Abstract

The Tonal Comparative Method: Tai Tone in Historical Perspective

Rikker Dockum

2019

To date, the majority of attention given to sound change in lexical tone has focused on how an atonal language becomes tonal and on early stage tone development, a process known as tonogenesis. Lexical tone here refers to the systematic and obligatory variation of prosodic acoustic cues, primarily pitch height and contour, to encode contrastive lexical meaning. Perhaps the most crucial insight to date in accounting for tonogenesis is that lexically contrastive tone, a suprasegmental feature, is born from segmental origins. What remains less studied and more poorly understood is how tone changes after it is well established in a language or language family. In the centuries following tonogenesis, tones continue to undergo splits, mergers, and random drift, both in their phonetic realization and in the phonemic categories that underlie those surface tones. How to incorporate this knowledge into such historical linguistic tasks as reconstruction, subgrouping, and language classification in a generally applicable fashion has remained elusive.

The idea of reconstructing tone, and the use of tonal evidence for language classification, is not new. However, the predominant conventional wisdom has long been that tone is impenetrable by the traditional Comparative Method. This dissertation presents a new methodological approach to sound change in lexical tone for languages where tone is already firmly established. The Tonal Comparative Method is an extension of the logic of the traditional Comparative Method, and is a method for incorporating tonal evidence into historical analyses in a manner consistent with the first principles of the longstanding Comparative Method.

The Tonal Comparative Method is developed and modeled using data drawn from

hundreds of doculects (Good & Cysouw 2013) of Tai languages, a branch of the Kra-Dai language family. The Tai languages make an ideal testing ground for advancing the theory of sound change in tone systems because they are robustly tonal, relatively young and well documented, and the segmental origins of the tones are very regular and well understood. The regularity of tonal change within Tai allowed for the creation of the 'tone box' (Gedney 1972), a compact visualization of the mapping between the modern tones of any Tai language and the posited historical environments that conditioned tone splits and mergers in that language. The tone box has been in wide use for historical analysis, language documentation, and dialectology in Tai languages for half of a century. Using tone boxes from hundreds of Tai doculects, this dissertation demonstrates that tone systems contain strong phylogenetic signal, a statistical measure of their historical informativity.

This dissertation advances theory and practice in historical linguistics, while demonstrating concrete advances in Tai historical linguistics. The Tai languages thus serve as a model for (1) a more generalized reasoning of why tonal evidence is not only possible to incorporate into a historical analysis, but will be a crucial element of the best historical analyses going into the future, and (2) how tonal evidence can resolve outstanding issues where predominantly segmental evidence has may have failed to do so. Using the insights of Tonal Comparative Method, we can expect the diachronic explanatory power of tone to extend well beyond the level achieved to date.

The Tonal Comparative Method: Tai Tone in Historical Perspective

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Rikker Dockum

Dissertation Director: Dr. Claire Bowern

December 2019

Acknowledgments

Above all I am deeply grateful to my wife, Kaw, and our children, Leslie and Henry, for all of their patience through the years of my doctoral program. From putting up with the penury of grad school to the many late nights and missed bedtime tuck-ins, I truly cannot imagine surviving this process, let alone thriving in it, without their support.

I also owe great thanks to my mother, Lori, and my sisters, Nanette, Nollie, Bekka, Kira, and Kaylee. They have been an unflagging source of positivity and encouragement throughout my graduate school years. I also owe so much of who I am to my late father, John, who has been gone 21 years, and finished his doctorate degree exactly 40 years ago.

It is impossible to overstate how much my experience at Yale has been defined by the influence of Claire Bowern. As my mentor from the start of my degree until chairing my dissertation committee, I have had the privilege to work with Claire in her lab and to co-teach two courses with her. I have gained immensely from her mentorship. As the Director of Graduate Studies for most of my years in the program, she also shaped the entire linguistics PhD degree at Yale for the better in numerous ways that I benefited from. Under her supervision I became a better researcher, a better writer, a better linguist. It has been an immense joy to be her student, and I hope ever to be her mentee.

