide position, as stated by Ratliff (2010). The difference between the two analyses is that the onset is divided into two parts rather than into three parts.

templates are language dependent. Different templates would influence the tokenization, and may influence the method of conversion. Consider the previous example “hair” (here we use 毛), written as $qa^{33}\text{tj}u^{33}$. The qa^{33} is a prefix; our analysis considers it to be a different morpheme.⁴ The $\text{tj}u^{33}$ contributes to the semantic meaning of “hair”. Thus, we show the tokenization of qho^{33} according to different templates. Tables 5.1 and 5.2 show how the word forms are converted and aligned in different analyses. Table 5.1 shows that the j represents palatalization, and the j in the first half of Table 5.2 shows that the j is a medial. The second half of Table 5.2 also treats the j as a symbol for palatalization. The treatment of j depends on the data collector’s use of graphemes or the user’s research purpose. Chen (2012, p. 50) explained that j was a representation of palatalization. Therefore, the tj in this particular data set is better treated as t^j , as it occupies the initial position in the five-part template (as in the second half of Table 5.2).

	Onset	Rime	Tone
Grampheme	t^j	u	³³
Phoneme	t^j	u	³³

Table 5.1: Onset-rime template. Although the tone is attached to the vowel, the tone is listed in another column for purposes of computational analysis.

	I	M	N	C	T
Grampheme	t^j	j	u	-	³³
Phoneme	t^j	j	u	-	³³
Alternative analysis					
Grampheme	t^j	-	u	-	³³
Phoneme	t^j	-	u	-	³³

Table 5.2: Five-part template. In the introduction and the first project, we indicated that we did not consider the medial counts to be part of the onset or the rime.

Another example shows that the treatment of diphthongs can sometimes be difficult: Hmong-Mien languages have complex diphthongs, such as *ia*, *ua*, *ie*, *ue*, and so on. Therefore, we used the word “house” in the Nunu language (Glottolog: nunu1247) *pia*³⁵ as an example. Tables 5.3 and 5.4 show the treatment of diphthongs in two different templates. The difficult part is the analysis in Table 5.4. Diphthongs use two simple vowels to describe the movement of the tongue or the oral cavity. Therefore, a diphthong can be seen as one unit, and can be placed in the slot for a nucleus (see the first half of Table 5.4). Nevertheless, the *i* can also be treated as the head of rime, and one then moves on to pronounce *e*. Therefore, the *e* is the nucleus (rime body). As Chen (ibid.) stated, the medial contains the rime-head and the onset-tail, and the *i* can be analyzed as a medial. However, the medial in HM languages only allows *-j-*, *-l-*, and *-w-* (Ratliff, 2010). Therefore, the *i* is analyzed as having the same phonetic position as *j* in this template.

Although the templates will not change the surface pronunciation, the examples above show that the template on which the tokenization method is based will influence the inference of sound

⁴It can be analyzed as a sesquisyllable according to Matisoff’s analysis.

	Onset	Rime	Tone
Grampheme	p	ia	³⁵
Phoneme	p	ia	³⁵

Table 5.3: Onset-rime template.

	I	M	N	C	T
Grampheme	p	-	ia	-	³⁵
Phoneme	p	-	ia	-	³⁵
Alternative analysis					
Grampheme	p	i/j	a	-	³⁵
Phoneme	p	i/j	a	-	³⁵

Table 5.4: IMNCT template.

correspondence. Table 5.6 and 5.8 use the word “house” in different Hmongic languages to show the alignments when using different templates. Table 5.6 shows that *pl-* corresponds to *p^j-*, and the rime correspondences are *-ε*, *-o*, *-e*, *-ia*, *-ei*, and *-ui*. The first half of Table 5.8 shows that *p-* corresponds to *p^j-*, *-l-* corresponds to medial deletions, and nucleus correspondences are the same as shown in Table 5.6. The second half of Table 5.8 shows that the *p-* is constant, the medial *-l-* corresponds to *-j-*, and the nucleus mainly remains the same except that the *-ia* is not *-a*.

Doculect	Grapheme	Phoneme	Onset	Rime	Tone
Central Guizhou Chuanqiandian 高坡	ple ¹³	p ^j le ¹³	pl	ε	¹³
Eastern Baheng 毛坳	pjo ³¹³	p ^j o ³¹³	p ^j	o	³¹³
Dongnu 七百弄	pje ⁵³	p ^j e ⁵³	p ^j	e	⁵³
Nunu 西山	pia ³⁵	p ^j ia ³⁵	p	ia	³⁵
Numao 瑶麓	pjei ¹³	p ^j ei ¹³	p ^j	ei	¹³
Younuo 優諾	pui ³³	p ^j ui ³³	p	ui	³³

Table 5.6: Onset and rime template.

Doculect	Grapheme	Phoneme	I	M	N	C	T
Central Guizhou Chuanqiandian 高坡	ple ¹³	p ^j le ¹³	p	l	ε	-	¹³
Eastern Baheng 毛坳	pjo ³¹³	p ^j o ³¹³	p ^j	-	o	-	³¹³
Dongnu 七百弄	pje ⁵³	p ^j e ⁵³	p ^j	-	e	-	⁵³
Nunu 西山	pia ³⁵	p ^j ia ³⁵	p	-	ia	-	³⁵
Numao 瑶麓	pjei ¹³	p ^j ei ¹³	p ^j	-	ei	-	¹³
Younuo 優諾	pui ³³	p ^j ui ³³	p	-	ui	-	³³
Alternative analysis							
Central Guizhou Chuanqiandian 高坡	ple ¹³	p ^j le ¹³	p	l	ε	-	¹³
Eastern Baheng 毛坳	pjo ³¹³	p ^j o ³¹³	p	j	o	-	³¹³
Dongnu 七百弄	pje ⁵³	p ^j e ⁵³	p	j	e	-	⁵³
Nunu 西山	pia ³⁵	pi/ja ³⁵	p	i/j	a	-	³⁵
Numao 瑶麓	pjei ¹³	p ^j ei ¹³	p	j	ei	-	¹³
Younuo 優諾	pui ³³	p ^j ui ³³	p	-	ui	-	³³

Table 5.8: IMNCT template

Consistency is the most important aspect at this stage; therefore, we do not specify which template is the most “accurate”. Users are free to choose whichever template is best suited to their research purposes. If necessary, one can even establish a customized tokenization template, as long as users remain consistent throughout the entire research project.

Our four Hmong-Mien data sets are tokenized according to the IMNCT template. One critical point in our analysis is that Chen (2012) analyzed the onset as having three different parts. Hence, the onset cluster *pz* (or similar types) in Chen (ibid.) should be analyzed as the onset head *p* and the onset body *z*. Our IMNCT template does not have a space for the onset body; therefore, the *z* is placed in the medial position. Therefore, we would expect that the sound correspondence of the medial in our analysis would be *-j-*, *-l-*, *-w-*, and *-z-* (or similar).

5.1.3 Tones

There are a few systems of tonal marker annotations; the two most common annotations are numbers and diacritics. Depending on the factors that are actually annotated, the markers are used to annotate the tone category or the tone value. Tone category means labeling the tones, but the labeling does not necessarily reflect pitch changes. A classic example is Chinese tones with the combination of *ying* and *yang* and 平上去入 *píng shàng qù rù*, or the first to fourth tones in Mandarin Chinese. The tone values aim to describe the rising or falling pitch, or any pitch change, the most significant system being the five-level tone mark (五度標記法). Linguists often use numbers: ⁵⁵ means a high-level, ³³ means a mid-level, and ³⁵ means a mid-rising tone. The diacritics can also be used to represent either the category or the values.

A less common tonal marking system is the use of consonants to annotate the tones (see Table 5.9). For example, Heimbach (1969) uses the *Ntawv Hmoob* to present White Hmong vocabulary. The reason for developing such a system is the tendency toward vowel ending in White Hmong morphemes, except for *ŋ*. Therefore, the consonants that are treated as tonal markers are placed after the vowels. For example, the White Hmong word for “molar tooth” in Heimbach (ibid.) is written as *hniav puas* (lit. tooth molar); correspondingly, the *v* and the *s* are the mid-rising and the low-level tones. The *Ntawv Hmoob* orthography has continuously been used by different White Hmong societies across the world, including at present.

Tone category	Tone value	My conversion to tone value
b	┐	55
j	ㄣ	53
v	┐	34
	┐	33
g	┐	21
s	┐	11
m	┐?	11?

Table 5.9: The tone category and the tone value in Heimbach (1969).

The word “molar tooth” in the White Hmong data in Ratliff (2010) is written as *pua*¹. In this example, the *s* in Heimbach (1969) corresponds to the digit 1 in Ratliff (2010). However, the tonal markers between Ratliff (ibid.) and Heimbach (1969) are not one-on-one relationships.