I owe a great debt of gratitude to the Tai Khamti community at Khamti Township, Sagaing Division, Myanmar. My primary fieldwork consultants, U Kavipala and Sai Aung Mung, generously shared their language with me over the course of so many hours of recording, transcription, analysis, and translation. They also constantly went above and beyond my expectations in taking such good care of me in the face of torrential rains, flooded guesthouses, and numerous logistical challenges. The entire community welcomed me warmly with a generosity that I cannot forget. After hundreds of days worked and hundreds of mangoes eaten, I could never thank them enough. Special thanks also to the Yale Southeast Asia Council, the Yale MacMillan Center, and the National Science Foundation for funding my fieldwork and archival work in Myanmar and Thailand.

Thanks also to the linguists who set me irrevocably on the path to this dissertation: Lindsay Whaley and Ann Charity Hudley taught the first two linguistics classes I ever took, and Dave Peterson and Tim Pulju taught many of the ones that followed. Dave changed my brain when he introduced me to linguistic fieldwork. Thanks to Doug Cooper who saw an eager undergrad and channeled that energy into so many interesting projects.

Thanks are also due to the many professors, lecturers, post-docs, researchers, and fellow students who I have come to know at Yale and at conferences around the world. In particular, thanks to Ryan Bennett, who was an important early mentor to me in graduate school. My thanks also to my other committee members: Natalie Weber and Pittayawat (Joe) Pittayaporn. Through his stellar historical work, Joe was a huge influence on me well before I started the PhD, and it has been a privilege to have his feedback during this process.

This entire experience would not have been possible for me without the camaraderie and commiseration of my peers. I am grateful to have come to Yale with my original cohort: Luke Lindemann, Matt Tyler, and Patrick Patterson. And I'm grateful for the many colleagues and friends who had an impact on me along the way. Both those who came before me, including Emily Gasser, Jason Zentz, Natalie Schrimpf, Ryan Kasak, Alysia Harris, and Sean Gleason, and those who came after me: Josh Phillips, Martín Fuchs, Chris Geissler, Manu Quadros, Sarah Babinski, Andy Zhang, and many others.

Thank you, one and all.

For my wife and children, who have been so very patient.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

This dissertation presents a new historical approach to sound change in lexical tone for languages where tone is already firmly established. The terms *tone* and even *lexical tone* have been used in varying senses over the decades to describe the use of pitch or fundamental frequency in language. In this dissertation I adopt the following definition of lexical tone: the systematic and obligatory variation of prosodic acoustic cues, primarily pitch height and contour, to encode contrastive lexical meaning. Throughout this work, whenever the term *tone* is used alone, it is this sense of *lexical tone* that is intended, unless otherwise specified.

To date, the majority of attention given to sound change in lexical tone has focused on how an atonal language becomes tonal and on early stage tone development, a process known as tonogenesis.[1] Although tone has long been an object of direct linguistic study (e.g. Bradley 1911) and a connection between tones and segmental aspects of syllables is likewise a longstanding observation (e.g. Karlgren 1915), the first full account of tonogenesis appeared in the mid-20th century with Haudricourt's (1954) account of the

[1] The term 'tonogenesis' was coined by Matisoff, first fleshed out in Matisoff (1973), though first used in publication in Matisoff (1970).

origins of tone in Vietnamese. Perhaps the most crucial insight in accounting for tonogenesis is that lexically contrastive tone, a suprasegmental feature, is born from segmental origins. For instance, when differences in the VOT of onset consonants condition variation in pitch height. The process through which this happens is one example of what Hyman (1976) termed *phonologization*, his term for the process whereby intrinsic, mechanical effects of speech articulation come to be under conscious control and eventually are phonemically contrastive. A similar term is *transphonologization*,[2]. The two terms have slightly different perspectives on tonogenesis: as Hyman notes, phonologization does not imply transphonologization (2013: 9). The former is when an intrinsic effect becomes phonemic; the latter is when a set of phonemic contrasts is preserved in a language, but shifted onto different features. As such, tonogenesis is related to other examples of diachronic transphonologization like compensatory lengthening, where lost consonantal contrasts shift to vowels, or nasalization, when nasal segments disappear but survive as contrastive nasalization on adjacent vowels.