A White Hmong speaker who has learned the orthography system can easily differentiate between the tonal markers and the alphabets. However, computers may struggle to work with such orthography if there no morpheme boundary is marked. For example, the word “people” is written as *tibneeg* “people, a person”. The *tib* means one (as a verb, the meaning changes to “to pile up”⁵), and that the *neeg* means “person” (ibid., p. 316). This disyllabic word has two tones: high-level *b* and low-falling *g*. Nevertheless, the morpheme boundary between *tib* and *neeg* is not marked. During the data standardization, assuming that we standardized the graphemes according to standard IPA, the computer program would treat the *b* as a regular consonant. The conversion would then be *tibnē*²¹ instead of *tī*⁵⁵*nē*²¹. We found other words in the dictionary that were written as disyllabic words; for example, *menyuam* “children” (lit. small-little) and *pojníam* “[married] woman” (lit. woman, the wife of) (ibid., pp. 125, 232). To improve the accuracy of data standardization, users should separate each morpheme using a symbol to ensure that computer programs are able to convert the graphemes correctly.

During our time working with tonal marker standardization, the question of *whether tone values could really be compared across languages* lingered constantly. Linguists determine the number of tones or contours using a set of principles without a universal standard. Linguists compare a list of words to test whether the pitch difference changes a word’s semantics. This contrast may be subjective. The speaker’s articulation and the receiver’s internal interpretation determine the correlation between pitch and semantics. Furthermore, the high, mid, and low tones are determined within the same language. The high and low tones are never compared across different languages. Therefore, the tone value ⁵⁵ in the White Hmong language may not be the same as ⁵⁵ in the Eastern Luopohe language on a frequency-related scale. Furthermore, various factors influence the articulations, such as gender, age, and language proficiency. It appears to be less than feasible to employ a universal standard to test the tones’ real values; therefore, it also appears to be unreasonable to compare tones across languages. Standardizing the tonal range also appears to be redundant.

In the realm of comparative linguistics, we should not dwell on the accuracy of tone values; that is, questions such as “How many Hz is high-level tone?”, “Is the high-level tone the same frequency (Hz) between A and B languages?”, and so on. These should be questions for other linguistic fields; for example, those seeking to improve sound recognition techniques. The purpose of comparative linguistics is to model the language changes within a community or under a condition; thus, we should only see the tonal marker as a property of a word that has the function of distinguishing

⁵A side note about the *tib*: The meaning “to pile up” is written as *teeb* in WOLD (<https://wold.clld.org>), but as *teeb* in Heimbach (1969) means lamp or light. Heimbach (ibid.) stated that the *ee* represented a nasalized vowel, and that *i* was pronounced like the *e* sound in the English word “we”. It is assumed that the *teeb* in WOLD might stem from a different source.

semantics, at least from the perspective of MSEA languages. Tone is an abstraction that assists us to understand the differences in languages. Therefore, standardizing the tonal markers can be beneficial from the perspective of data analysis. As stated in the previous subsection, *ng* and *ŋ* do not represent the same phoneme. The word 姨 “aunt” in Standard Mandarin, which is spelled *í*, is not the same as *i*³⁵ or *i*² (second tone) for a computer program or for a person who does not understand the Standard Mandarin tonal system. Imagining a data set such as Table 5.10 will create confusion for someone who is a beginner in Mandarin and is not yet familiar with converting the tones among diacritics, tone categories, or tone values. As another example, imagine that we were to run a cross-linguistic analysis of the merged data set of the White Hmong lexical data from Heimbach (1969) and the Hmong-Mien lexical data from Chen (2012). Without standardizing the tones in the two data sets, the word-final consonants would always be analyzed as consonants instead of tones. Therefore, standardizing the tonal markers as either diacritics or as numbers can improve the accuracy of analysis; this could also avoid confusion. Another benefit of tonal marker standardization is that it increases data re-usability. Numerous cross-linguistic data sets are now using either diacritics or numbers; for example, the raw data in Clark (2008), Hsiu (2015), and Mão (2004) were all marked using the five-level tone mark system. In addition, many published CLDF data sets also mark the tones using the same system. Therefore, the data sets can easily be extended.

Doculect	Concept	Value
Mandarin 1	head	t ^h ou ²
Mandarin 2	head	t ^h ou ³⁵
Mandarin 3	head	tóu
Mandarin 1	hand	ʂou ³
Mandarin 2	hand	ʂou ²¹³
Mandarin 3	hand	shǒu

Table 5.10: An imaginary lexical data set. The value column displays the standardized raw form. Mandarin 2 is based on a linguistic data set.

The WOLD is a concrete example indicating that data re-usability is increased by standardizing the tonal markers. Prior to the conversions, the tones were marked with diverse tonal markers. Such a data set can be used for visual inspections, but it cannot be analyzed by computer programs. The same data set has now finely been curated and is presented in a CLDF format, in which the tones have all been converted into numbers.

Finally, the knowledge about tones is still limited. Different languages may have different ways of using tones. Furthermore, what we currently know is still based on observations of a small number of languages. We are far from being able to quantify the frequency of tones accurately. Moreover, there are no computational programs that can assist us to disentangle all the linguistic topics surrounding tones. It is for exactly this reason that we must standardize the tonal markers at this point, and prepare a well-curated data set for future studies. Having a good data set is always

a starting point for discovering new fields for quantities analyses.

5.1.4 Pros and Cons of an Orthographic Profile

The GTP technique in computational linguistics converts graphemes into phonemes via either rule-based or statistical approaches (Yolchuyeva et al., 2019). The statistical GTP methods have extremely wide applications in the domain of natural language processing, such as sound recognition and speech synthesis systems (*ibid.*). The statistical approach relies on large-scale data and machine learning algorithms, according to which the conversion rules are automatically inferred from the finely curated training data. This method does not require experts to establish the conversion rules manually. Nevertheless, the data-driven approach is not suitable for the CALC framework for two main reasons. First, the application can only apply to languages that have a lot of available data. Second, the conversion rules are not presented clearly.

The orthographic profile in the CALC framework is a rule-based GTP approach that relies solely on the input of linguistic experts. The rule-based approach frees us from the requirement of having large data sets. Hence, this method can be applied to all languages, even understudied languages or those with small-scale data sets.

The orthographic profile is a comma- or tab-separated format; thus, it can be created in plain text editing software, which will enable linguists to create data sets on an ad-lib basis. Furthermore, each rule is displayed in an independent row to ensure that both the machine and the user are aware of the original and the converted phonemes. In addition, users can optimize individual rules in the rule-based GTP approach, which cannot be done in a straightforward manner in the statistical GTP approach.

An orthographic profile in the CALC framework addresses both conversion and tokenization. The number of rules is determined by the complexity of the graphemes, as well as by the number of languages and concepts. As the number of rules increases, the orthographic profile may become less readable, and it may be less possible to create consistent conversion and tokenization.

In the workflow, the computer program searches for the rules in the orthographic profile from top to bottom. If two rules contradict each other, the later rule can be ignored. Therefore, users have to pay attention to compatibility among the rules.

5.2 Aggregation

Although data aggregation is the common approach in data-driven studies, the principles have not been sufficiently addressed. Scholars consider data aggregation to be the simple combination of sparse data into a data set with rich information. However, the key principle is whether the data can retain the flexibility to be expanded or extracted. Maintaining flexibility requires well-prepared metadata, which entails an additional document to describe the data contained in the set. Each metadata entry is assigned a globally unique and consistent identifier to ensure that both

human and machine can use the identifier to connect multiple data sets.

After standardizing the data set into CLDF, the metadata include the language information and the definition of the concepts. The languages are annotated using the Glottolog code. The concepts are defined by the Concepticon database. Both Glottolog and Concepticon are our “links” to connect multiple CLDF data sets.

Three online language catalogs are commonly used in computational historical linguistics: Ethnologue, ISO 639, and Glottolog. The CLDF dataset takes the Glottolog code as the primary source for annotating languages’ metadata. Language varieties in the Glottolog database are linked to ISO 639-3 if the languages are also identified in the ISO 639-3 database. Glottolog also provides alternative names and the levels of endangerment. Therefore, annotating the languages in the CLDF data set using Glottolog codes almost simultaneously combined numerous pieces of information that may not have been provided by previously by the data set itself.

The Concepticon database is constantly growing, and currently features approximately 3,800 commonly used concepts from various concept lists, including the Swadesh list and variants thereof (Swadesh, 1955; Swadesh, 1964), as well as large concept lists that contain more than 800 unique glosses (Chen, 2012; Huáng and Dài, 1992). The database provides a unique identification number, a detailed description, and additional information for each concept. Mapping the vocabularies in a data set onto the Concepticon database can be seen as turning the implicit glosses into explicit concepts. It also creates a link to many existing data sets that are also linked to Concepticon, thus significantly increasing the potential benefits that one may extract from a single data set.