Despite longstanding attention to tonogenesis, what remains less studied and more poorly understood is how tone changes after it is well established in a language or language family. In the centuries following tonogenesis, we know that tones continue to undergo splits, mergers, and random drift, both in their phonetic realization and in the phonemic categories that underlie those surface tones. This dissertation advances our knowledge on this front, using data from the Tai languages, a branch of the Kra-Dai language family. See Pittayaporn 2009: 298 for the best current subgrouping of the Tai languages. The position of the Tai branch within the larger family is still a matter of open debate (Pittayaporn 2009: 5).

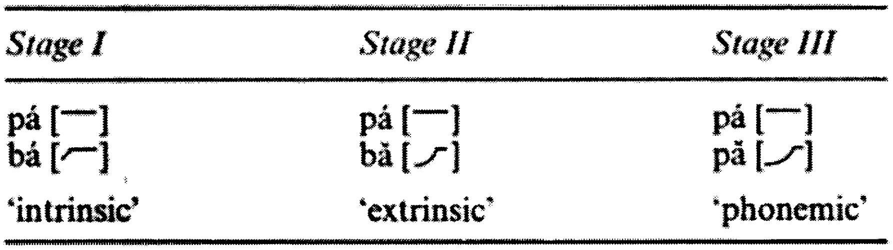
The Tai languages make an ideal testing ground for advancing the theory of sound change in tone systems because they are robustly tonal, relatively young and well docu-

[2] Related terms include *rephonologization* and *cheshirization*, the latter coined by Matisoff (1991), so called because segmental contrasts disappear but leave a trace of themselves behind, much like the smile of the Cheshire Cat in the Lewis Carroll novel.

mented, and the segmental origins of the tones are extremely regular and well understood. This allowed for the creation of the Tai tone box by Gedney (1972), which has been in wide use for historical analysis, language documentation, and dialectology in Tai languages for half of a century. Indeed, the concept of a tone box is so foundational in making this dissertation possible that I introduce it here in this introductory chapter, in §1.3. A more thorough account of its origins is given in §3.3 as well.

In this work I bring together two methodological approaches to the study of tone change. The first approach is the traditional Comparative Method (Hoenigswald 1993; Hock & Joseph 2009; Rankin 2003; Weiss 2014, *inter alia*), in which contemporary data from many languages is compared in order to classify those languages and reconstruct facts about their common ancestor languages. The second approach is the use of tools originally developed for Evolutionary Biology to enable the quantitative study of descent and change in biological systems, often referred to collectively as computational phylogenetics (Bowern 2018). Rather than being a single method, this is a family of tools and techniques that can be used to generate and test hypotheses about language history and relatedness.

Tonogenesis begins when some segmental contrasts of a language have pitch as a secondary phonetic cue below the level of speakers' awareness. Subsequent generations of speakers hear both types of phonetic cues, segmental and suprasegmental. Over time this initially redundant pitch cue comes to be the primary cue. The segmental cues, now secondary, are frequently lost in the process, eventually leaving each tone as the most salient cue for lexical contrastiveness in its respective subset of the lexicon. Hyman (1976) schematized this project as seen in Figure 1.1:



Stage I	*Stage II*	*Stage III*
pá [˥] bá [˦˥]	pá [˥] bǎ [˨˥]	pá [˥] pǎ [˨˥]
'intrinsic'	'extrinsic'	'phonemic'

Figure 1.1: Tonogenesis as illustration of phonologization (from Hyman 1976).

See also Hyman's breakdown into more detailed steps in Table 1.1, going from intrinsic and mechanical to speaker-controlled and phonemic:

Stage	Correlates
I. Intrinsic	(a) production - F_0 variations automatic (b) perception - F_0 variation a voicing cue
II. Extrinsic	(a) production - F_0 variations not automatic (b) perception - F_0 variation a tone cue
III. Phonemic	(a) production - same as stage IIa (b) perception - same as stage IIb

Table 1.1: Detailed breakdown of tone phonologization (from Hyman 1976).

This detailed breakdown is still a simplification, as there are always additional acoustic correlates of tone available to act as secondary cues to tonal contrasts, as evidenced by cue differences measured in whispered tone (Chang & Yao 2007). These additional cues may include amplitude, intensity, duration, and phonation (Rivera-Castillo & Pickering 2004; Yeh 2009). Given the variety of cues present in different tone languages, what we think of as phonologization of pitch is always a bundle of acoustic features which take varying degrees of primacy in different tonal languages, and the balance of which can shift over time. This is the basic diachronic process as generally accepted, however.

(Pittayaporn 2009: 248) describes a version of this process for subsequent tone splitting with accompanying loss of laryngeal contrasts, in which the stages are labeled Stage I to IV (see Figure 1.2). Importantly, evidence from Cao Bang Tai indicates that this kind of change can propagate partially without necessarily completing. Cao Bang Tai has neutralized voicing on initial sonorants but not on initial obstruents (Pittayaporn & Kirby 2017).

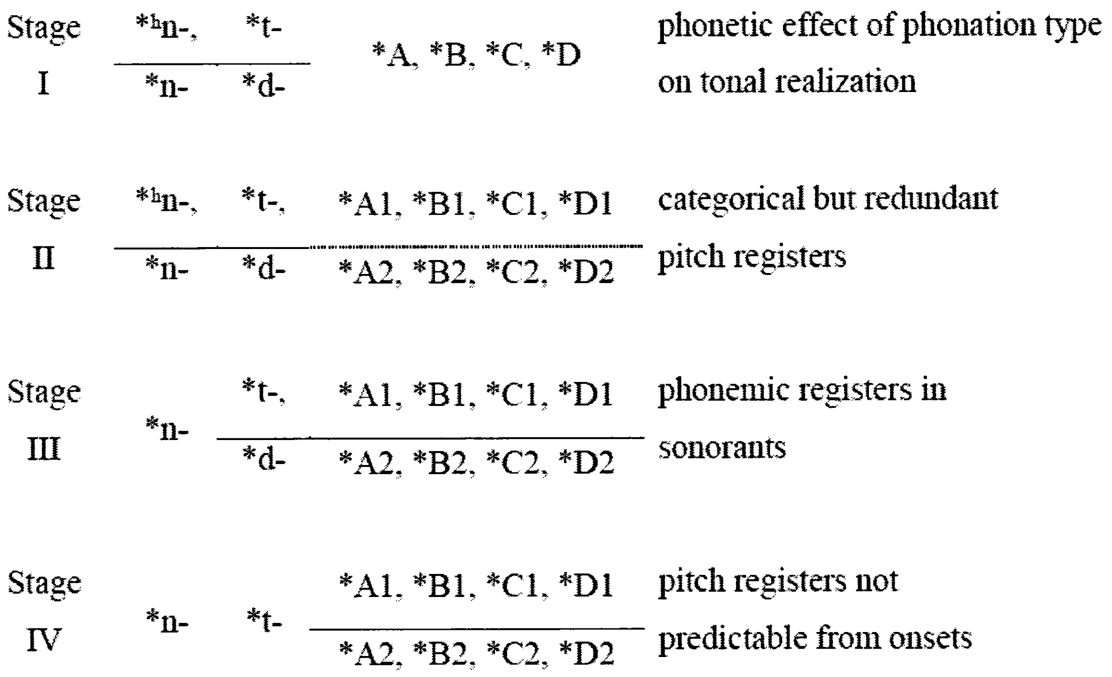


Figure 1.2: Tone split with phonation neutralization (from Pittayaporn 2009: 248).