Overall, merging multiple CLDF data sets entails the following steps:

1. Extract the overlapping concepts.
2. Remove the concepts with low coverage.
3. Establish a list of sampled languages.
4. Prepare a script to use CLDFBench API (Forkel and List, 2020) to iterate through the concepts and the languages in each data set.

We demonstrate aggregating data sets into a full lexical data set in Wu et al. (2022); the Python script could be found in the supplementary materials.

5.3 Annotation

Annotating data means annotating the semantic meaning or syntactic features of morphemes, as well as computer-assisted cognate judgments. This step is essential when converting partial cognates into a binary matrix for a Bayesian phylogenetic analysis. Furthermore, the annotation

assists others to better understanding the data, thus increasing the odds of the data being re-used by others.

5.3.1 Morpheme Annotation

There are two annotation tasks at this stage. The first is the semantic meaning or the syntactic role of the morphemes. The second is highlighting the morpheme that linguists believe is “salient” in each word. Our analysis determined the salient morphemes based on whether the morpheme determined the semantic meaning or not. Our annotation scheme was introduced in Chapter 3. The material that was used to demonstrate the annotation in Chapter 3 was taken from the Sinitic languages. To show that the annotation method could also be used in other MSEA languages, we present our analysis of a Hmong-Mien language data set. We present the examples in tabular format in this section, but we encourage users to annotate morphemes with the assistance of *Edictor*.

During the revision of our paper, we received the following feedback: “[A]nnotating the compound words with head and modifier, and the head morpheme represents the salient morpheme”. In our study, we showed that always treating the head morpheme as the salient part was arbitrary. The subsections below provide a different analysis of compound words by categorizing compound words according to four different categories; we annotate the salient parts in bold font.

5.3.1.1 The Coordinative Type of Compound Words

The coordinative compound can be further divided into copulative and appositional. With regard to copulative compound words, all the morphemes have similar meanings. Our strategy was to select only one morpheme as the salient morpheme. Table 5.11 shows that the word “sieve” in Iu Mien is a copulative word that comprises two monosyllabic words with similar meanings. Sieve and Dustpan in Zao Min are synonyms, both of which are written as *kɛŋ* in Chen (2012). Therefore, we highlight the morpheme *kɛŋ* in this variety using bold font. The highlighted morpheme is called the salient morpheme in our framework.

DOCULECT	VALUE	TOKEN	MORPHEMES
Iu Mien	sjaŋ ³³ tɕei ³³	s j a ŋ ³³ + tɕ ei ³³	dustpan + sieve
Kim Mun	gjaɪ ³⁵ tθai ²¹	g j ai - ³⁵ + tθ ai ²¹	Plough + sieve
Biao Min	sɛ ³³	s - ɛ - ³³	sieve
Zao Min	hɛi ⁴⁴ kɛŋ ⁴⁴	h ɛi ⁴⁴ + k ɛ ŋ ⁴⁴	Plough + sieve

Table 5.11: The concept “sieve” in four Mienic varieties (Máo, 2004). The “-” means that the positions corresponding to the IMNCT template are empty.

However, the appositional cases are treated differently. Appositional compounding means that the two morphemes belong to the same semantic category, but have opposite meanings. For example, the word “parent” in Northern Qiandong is *maŋ¹³pa³⁵*, which comprises “mother” and “father”. Since all the parts contribute to the semantic meanings, all the morphemes in the words

are considered to be salient. Therefore, all the parts are highlighted.

5.3.1.2 Subordinate Type of Compound Words

The semantic category is determined by the head morpheme in the subordinate compound words. If the research purpose is only to identify the types of compound words suggested by Kratochvíl (1970), it is extremely straightforward. Users can design the morpheme tags as *head*, *attribute*, *referent*, *modifier*, and *measure*, always highlighting the head morphemes. These tags correspond to the categories proposed by Kratochvíl (ibid.), namely *attribute-head*, *head-referent*, *head-modifier*, and *head-measure*.

Our research purpose is to derive the binary vectors from the partial cognates and pipe them to the Bayesian phylogeny algorithms. However, the definition of salience is still abstract at this point. Highlighting the salient part in the subordinate compound words is non-trivial. Thus, our annotation focuses on the morphemes' semantic meanings. The salient morphemes are highlighted in a case-by-case manner. Currently, we have two situations in which the salient morphemes can be easily determined. First, the compounding type is in the form of *root + suffix* or *prefix + stem*. The root and the stem are the salient morphemes; because the majority of prefixes in Hmong-Mien languages are only functional, their semantic meanings are vague. We can determine the root or the stem of the main morphemes that determine the semantic meanings. Second, in the *morpheme + loan morpheme* type, Hmong-Mien speakers have had long-term co-habitation with other language speakers. Therefore, many Hmong-Mien words are combinations of Hmong-Mien words and loanwords. Since the language phylogeny only considers the genealogical relationship, loanwords or loan morphemes should not be included in the data set.

The reason for focusing on the annotation of morphemes' semantic meanings is that the same compound word in different Hmong-Mien languages may be expressed via different orders. For example, the word “cow” in Chuanqiandian variety is written as $na^{31}no^{31}$, in which the morpheme used to represent “bovine” is the no^{31} . However, the same word in Zao Min variety is written as $\eta\eta^{53}pja^{53}$, in which the $\eta\eta^{53}$ is the morpheme representing “bovine”. If users decide to extract only the bovine part, the annotation can provide a keyword for rapid retrieval.

5.3.1.3 Reduplicated Types of Compound Words

Reduplicated words are the easiest cases to address. Even though the words have the purpose of emphasizing something or working as an English comparative, all of the morphemes have the same meaning. Therefore, we only highlight one of the morphemes to represent the word.

5.3.2 Considering the Semantic Shift

Chapter 4 mentioned that the mechanisms of semantic shift results in missing cognates in the sampled languages. We can also find some examples in Sagart et al. (2019). Some cognate decisions in Sagart et al. (ibid.) were found to be forcefully fitted into the cognate coding because they (and subsequently we in the third project), were working under the underlying assumption

that a *semantic shift did not exist in the core vocabulary*. For example, Baxter and Sagart (2014, p. 101) reconstructed the Old Chinese word 舐 “lick” as **Cə.leʔ* and the word 食 “eat” as **mə-lək*. The study annotated the Lushai word *lick liak* and the Cantonese word *lick lai* as cognates. However, putting the three words **Cə.leʔ*, **mə-lək*, and *liak* together, we can see that *liak* and **mə-lək*, rather than *liak* and **Cə.leʔ*, are cognates (Lai, 2021).

In this project, a more appropriate treatment would have been to leave the cognate coding of the Lushai word “lick” *liak* blank, indicating that the cognates within the definition of the Concepticon concept “lick”⁶ could not be found.

The assumption that *no semantic shift occurred in the core vocabulary* was unrealistic, but there was no other way to bypass this assumption when the team was working on this large-scale phylolinguistic project. The limitations stemmed from the computational methodology, including automatic data merging, cognate decisions, and the underlying data matrix for a Bayesian analysis. We also discussed the possibility of employing the cognate coding method, which allows linguists to annotate cognates cross-semantically, in Chapter 4. In this section, we discuss the obstacles in more detail.

Linking various data sets using Concepticon concepts requires scholars to agree on the definition of each concept. For example, a Concepticon concept “sky” defines the concept as “The part of the earth’s atmosphere and space outside it that is visible from earth’s surface. During the day it is perceived as blue, and at night as black” (List et al., 2020a). If one links the word “sky” in their data set to the Concepticon concept “sky”, the user automatically agrees that the word matches the definition. Using this agreement, we are able to link the words belongs to “sky” in various languages and, potentially, to numerous data sets. Therefore, linking the data sets via the computational method and the Concepticon concept, means that we would not be able to consider the factor of semantic shift—for example, if a word “above” in a language represents both “sky” and “above”. Assuming that this word is a cognate with “sky” in other languages but is annotated with the Concepticon concept “above”, this word will not be retrieved if the user only requests merging the “sky” words.

Another limitation is based on the cognate judgments in large-scale, cross-linguistic data. In Chapter 2 and 3, we explained the usefulness of *Edictor* in determining word or morpheme cognacy among several language varieties. Compared to the old method, which used Excel, *Edictor* provides much more assistance when making cognate annotations. However, the feature of displaying cross-semantic cognates was only incorporated after 2020 (Wu et al., 2020). The previous version only allowed users to annotate the cognate sets within the same semantic category. Therefore, annotating cross-semantic cognates manually via *Edictor* was not possible.

The last issue pertaining to cross-semantic cognates is the challenge of data transformation. We elaborate on this particular issue in Section 5.4.

⁶Concepticon id: 319, definition: to stroke with the tongue.

Linguists may argue that a model that assumes no semantic shift is not ideal. However, these assumptions provide a baseline. Future phylolinguistic analyses can employ more complex models, and can compare the differences between two different scenarios.

5.4 Transformation

Transformation means extracting information from the annotated data sets and forming new formats for follow-up applications. In our case, the goal is to transform the cognate sets into a matrix on which Bayesian phylogenetic algorithms can operate. The matrix usually organizes the binary sequences according to whether the cognate is present or absent in a given language. The orders of the cognate sets imply that the semantic categories are also taken into account. Since language changes do not always follow a constant rate, neither does the word form. One of the advantages of the Bayesian phylogenetic algorithm is that we can incorporate different evolutionary rates for each of the concepts into our presumption in the model.

Table 5.14 shows the structure of the binary matrix. We identified two critical points that needed to be considered during the transformation stage by observing the data structure.

Taxa	Concept 1			Concept 2		
	C1	C2	C3	C4	C5	C6
Taxon 1	0	1	0	1	0	0
Taxon 2	1	0	0	0	1	0
Taxon 3	1	0	0	0	0	1

Table 5.12: A representation of the binary matrix.

5.4.1 Partial Cognates

Each word is assigned one cognate set in the binary matrix shown in Table 5.14. The question is how partial cognates are included in this type of format.

The binary vector transforms each slot as the presence or absence of the cognate. This means that the transformation only considers one of the evolutionary events, namely the morphological changes. Assuming there is a finite set of morphemes that can be used to express a particular concept in a proto-language, the daughter languages will express the same concept by changing the morphemes' forms or by using a subset of the morphemes. However, other evolutionary factors, such as loanwords, semantic shifts, or morpheme replacement, have not been considered.

The following is an example from the Sinitic languages, assuming that the concept “wife” is expressed as 內人 *nèi rén*, 妻子 *qī zǐ*, and 娘子 *niáng zǐ* in three different languages (see Table 5.13). We assign this concept a total of five morphemes: 內 *nèi* “inside”, 人 *rén* “person”, 妻 *qī* “wife”, 娘 *niáng* “female or mother”, and 子 *zǐ* “a suffix to indicate a person”. The underlying assumption shows that all these morphemes existed in the proto-language under the “wife” concept. All these morphemes have the same opportunity to appear in the daughter languages, and

vice versa. However, this may not be true. Many factors can contribute to the result that Taxon 1 has a different set of cognates from the other languages, such as semantic shifts, analogy, and loanwords.

Taxa	Concept 1					Concept 2		
	妻	娘	子	內	人	C6	C7	C8
Taxon 1	0	0	0	1	1	1	0	0
Taxon 2	1	0	1	0	0	0	1	0
Taxon 3	0	1	1	0	0	0	0	1

Table 5.13: A representation of the partial cognates.

As a result, we introduced the annotation method and four different transformation methods to convert the partial cognate sets into word cognate sets. The figures below demonstrate the Hmong-Mien phylogeny generated from the four different transformations in contrast to the result generated from the partial cognates.

5.4.2 Cross-Semantic Cognates

Working with cross-semantic cognates would mean that we grouped morphemes that appeared in different concepts. The algorithm for detecting cross-semantic cognates groups morphemes according to the morphemes' forms (Wu et al., 2020). The purpose is to provide a quick summary to assist linguists to infer cross-semantic cognates in an efficient way.

It has been suggested that, instead of ordering cognates according to concepts, one could order the cognates according to the morphemes. Take three Chinese concepts as an example: 魚子 *yú zǐ* “fish egg”, 妻子 *qī zǐ* “wife”, and 子嗣 *zǐ sì* “descendants”. These words can be divided into the morphemes 魚 *yú* “fish”, 妻 *qī* “wife”, 嗣 *sì* “heirs”, and 子 *zǐ* “egg, suffix, or descendants”. The 子 *zǐ* will be detected as the cross-semantic cognate in this example. The cognate sets can be displayed as follows:

Taxa	魚		妻		嗣		子	
	C1	C2	C3	C4	C5	C6	C7	C8
Taxon 1	0	1	1	0	0	1	1	0
Taxon 2	1	0	1	0	0	1	0	1
Taxon 3	0	1	0	1	1	0	1	0

Table 5.14: A representation of the cross-semantic cognate sets.

The 子 *zǐ* originally meant descendants. The word has differentiated into multiple meanings, such as son, egg, and a suffix to refer to someone. Moreover, these compound words may not appear in the languages at the same time. Therefore, transforming the cognate sets in such a way only considers the methodological point of view instead of starting from a linguistic perspective.

5.5 Application

Cognate sets can be applied in various analyses, but our dissertation only focuses on deriving the language phylogeny from the cognate sets. The definition of the distance-based matrix and the character-based matrix has been explained in previous chapters. Here, we would like to introduce the applications that are available to reconstruct phylogenies from the two different type of matrices.

5.5.1 Distance-based Study

Inferring and visualizing phylogeny from distance-based matrices can be achieved via phylogenetics software or packages. Users have a wide selection of software and packages to reconstruct phylogenies from distance-based matrices. Throughout the three projects, we have used or tested phylogenetics software that is provided by other research groups, such as *SplitsTree* (Huson, 1998; Huson and Bryant, 2005), *Figtree* (Rambaut, 2010), or *iTOL* (v5) (Letunic and Bork, 2021). We also used Python and R programming languages to reconstruct and annotate phylogenies from distance-based matrices. Users should select a tool based on the complexity of the tasks.

Desktop or web applications are a fast solution for generating or inspecting a tree. The downside of using these applications is that users have limited options regarding the phylogenetic algorithms used to infer a tree. Phylogenetic software is typically a research output, and is not constantly maintained or updated if the projects are no longer funded. Newer algorithms are usually not available, and are often not compatible with newer operating systems. The software is also very strict in terms of input formats. Users need to transform their data sets into *Phylip* or *Nexus* before using the software. The software will report errors without any further indication of whether the format contains a typo, such as a space or a tab. A lack of informative feedback usually results in people wasting time debugging.

We recommend that users should write their own programming scripts for complex tasks, despite the fact that writing a script from scratch is often time consuming. Programming languages with open-source packages provide users with various algorithms to infer phylogenies from distance-based matrices. Users have a wider selection of phylogenetic algorithms. One can always find a package, even if the algorithm is newly proposed. For example, *LingPy*, a Python library developed for handling lexical data, provides several methods for computing distances among languages. The Python package also provides two options that allow people to reconstruct NJ or UPGMA trees. Users can also overlay several mathematical models; for example, an NJ tree with a bootstrapping method. The programming scripts can also report errors in an informative way when the script does not run as expected. Replicability is another advantage of using programming languages for phylogenetic inference. If a study provides the programming script and data set, it is easy for others to obtain the same results. A research result that can be replicated is more trustworthy. The only downside of generating trees via programming languages is that

users need to familiarize themselves with the programming language in order to work efficiently.

A module in our pipeline in Chapter 3 was written in Python language, and made use of the aforementioned Python libraries for inferring and visualizing the phylogeny. This module can be taken as a Python example if people who are not familiar with Python language want to adopt this approach.

5.5.2 Character-based Study

For the purpose of our experiments, writing a script to generate a Bayesian phylogeny was not feasible. A Bayesian model consists of several components, and each component is established based on a solid statistical background, which we lacked. Because the focus of this dissertation was on improving and applying the existing methods, we relied on the currently available Bayesian phylogenetic software (Bouckaert et al., 2019; Ronquist et al., 2012) to infer Bayesian phylogenetic trees. We used *MrBayes* (Ronquist et al., 2012) to infer the Sinitic phylogeny in the second project (Wu and List, 2022) and *Beast 2* (version 2.5) (Bouckaert et al., 2019) to infer the phylogeny (Wu et al., 2022).

MrBayes is a light-weight Bayesian phylogenetic software package. Users can generate a *Nexus* and set up the models in the *MrBayes*' terminal. Alternatively, users can set up a block in the *Nexus* file to list all the parameters for the model. It takes a very short time to prepare the input file because a *Nexus* file can be generated directly by *LingPy*. Users only need to modify the prior models. The input file format is much easier for users to prepare and understand. We used *MrBayes* because we could generate different input formats more rapidly. It is a powerful tool when the purpose of the experiment is to compare the topology of cognate sets that have been derived from the four conversion methods based on the same statistical model. However, *MrBayes* integrates fewer and older Bayesian phylogenetic models than does *Beast 2*. It may not be the best tool for users who would like to try the most recently developed Bayesian models.

Beast 2 is the most popular Bayesian phylogenetic analysis software in the domain of historical linguistics. The software incorporates a wider selection of Bayesian models than does *MrBayes*, including fossilized birth-death and skyline birth-death models, the two most frequently used models in the field of historical linguistics. The software adopts an Extensible Markup Language (XML), which contains all the parameters of a model, and a character-based matrix. Obtaining such an input framework in our framework is not straightforward because *LingPy* cannot directly produce an XML file. Therefore, we generated a *Nexus* using *LingPy*, and then used *BEAUti* (Drummond et al., 2012), desktop software with a user-friendly interface, to establish the models and to generate the XML file.