Since the first descriptions of tonogenesis, the list of languages whose tone contrasts have known segmental origins has continued to grow, and now includes languages from many families. Tone is a major topic in linguistics for good reason: roughly half of the world's languages are tonal to one degree or another (Hyman 2018; Dryer & Haspelmath 2013).

1.2 Roadmap of the dissertation

In the remaining sections of chapter 1, I present the Gedney tone box, which is fundamental to this entire dissertation, and then I discuss a couple of issues in terminology and language classification. Following that, the remainder of the dissertation is divided into six chapters. A summary of each chapter is as follows:

In chapter 2, I present some of the results of an extensive survey of descriptive fieldwork on Tai languages and dialects conducted in Thailand over the past half century. Predominantly masters and doctoral theses, this body of work includes hundreds of works, little known and seldom cited by Anglophone linguists. I describe some of the coverage and trends in this area and samples of metadata from these works.

In chapter 3, I describe the history of language documentation conventions for lexical tone in Southeast Asia, and the development of the ubiquitous tone box approach in Tai linguistics. I also argue that much wider awareness is needed for regional differences in tone documentation conventions, especially with respect to what constitutes distinct tonal categories, as the adoption of one convention or the other can introduce implicit bias into our theory and analyses, both synchronic and diachronic.

In chapter 4, I discus the past use of various kinds of linguistic data in computational phylogenetic analyses. I demonstrate the presence of phylogenetic signal in phonological features of language. Using Tai data, I first show using segmental traits that there is phylogenetic signal in even coarse phonological data, confirming and expanding upon earlier work. Phylogenetic signal in segments is a statistical confirmation of the general validity of the traditional Comparative Method (CM). I then show that there is also strong phylogenetic signal in patterns of tone splits and mergers in Tai languages. This provides statistical confirmation for the theory of the Tonal Comparative Method (TCM), and lays the foundation for the fruitful use of tonal traits in computational phylogenetics in the future.

In chapter 5, I outline the theoretical foundation for the Tonal Comparative Method, a proposed extension of the traditional CM. I discuss how and why tone has been met with skepticism in diachronic linguistics, and generally excluded from the CM in the past. I then lay out an argument for how we can incorporate tonal evidence in a scientifically rigorous way that is consistent with the logic of the traditional CM. I explain how to apply the Tonal Comparative Method at two different stages of linguistic analysis: first in the early stage, how to identify tonal correspondence sets, and discover the segmental origins of tone in a set of related languages where nothing is previously known; and second, in an advanced stage, where segmental origins are well understood and uncontroversial, as with the Tai languages, I discuss a method for categorizing tone changes to differentiate shared innovations from parallel innovation and shared inheritance. Finally, I discuss limitations of the method.

In chapter 6, I present a case study with data from three tonally and temporally disparate dialects of Tai Khamti, demonstrating how to use the logic and methods outlined in chapter 5 to reconcile tonal variation in closely related dialects, how to reconstruct the tone system of their nearest common ancestor, and how to resolve fine-grained classification issues where the traditional CM has as yet not yielded a consensus, or is poorly equipped to do so due to segmental homogeneity.

Chapter 7 provides concluding comments to the dissertation, and outlining directions for future work.

1.3 The Gedney tone box

This entire dissertation is made possible by the discovery of the historical factors that conditioned tone splits in Tai languages, and the equally important discovery of the extreme regularity of the relationship of modern tone categories and the environments that conditioned them. The relationship between the two is conventionally visualized using the Gedney (1972) tone box, shown in Figure 4.11. (See also 3.3 for a fuller account of the development of this convention.)