Software for Bayesian phylogenetic analysis usually produces several large files containing thousands of sampled trees and numerous lines of running logs. Navigating all these numbers is not an easy task; hence, the software is usually accompanied by programs to visualize all the data. *Tracer* (Rambaut et al., 2018), *DensiTree* (Bouckaert, 2010), and *TreeAnnotator* (Drummond and

Rambaut, 2007) are programs that are provided by the same research team to inspect the progress, assess the outcome, and generate a consensus tree.

5.6 Interpretation

The term “interpretation” refers not only to interpreting the results of Bayesian phylogeny using knowledge from various disciplines, but also to selecting the appropriate calibration dates for the languages and reports; that is, the log output from the Bayesian phylogenetic software.

5.6.1 The Reports on Bayesian Phylogenetic Analyses

A Bayesian phylogenetic analysis compares the prior and the posterior throughout the entire process. The comparisons are all written in a log file. It is impossible to inspect the quality of the inference process if we are only scanning the entries. Fortunately, *Tracer* is a powerful tool for examining whether the algorithms reach a high posterior stage (that is, converge) or how effective the parameters we assign to the models are. This tool assisted us to modify the previous settings for each run.

At first, it was extremely difficult to understand the figures presented by *Tracer* because the figures are lines in different colors that run up and down from one side of the figure to the other. We sought the help of mathematicians to provide some general principles for interpreting the patterns and adjusting our previous settings.

5.6.2 Selecting Calibration Dates

Another challenge regarding interpretation that we encountered was selecting the calibration dates for inferring Bayesian phylogeny. The best material for dating a Bayesian phylogeny is the dates in written records. However, the languages were mainly spoken languages in the Himalayan mountain region. No inscriptions were available to date the birth of the Kho-Bwa or Hrusish languages. Surveying evidence from other disciplines, such as archaeogenetics or environmental archaeology, is an alternative method. However, the information from the other disciplines for calibrating the Bayesian phylogeny was extremely limited. As mentioned in Wu et al. (2022), we were unable to use the calibration dates from the archaeological discoveries because we could not confirm that the archaeological findings were associated with the ancestor of Kho-Bwa, Hrusish, or Tani speakers. We used the calibrations dates from Sagart et al. (2019) in another branch, and used a Bayesian phylogenetic analysis in two different stages to bypass these difficulties.

In principle, we recommend that scholars should treat written records as the first priority, materials from other disciplines as the second priority, and use some other approaches to circumvent the limitation of missing calibration dates. Moreover, scholars should use resources that are reliable and not forcibly calibrate the internal nodes using the archaeological findings.

5.6.3 Bayesian Phylogeny

Linguists infer the speakers' ancestors' lifestyle by reconstructing the proto-language and then inferring the lifestyle from the reconstructed words. Linguists have reconstructed agricultural words (for example, “rice” **mblau_A*) in the proto-Hmong-Mien language, and inferred that the common ancestor of the Hmongic and Mienic people lived in a region in which rice farming was practiced. Because rice plants were cultivated in the Yangtze River region in prehistoric times, some linguists believe that the proto-Hmong-Mien speakers lived there (Ratliff, 2010). Some linguists may overly interpret the genetic analysis and forcibly link the research outcome with their linguistic findings (Robbeets et al., 2021; van Driem, 2001). We consider this type of inference to be similar to cherry-picking the archaeological or genetic findings that suit their narratives. In one of the unpublished versions of Wu et al. (2022), we overly interpreted our research results. It was only through expanding the collaborations with linguists and consulting with archaeologists that we were able to review our work and modify the manuscript in a more humble manner.

Cross-disciplinary frameworks for researching the association between human and language family expansion are underdeveloped. It takes time to establish a solid framework by combining data or research findings from different disciplines in one study. Despite the fact that many published studies have used the language and gene co-evolution framework since it was first proposed (Cavalli-Sforza et al., 1992), the concept is immaturely developed. Take Cavalli-Sforza et al. (ibid.) as an example. The study compares populations living in South China who speak Sino-Tibetan languages. This study incorrectly sampled populations according to geography instead of according to language, because the population in South China speaks Sino-Tibetan, Hmong-Mien and Tai-Kadai languages. Another issue is that studies such as this lack common definitions that are shared across various disciplines. Languages, cultures, and populations all form their own hierarchical structures. The Sino-Tibetan language family is a large language family with Tibeto-Burman and Sinitic as the two major divisions under which multiple subgroups can be identified. Population studies often use nationalities, cities, geographical regions or certain cultural labels. Accordingly, it is clear that ways of forming the hierarchical structures in different disciplines differ significantly. Another notable point is population migration, population shifts, new populations integrating with local groups, and many other factors that could trigger language or cultural changes or shifts. There are more factors to consider than is possible when imperiously marking language changes or cultural features with labels such as “present” and “absent”. Finally, a quantitative method for measuring the differences between two or among three Bayesian trees is still lacking.

A valuable lesson we learned through working on the project with Wu et al. (2022) is that we need to be careful when selecting the calibration dates for dating Bayesian phylogeny. In addition, we should interpret different threads of evidence or data sets as carefully as possible. Lastly, including perspectives from different research fields is highly recommended. The best scenario entails communicating with experts in different fields of study.

5.7 Conclusion

This dissertation discussed the adequacy of the classical comparative method for analyzing MSEA languages. We developed a computer-expert hybrid approach to improve the classical comparative method to produce a methodology that is better suited to MSEA languages. Moreover, the works in the dissertation demonstrate the significance of open data. Finally, we addressed the importance of combining multiple forms of evidence to reconstruct human prehistory.

Bibliography

- Bandelt, H.-J. and A. W. M. Dress (1992). “Split decomposition: A new and useful approach to phylogenetic analysis of distance data.” In: *Molecular Phylogenetics and Evolution* 1.3, pp. 242–252. doi: 10.1016/1055-7903(92)90021-8.
- Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Reprint in 2010. Berlin and New York: De Gruyter Mouton. doi: 10.1515/9783110857085.
- Baxter, W. H. and L. Sagart (2014). *Old Chinese: A new reconstruction*. Oxford University Press. doi: 10.1093/acprof:oso/9780199945375.001.0001.
- Bellwood, P. S. (2005). *First Farmers: The origins of agricultural societies*. Malden, MA: Wiley.
- Blench, R. and M. W. Post (2013). “Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages.” In: *Trans-Himalayan Linguistics: Historical and Descriptive Linguistics of the Himalayan Area*. Ed. by T. Owen-Smith and N. Hill. Berlin and New York: De Gruyter Mouton, pp. 71–104. doi: 10.1515/9783110310832.71.
- Bloomfield, L. (1933). *Language*. University of Chicago. London: George Allen & Unwin.
- Bodt, T. A. (2020). *Duhumbi dictionary*. Arnhem, the Netherlands: Monpasang Publications.
- Bouckaert, R. et al. (2019). “BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis.” In: *PLOS Computational Biology* 15.4, e1006650. doi: 10.1371/journal.pcbi.1006650.
- Bouckaert, R. R. (2010). “DensiTree: making sense of sets of phylogenetic trees.” In: *Bioinformatics* 26.10, pp. 1372–1373. doi: 10.1093/bioinformatics/btq110.
- Brass, T. (2012). “Scott’s “Zomia”, or a populist post-modern history of nowhere.” In: *Journal of Contemporary Asia* 42.1, pp. 123–133. doi: 10.1080/00472336.2012.634646.
- Brunelle, M. and J. Kirby (2016). “Tone and phonation in southeast Asian languages.” In: *Language and Linguistics Compass* 10.4, pp. 191–207. doi: 10.1111/lnc3.12182.
- Bryant, D. and V. Moulton (2004). “Neighbor-Net: An agglomerative method for the construction of phylogenetic networks.” In: *Molecular Biology and Evolution* 21.2, pp. 255–265. doi: 10.1093/molbev/msh018.
- Burling, R. (1961). *A Garo grammar*. Vol. 25. Deccan College Monograph Series. Deccan College Postgraduate and Research Institute.
- Bußmann, H. and H. Bußmann (2006). *Routledge dictionary of language and linguistics*. Ed. by G. Trauth. Transferred to digital print. Routledge linguistics / reference. London and New York: Routledge.
- Campbell, L. (1999). *Historical linguistics: An introduction*. First MIT Press. Cambridge, Massachusetts: The MIT Press.
- Campbell, L. (2013). *Historical linguistics: An introduction*. 3rd edition. Edinburgh: Edinburgh University Press.

- Catford, J. C. (1977). *Fundamental problems in phonetics*. Vol. 55. Edinburgh University Press.
doi: 10.2307/412751.
- Cavalli-Sforza, L. L., E Minch, and J. L. Mountain (1992). "Coevolution of genes and languages revisited." In: *Proceedings of the National Academy of Sciences* 89.12, pp. 5620–5624. doi: 10.1073/pnas.89.12.5620.
- Chechuro, I., M. Daniel, and S. Verhees (2021). "Small-scale multilingualism through the prism of lexical borrowing." In: *International Journal of Bilingualism* 25.4, pp. 1019–1039. doi: 10.1177/13670069211023141.
- Chen, B. (1996). *Lùn yǔ yán jiē chù yǔ yǔ yán lián méng: Hàn yuè (dòng tái) yǔ yuán guān xì de jiě shì* 《論語言接觸與語言聯盟: 漢越 (侗台) 語源關係的解釋》 [*Language contact and language union: Explaining the linguistic relationship between Chinese and Yue (Kam-Tai)*]. 1st ed. Beijing: Yǔ Wén Chū Bǎn Shè 語文出版社.
- Chen, Q. (1984). "The status of the She language in the Miao-Yao stock language 〈畬語在苗瑤語族中的地位〉." In: *Studies in Language and Linguistics* 《語言研究》1, pp. 200–214.
- Chen, Q. (2002). "Hàn yǔ miáo yáo yǔ bǐ jiào yán jiū 〈漢語苗瑤語比較研究〉 [The comparative study of Sinitic and Miao-Yao dialects]." In: *Hàn cáng yǔ tóng yuán cí yán jiū* (2) 《漢藏語同源詞研究 (二)》 [*Research of Sino-Tibetan cognates*]. Ed. by b. Dīng. 1st ed. Guǎng xī mín zú chū bǎn shè 廣西民族出版社.
- Chen, Q. (2012). *Miáo yáo yǔ wén* 《苗瑤語文》 [*Mao and Yao language*]. 1st ed. Beijing: Zhōng yāng mín zú dà xué chū bǎn shè 中央民族大学.
- Clark, E. R. (2008). "A phonological analysis and comparison of two Kim Mun varieties in Laos and Vietnam." MA thesis. Chiang Mai, Thailand: Payap University.
- Cui, L., F. Cong, J. Wang, W. Zhang, Y. Zheng, and J. Hyönä (2018). "Effects of grammatical structure of compound words on word recognition in Chinese." In: *Frontiers in Psychology* 9, p. 258. doi: 10.3389/fpsyg.2018.00258.
- DeLancey, S. (2013). "Creolization in the divergence of the Tibeto-Burman languages." In: *Trans-Himalayan Linguistics: Historical and Descriptive Linguistics of the Himalayan Area*. Ed. by T. Owen-Smith and N. Hill. Berlin and New York: De Gruyter Mouton, pp. 41–70. doi: 10.1515/9783110310832.41.
- Deng, X. and W. S.-Y. Wang (2003). "A quantitative study on the genetic relationship of Miao-Yao languages: The lexicostatics approach." In: *Chinese Language* 3.
- Diamond, J. (2003). "Farmers and their languages: The first expansions." In: *Science* 300.5619, pp. 597–603. doi: 10.1126/science.1078208.
- Diller, A. V. N., J. A. Edmondson, and Y. Luo (2008). *The Tai-Kadai languages*. London and New York: Routledge.
- Dolgopolsky, A. B. (1964). "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točki zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia]." In: *Voprosy Jazykoznanija* 2, pp. 53–63.

- Dolgopolsky, A. B. (1986). "A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia." In: *Typology, Relationship and Time. A collection of papers on language change and relationship by Soviet linguists*. Originally published in 1964 as "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točki zrenija" and translated from the Russian by V. V. Shevoroshkin. Karoma Publisher, pp. 27–50.
- Drummond, A. J. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." In: *BMC Evolutionary Biology* 7.1, p. 214. doi: 10.1186/1471-2148-7-214.
- Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut (2012). "Bayesian Phylogenetics with BEAUti and the BEAST 1.7." In: *Molecular Biology and Evolution* 29.8, pp. 1969–1973. doi: 10.1093/molbev/mss075.
- Dryer, M. S. and M. Matthew S., eds. (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Durie, M. and M. Ross, eds. (1996). *The comparative method reviewed: Regularity and irregularity in language change*. New York: Oxford University Press.
- Enfield, N. J. and B. Comrie (2015). *Mainland southeast Asian languages*. Berlin and New York: De Gruyter Mouton.
- Enfield, N. J. (2011). "Dynamics of human diversity in mainland southeast Asia: Introduction." In: *Dynamics of human diversity: The case of mainland southeast Asia*. Ed. by N. J. Enfield. Canberra: Pacific Linguistics, pp. 1–8.
- Fitch, W. M. (1997). "Networks and viral evolution." In: *Journal of Molecular Evolution* 44.1, S65–S75. doi: 10.1007/PL00000059.
- Forkel, R. and J.-M. List (2020). "CLDFBench: Give your cross-linguistic data a lift." In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. Luxembourg: European Language Resources Association (ELRA), 6997-7004.
- Forkel, R., J.-M. List, S. J. Greenhill, C. Rzymiski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. A. Kaiping, and R. D. Gray (2018). "Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics." In: *Scientific Data* 5.1, p. 180205. doi: 10.1038/sdata.2018.205.
- Geisler, H. and J.-M. List (2010). "Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics." In: *Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik*. Ed. by H. Hettrich. Document has been submitted in 2010 and is still waiting for publication. Wiesbaden: Reichert.
- Gong, H.-C. (2006). "The Sino-Miao-Yao relationship revisited." In: *Bulletin of Chinese Linguistics* 1.1, pp. 225–270. doi: 10.1163/2405478X-90000013.
- Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). "Language phylogenies reveal expansion pulses and pauses in Pacific settlement." In: *Science* 323.5913, pp. 479–483. doi: 10.1126/science.1166858.

- Gray, R. D., D. Bryant, and S. J. Greenhill (2010). "On the shape and fabric of human history." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1559, pp. 3923–3933. doi: 10.1098/rstb.2010.0162.
- Greenhill, S. J., P. Heggarty, and R. D. Gray (2020). "Bayesian phylolinguistics." In: *The Handbook of Historical Linguistics*. Ed. by R. D. Janda, B. D. Joseph, and B. S. Vance. West Sussex: John Wiley & Sons, Ltd. Chap. 11, pp. 226–253. doi: 10.1002/9781118732168.ch11.
- Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel (2015). "Bantu expansion shows that habitat alters the route and pace of human dispersals." In: *Proceedings of the National Academy of Sciences* 112.43, pp. 13296–13301. doi: 10.1073/pnas.1503793112. eprint: <https://www.pnas.org/content/112/43/13296.full.pdf>.
- Hammarström, H., R. Forkel, M. Haspelmath, and S. Bank (2020). *Glottolog database. Version 4.3*. Leipzig: Zenodo. doi: 10.5281/ZENODO.4061162.
- Heimbach, E. (1969). *White Hmong-English dictionary*. Ithaca: Southeast Asia Program Cornell University.
- Hill, N. W. and J.-M. List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages." In: *Yearbook of the Poznan Linguistic Meeting* 3.1, pp. 47–76. doi: 10.1515/yp1m-2017-0003.
- Holm, H. J. (2007). "The new arboretum of Indo-European "trees". Can new algorithms reveal the phylogeny and even prehistory of Indo-European?" In: *Journal of Quantitative Linguistics* 14.2-3, pp. 167–214. doi: 10.1080/09296170701378916.
- Holman, E. W., S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, and D. Bakker (2008). "Explorations in automated language classification." In: 42.3-4, pp. 331–354. doi: 10.1515/FLIN.2008.331.
- Hsiu, A. C. (2015). *The classification of Na Meo, a Hmong-Mien language of Vietnam*. Bangkok. doi: 10.5281/zenodo.1127804.
- Huson, D. H. (1998). "SplitsTree: Analyzing and visualizing evolutionary data." In: *Bioinformatics* 14.1, pp. 68–73. doi: 10.1093/bioinformatics/14.1.68.
- Huson, D. H. and D. Bryant (2005). "Application of Phylogenetic Networks in Evolutionary Studies." In: *Molecular Biology and Evolution* 23.2, pp. 254–267. doi: 10.1093/molbev/msj030. eprint: <https://academic.oup.com/mbe/article-pdf/23/2/254/3894375/msj030.pdf>.
- Huáng, B. and Q. Dài, eds. (1992). *Zàngmiǎn yǔzú yǔyán cíhuì* 《藏緬語族語言詞匯》 [A Tibeto-Burman Lexicon]. Běijīng: Zhōngyāng Mínzú Dàxué [中央民族大學].
- Ji, X., K. Kuman, R. J. Clarke, H. Forestier, Y. Li, J. Ma, K. Qiu, H. Li, and Y. Wu (2016). "The oldest Hoabinhian technocomplex in Asia (43.5 ka) at Xiaodong rockshelter, Yunnan Province, southwest China." In: *Quaternary International*. Peking Man and related studies 400, pp. 166–174. doi: 10.1016/j.quaint.2015.09.080.
- Jäger, G. (2018). "Global-scale phylogenetic linguistic inference from lexical resources." In: *Scientific Data* 5.1, p. 180189. doi: 10.1038/sdata.2018.189.