	Proto-Tai tonal categories				
Proto-Tai initials	A	B	C	D-short	D-long
Voiceless friction *p^{h}, *t^{h}, *k^{h}, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

Figure 1.3: Tone box for Tai historical analysis, adapted from Gedney (1972)

The tone box is a compact way of mapping surface tones back to the features of the proto-onsets in Proto-Tai that conditioned splits and mergers in thc tonc system. The first three columns A, B, and C each represent one of the original tones of Proto-Tai, found on open and sonorant-coda syllables. The fourth column, D, represents syllables with stop codas, which pattern together tonally, and can be further subdivided in some Tai languages based on vowel length. Each row represents a former natural class in Proto-Tai, the laryngeal configuration that conditioned tone, followed by some neutralization of the former segmental contrast in most cases. The end result is a grid of 20 potential conditioning environments, each representing a subset of the native lexicon that patterns together tonally in modern languages. While Gedney built on the work of others, it is hard to overstate the impact of this work on Tai linguistics, as the 'Gedney box' remains a ubiquitous feature of language documentation in the family. The Gedney box is so central to this dissertation that it merits immediate introduction.

Even before the paper that introduced it was published, students and colleagues of Gedney were using the tone box. It quickly became a near ubiquitous feature of Tai language documentation, which continues to the present. It has lesser adoption among scholars trained in the Sinological tradition, which has its own set of conventions for studying historical tone.

1.4 Terminology and classification

1.4.1 Language, dialect, lect, doculect

Throughout this dissertation I use the term *lect* to refer to a language variety, in order to remain agnostic on the status of a particular variety as a language or dialect, and the problems that accompany those terms. Certainly much of the data used herein is at the dialect level, just as much is from undoubtedly distinct languages. To avoid having to define the border area between them, I prefer to use *lect* wherever I need to avoid that ambiguity. I also use the term *doculect* (Good & Cysouw 2013). This term was coined to be able to refer to any documented variety of spoken language, tied to a particular documenting linguist, a particular speech community, and a specific place and point in time. In other words, a doculect is an observation of a lect.

1.4.2 'Tonogenesis' vs. 'tone change'

Use of terminology for describing changes in tone systems has varied. The original use the term 'tonogenesis' has been to describe the process by which a language goes from being atonal to tonal, that is, the acquisition of tonemic (phonemic tone) contrasts in a language (Matisoff 1973). Others (e.g. Hyslop 2007) have extended the term to include any change in the tone system of a language where phonetic variation becomes phonologized into a new tonal category, even when this change results in the split or merger of existing tone categories. In this sense, changes to the tone systems are often referred to collectively as 'tonogenetic events.'

The use of the term 'tonogenesis' in both senses is clearly justifiable, but it is useful to be able to differentiate initial tonogenesis, the atonal to tonal stage, from sound change within a well-established tone system, which, so far as we know, continues in perpetuity for as many centuries after tonogenesis as the language remains tonal. Tone surveys in Tai dialectology suggest that tone is an extremely dynamic area of phonology, and that sound change in tones and tonal categories may happen more quickly than segmental change. Surveys that support this notion include studies such as Sopheap (2017), in which the phoneme inventory remains the same between closely related doculects, and tone appears to be the main locus of variation.

Assuming for the time being that tones and tonemes change no slower than segments, and are subject to the same perpetual sound change as the segmental domain, it is useful and necessary to distinguish these two types of sound change. I follow Ratliff (2015) in making this distinction, using tonogenesis to refer exclusively to initial onset of toneme contrasts, and tone change to refer to subsequent change.

1.4.3 ISO 639-3 codes and glottocodes

There is no sole standard for how many Tai languages there are. The two standards for language identification in wide use in linguistics are ISO 639-3 and Glottolog. Both are standards for assigning a unique and persistent identifier to all natural languages. Within these two standards, an individual code is known as an ISO code or a glottocode, respectively. The registration authority for ISO codes is SIL International, since the standard originally grew out of the codes used in SIL's Ethnologue resource Eberhard et al. (2019). Glottolog is administered by the Max Planck Institute for the Science of Human History (Hammarström et al. 2019b).