- Jäger, G. (2019). "Computational historical linguistics." In: *Theoretical Linguistics* 45.3-4, pp. 151–182. doi: 10.1515/tl-2019-0011.
- Klaproth, J. von (1823). *Asia Polyglotta*. A. Schubart.
- Kratochvíl, P. (1970). *The Chinese language today: features of an emerging standard*. London: Hutchinson.
- Ladefoged, P. (1982). *A course in phonetics*. Los Angeles: University of California.
- Lai, Y. (2021). *private communication*. private communication.
- Letunic, I. and P. Bork (2021). "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation." In: *Nucleic Acids Research* 49.W1, W293–W296. doi: 10.1093/nar/gkab301. eprint: <https://academic.oup.com/nar/article-pdf/49/W1/W293/38842284/gkab301.pdf>.
- Leyden, J. (1808). "On the languages and literature of the Indo-Chinese nations." In: *Asiatic Researches* 10, pp. 158–289.
- Li, F. (1937). "Languages and dialects of China." In: *The People's Republic of China Yearbook* 1.1.
- Li, F. (1973). "Languages and dialects of China." In: *Journal of Chinese linguistics* 1.1.
- Lieberman, V. (2010). "A zone of refuge in Southeast Asia? Reconceptualizing interior spaces." In: *Journal of Global History* 5.2, pp. 333–346. doi: 10.1017/S1740022810000112.
- List, J.-M. (2012a). "LexStat: Automatic detection of cognates in multilingual wordlists." In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon, France: Association for Computational Linguistics, pp. 117–125.
- List, J.-M. (2012b). "SCA: Phonetic alignment based on sound classes." In: *New Directions in Logic, Language and Computation*. Ed. by D. Hutchison et al. Vol. 7415. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 32–51. doi: 10.1007/978-3-642-31467-4_3.
- List, J.-M. (2016). "Computer-assisted language comparison: Reconciling computational and classical approaches in historical linguistics." In: *Max Planck Institute for the Science of Human History*. doi: 10.5281/ZENODO.842734.
- List, J.-M. (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Valencia: Association for Computational Linguistics, pp. 9–12.
- List, J.-M. (2019). "Automatic inference of sound correspondence patterns across multiple languages." In: *Computational Linguistics* 45.1, pp. 137–161. doi: 10.1162/coli_a_00344.
- List, J.-M., C. Anderson, T. Tresoldi, and R. Forkel (2021a). *CLTS. Cross-linguistic transcription systems*. Leipzig: Max Planck Institute for Evolutionary Anthropology. doi: 10.5281/ZENODO.4705149.
- List, J.-M., M. Cysouw, and R. Forkel (2016). "Concepticon: A Resource for the Linking of Concept Lists." In: *Proceedings of the Tenth International Conference on Language Resources*

- and Evaluation (LREC'16. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2393–2400.
- List, J.-M., S. J. Greenhill, and R. D. Gray (2017). “The potential of automatic word comparison for historical linguistics.” In: *PLOS ONE* 12.1, e0170046. doi: 10.1371/journal.pone.0170046.
- List, J.-M., S. J. Greenhill, T. Tresoldi, and R. Forkel (2019). *LingPy. A python library for quantitative tasks in historical linguistics*. Version 2.6.9. Leipzig. doi: 10.5281/ZENODO.3554103.
- List, J.-M., C. Rzymiski, S. J. Greenhill, N. Schweikhard, K. Panykh, A. Tjuka, A. Tjuka, M.-S. Wu, and F. R. Concepticon (2020a). *Concepticon. A resource for the linking of concept lists. Version 2.3.0*. Jena: Max Planck Institute for the Science of Human History.
- List, J.-M., C. Rzymiski, S. J. Greenhill, N. E. Schweikhard, K. Panykh, A. Tjuka, M.-S. Wu, C. Hundt, T. Tresoldi, and R. Forkel (2020b). *Concepticon. A resource for the linking of concept lists. Version 2.4.0*. Jena: Max Planck Institute for the Science of Human History. doi: 10.5281/ZENODO.4162002.
- List, J.-M., C. Rzymiski, S. Greenhill, N. E. Schweikhard, K. Panykh, A. Tjuka, C. Hundt, and R. Forkel (2021b). *Concepticon. A resource for the linking of concept lists. Version 2.5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. doi: 10.5281/zenodo.4911605.
- List, J.-M., M. Walworth, S. J. Greenhill, T. Tresoldi, and R. Forkel (2018). “Sequence comparison in computational historical linguistics.” In: *Journal of Language Evolution* 3.2, pp. 130–144. doi: 10.1093/jole/lzy006.
- Maddieson, I. (2009). *Calculating phonological complexity*. Berlin and New York: De Gruyter Mouton.
- Maddieson, I. (2013). “Tone.” In: *The World Atlas of Language Structures Online*. Ed. by M. S. Dryer and M. Haspelmath. Max Planck Institute for Evolutionary Anthropology.
- Máo, Z. (2004). *Yáozú miǎnyǔ fāngyán yánjiù 《瑶族勉语方言研究》 [Research on the Mien dialect of the Yao people]*. Di 1 ban. Běijīng: Mínzú Chūbǎnshè 民族出版社.
- Matisoff, J. (1973). “Tonogenesis in southeast Asia.” In: *Southern California Occasional Papers in Linguistics* Consonant types and tone.1, pp. 71–96.
- Matisoff, J. A. (2003). *Handbook of Proto-Tibeto-Burman: System and philosophy of Sino-Tibetan reconstruction*. Ed. by J. A. Matisoff. University Presses of California, Columbia and Princeton.
- Matisoff, J. A. (2015). *The Sino-Tibetan etymological dictionary and thesaurus project (STEDT)*. University of California.
- Matsumura, H. et al. (2019). “Cranio-metrics reveal ”two layers” of prehistoric human dispersal in eastern Eurasia.” In: *Scientific Reports* 9.1, p. 1451. doi: 10.1038/s41598-018-35426-z.
- Mei, Z. (1995). *Comparative studies among Wu and Min dialects*. Di 1 ban. Zhongguo dong nan fang yan bi jiao yan jiu cong shu di 1 ji. Shanghai: Shanghai jiao yu chu ban she.

- Michaud, A. (2012). "Monosyllabicization: patterns of evolution in Asian languages." In: *Mono-syllables*. Ed. by T. Stolz, N. Nau, and C. Stroh. Akademie Verlag, pp. 115–130. doi: 10.1524/9783050060354.115.
- Michaud, J. (2010). "Editorial –Zomia and beyond." In: *Journal of Global History* 5.2, pp. 187–214. doi: 10.1017/S1740022810000057.
- Pan, W. (2007). "Discuss the relationship between Miao-Yao and Sino-Tibetan from a few words 〈從幾個詞語討論苗瑤語與漢藏語的關係〉." In: *Studies in Language and Linguistics* 《语言研究》 2.
- Pollock, R. (2006). "The value of public domain." In: *Institute for Public Policy Research*.
- Post, M. W. and R. Burling (2017). "The Tibeto-Burman languages of northeast India." In: *The Sino-Tibetan Languages*. Ed. by R. J. LaPolla and G. Thurgood. London and New York: Routledge, pp. 213–242.
- Przyluski, J. and G. H. Luce (1931). "The number "a hundred" in Sino-Tibetan." In: *Bulletin of the School of Oriental and African Studies* 6.3, pp. 667–668. doi: 10.1017/S0041977X00093150.
- Rambaut, A. (2010). *Figtree v 1.4.4*. Edinburgh: Molecular Evolution, Phylogenetics and Epidemiology.
- Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard (2018). "Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7." In: *Systematic Biology* 67.5, pp. 901–904. doi: 10.1093/sysbio/syy032.
- Ratliff, M. S. (2010). *Hmong-Mien language history*. Studies in language change 8. Canberra: Pacific Linguistics.
- Remsangpuia (2008). *Puroik phonology*. Shillong: Don Bosco Centre for Indigenous Cultures.
- Reyes-Centeno, H., S. Ghirotto, F. D  troit, D. Grimaud-Herv  , G. Barbujani, and K. Harvati (2014). "Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia." In: *Proceedings of the National Academy of Sciences* 111.20, pp. 7248–7253. doi: 10.1073/pnas.1323666111.
- Robbeets, M. et al. (2021). "Triangulation supports agricultural spread of the Transeurasian languages." In: *Nature* 599.7886, pp. 616–621. doi: 10.1038/s41586-021-04108-8.
- Ronquist, F., M. Teslenko, P. Mark, D. Ayres, A. Darling, S. H  hna, B. Larget, L. Liu, M. Suchard, and J. Huelsenbeck (2012). "MrBayes 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space." In: *Systematic Biology* 61, pp. 539–542. doi: 10.1093/sysbio/sys029.
- Rosvall, M. and C. T. Bergstrom (2008). "Maps of random walks on complex networks reveal community structure." In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123. doi: 10.1073/pnas.0706851105.
- Rzyski, C. et al. (2020). "The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies." In: *Scientific Data* 7.1, p. 13. doi: 10.1038/s41597-019-0341-x.

- Rédei, G. P. (2008). "UPGMA (unweighted pair group method with arithmetic means)." In: *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*. Ed. by G. P. Rédei. Dordrecht: Springer, pp. 2068–2068. doi: 10.1007/978-1-4020-6754-9_17806.
- Sagart, L. (2011a). *Classifying Chinese dialects/Sinitic languages on shared innovations*. Paper presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28).
- Sagart, L. (2011b). "The homeland of Sino-Tibetan-Austronesian: Where and when?" In: *Communication on Contemporary Anthropology* 5.1. doi: 10.4236/coca.2011.51021.
- Sagart, L., G. Jacques, Y. Lai, R. J. Ryder, V. Thouzeau, S. J. Greenhill, and J.-M. List (2019). "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." In: *Proceedings of the National Academy of Sciences* 116.21, pp. 10317–10322. doi: 10.1073/pnas.1817972116.
- Saitou, N and M Nei (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees." In: *Molecular Biology and Evolution* 4.4, pp. 406–425. doi: 10.1093/oxfordjournals.molbev.a040454.
- Sánchez-Miret, F. (1998). "Some reflections on the notion of diphthong." In: *Papers and Studies in Contrastive Linguistics* 34, pp. 27–51.
- Schendel, W. van (2002). "Geographies of Knowing, Geographies of Ignorance: Jumping Scale in Southeast Asia." In: *Environment and Planning D: Society and Space* 20.6, pp. 647–668. doi: 10.1068/d16s.
- Scott, J. C. (2009). *The art of not being governed: an anarchist history of upland Southeast Asia*. Yale agrarian studies series. New Haven London: Yale University Press.
- Shafer, R. (1955). "Classification of the Sino-Tibetan languages." In: *Word* 11.1, pp. 94–111. doi: 10.1080/00437956.1955.11659552.
- Sidwell, P. and M. Jenny, eds. (2021). *The languages and linguistics of mainland Southeast Asia: a comprehensive guide*. 1st ed. The world of linguistics 8. Berlin and New York: De Gruyter Mouton.
- Smalley, W. A., C. K. Vang, and G. Y. Yang (1990). *Mother of writing: the origin and development of a Hmong messianic script*. Chicago: Univ. of Chicago Press.
- Strecker, D. (1987). "The Hmong-Mien languages." In: *Linguistics of the Tibeto-Burman Area* 10, pp. 1–11.
- Strecker, D. (2021). "The morphology and semantics of presyllables in Hmong-Mien languages." In: *Linguistics of the Tibeto-Burman Area* 44.1, pp. 55–74. doi: 10.1075/ltba.20007.str.
- Sun, T. (1993). "A Historical-comparative Study of the Tani (Mirish) Branch in Tibeto-Burman." PhD thesis. University of California at Berkeley.
- Swadesh, M. (1955). "Towards greater accuracy in lexicostatistic dating." In: *International Journal of American Linguistics* 21.2, pp. 121–137. doi: 10.1086/464321.
- Swadesh, M. (1964). "K voprosu o povyshenii tochnosti v leksikostatisticheskom datirovanii." In: *Novoe v Lingvistike* 1.

- Thurgood, G. (2003). "A subgrouping of the Sino-Tibetan languages: The interaction between language contact, change, and inheritance." In: *The Sino-Tibetan languages*. Ed. by G. Thurgood and R. J. LaPolla. London and New York: Routledge, pp. 3–21.
- Thurgood, G. and R. J. LaPolla, eds. (2003). *The Sino-Tibetan languages*. Routledge language family series 3. London and New York: Routledge.
- van Driem, G. (2001). *Languages of the Himalayas: An ethnolinguistic handbook of the greater Himalayan region containing an introduction to the symbiotic theory of language*. Handbook of oriental studies. Section two, India, Handbuch der Orientalistik. Indien. Leiden: Brill.
- van Driem, G. (2007). "The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese." In: *Bulletin of Chinese Linguistics* 1.2, pp. 211–270.
- van Driem, G. (2011). *Tibeto-Burman vs. Sino-Tibetan*. Berlin and New York: De Gruyter Mouton.
- van Driem, G. (2014). "Trans-Himalayan." In: *Trans-Himalayan Linguistics*. Ed. by T. Owen-Smith and N. Hill. Berlin and New York: De Gruyter Mouton, pp. 11–40. doi: 10.1515/9783110310832.11.
- van Driem, G. (2015). "Synoptic grammar of the Bumthang language HL Archive 6." In: *Himalayan Linguistics* 0, pp. 1–77. doi: 10.5070/H90025495.
- Wang, F. and Z. Mao (1995). *Miáo Yáo Yǔ Gǔ Yīn Gòu Nǐ* 《苗瑶語古音構擬》 [*Reconstruction of the proto-Hmong-Mien phonology*]. Beijing: China Social Sciences Press 中國社會科學出版社.
- Wichmann, S., E. W. Holman, and C. H. Brown (2016). *The ASJP database*. Max Planck Institute for the Science of Human History.
- Wilkinson, M. D. et al. (2016). "The FAIR guiding principles for scientific data management and stewardship." In: *Scientific Data* 3.1, p. 160018. doi: 10.1038/sdata.2016.18.
- Wu, M.-S., T. A. Bodt, and T. Tresoldi (2022). "Bayesian phylogenetics illuminate shallower relationships among Trans-Himalayan languages in the Tibet-Arunachal area." In: *Linguistics of the Tibeto-Burman Area*. Forthcoming.
- Wu, M.-S. and J.-M. List (2022). "Annotating cognates in phylogenetic studies of south-east Asian languages." In: *Language Dynamics and Change*. forthcoming. doi: 10.17613/3n9j-y345.
- Wu, M.-S. and J.-M. List (2023). "Annotating cognates in phylogenetic studies of south-east Asian languages." In: *Language Dynamics and Change*. doi: 10.1163/22105832-bja10023.
- Wu, M.-S., N. E. Schweikhard, T. A. Bodt, N. W. Hill, and J.-M. List (2020). "Computer-Assisted Language Comparison: State of the Art." In: *Journal of Open Humanities Data* 6.1, p. 2. doi: 10.5334/johd.12.
- Yang, X. (1999). "fāng yán běn zì yán jiū de tàn yì fǎ 〈方言本字研究的探義法〉." In: *Linguistic Essays in Honor of Mei Tsu-Lin: Studies on Chinese Historical Syntax and Morphology*. Paris: Ecole des Hautes Etudes en Sciences Sociales, Centre de Recherches Linguistiques sur l'Asie Orientale, pp. 299–326.

- Yip, M. (2002). *Tone*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. doi: 10.1017/CB09781139164559.
- Yolchuyeva, S., G. Németh, and B. Gyires-Tóth (2019). “Grapheme-to-Phoneme Conversion with Convolutional Neural Networks.” In: *Applied Sciences* 9.6, p. 1143. doi: 10.3390/app9061143.
- Zhang, H., T. Ji, M. Pagel, and R. Mace (2020). “Dated phylogeny suggests early Neolithic origin of Sino-Tibetan languages.” In: *Scientific Reports* 10.1, p. 20792. doi: 10.1038/s41598-020-77404-4.
- Zhang, M., S. Yan, W. Pan, and L. Jin (2019). “Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic.” In: *Nature* 569.7754, pp. 112–115. doi: 10.1038/s41586-019-1153-z.
- Zhang, Z. and F. Hu (2019). “Vowels and diphthongs in the Changde Mandarin Chinese.” In: *The 19th International Congress of Phonetic Sciences*. Melbourne, Australia.

Ehrenwörtliche Erklärung

Name: Wu-UrbaneK

Vorname: Mei-Shin

Adresse: Jena, Deutschland

Ich erkläre hiermit,

- (a) dass mir die geltende Promotionsordnung der Fakultät bekannt ist;
- (b) dass ich die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe;
- (c) dass mich ausschließlich die folgenden Personen bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben: Prof. Dr. Johann-Mattis List und Prof. Dr. Volker Gast,
- (d) dass die Hilfe einer kommerziellen Promotionsvermittlung nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für die Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- (e) dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;
- (f) dass ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe (wenn doch, bitte Ergebnis angeben).

Jena, March 13, 2023

Mei-Shin Wu-UrbaneK

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Ort, Datum

Unterschrift Verfasser/Verfasserin