ISO codes and glottocodes are not quite equivalent: Glottolog assigns glottocodes to everything it terms a *languoid*, a term that includes dialects, languages, and language families Hammarström et al. (2019a). Every node in the Glottolog tree, terminal and non-terminal alike, is assigned a glottocode, and each glottocode is classified as one of four types: family, subfamily, language, or dialect. Thus, ISO codes and language-type glottocodes are at roughly the same level of specificity.

As of May 2019 there are 58 ISO codes classified as Tai languages in Ethnologue.

For Glottolog, the equivalent branch is labeled Daic, with 58 language glottocodes, and approximately 50 more dialect glottocodes. Although the number 58 happens to coincide between the two at the time of this writing, these are not exactly the same in the two systems. For example, there are two ISO codes [nyw] and [tyj] for Nyo (also known as Yo), spoken in Thailand, and Tai Yo, spoken in Vietnam. Glottolog instead considers the former a variety of the latter, and assigns Tai Yo to [taid1248], a language-level glottocode, and Nyo to [nyaw1245], a dialect-level glottocode.

Chapter 2

Tai language documentation and dialectology

2.1 Introduction

Documentation of tonal languages has amassed for several decades since tonogenesis came to be well understood, much of it written in languages other than English and created by linguists working within the Sinospheric Tonbund. Much of this large body of documentary work is little known and seldom cited in western linguistics literature, even by specialists, creating a gap in the awareness of the larger linguistics community (Dockum 2018). Closing this gap is necessary for a more complete understanding of change and descent in established tone systems. Often produced a generation or more ago, many of these reports also describe endangered doculects, or speaker communities whose linguistic situation have changed significantly in the intervening decades, making them all the more valuable for building a more complete theory of tone system change. A disconnect of this magnitude also produces a negative feedback loop: our larger theory of sound change in tone systems suffers from lack of awareness and access to all of the new data being documented, and crucial language documentation tasks, which may prove to be the only work ever done in some speaker communities, cannot benefit from improved theories of tone that this lesser-known data could help inform.

This chapter reports some of the results from an extensive survey of language documentation on Tai tone systems that has been conducted over the last half century, with a focus on the vibrant language documentation tradition in Thailand's universities. The majority of the work surveyed thus deals with lects spoken within the geographical bounds Thailand, though some also deals with border areas, such as Lao dialects on the Laos-Cambodia border (Sopheap 2017), or Tai lects further afield, such as Zhang's (1999) documentation of Zhuang languages in China, part of the Northern Tai subgroup. Hundreds of theses, books, articles, and research reports were identified, giving access to a massively increased amount of data from Tai tone systems, and at a level of unprecedented detail. Each source reports on anywhere from one to dozens of doculects, varying in granularity from village-level to province-level tonal variation. Many sources also document the overall phonology, while others are more akin to sketch grammars with a chapter on phonology. Others still are dialectology surveys reporting on detailed lexical variation, or variation between generations of different generations of a single doculect, and thus contain little detail about tone. However, I have chosen to also include these in order to give a more complete accounting of the type of documentary work that has been done. Some are written in English, though most are in Thai, and wordlists are almost always recorded in the International Phonetic Alphabet with Thai or English glosses, or both. Tones are predominantly notated using one of the two systems stemming from Chao's (1930) five-pitch scheme, though this is variously represented by either the Chao tone *letters* (e.g. /˧˥/, mid-to-high rising tone), or Chao tone *numerals* (e.g. /35/, for the same tone). Others use symbolic numbering 1 through *n* for the distinct tonemes, though these symbolic numerals are in turn translated into either Chao letters or numerals in the surrounding discussion.

Raising awareness of lesser known fieldwork output in the wider linguistics community will yields benefits by increasing our ability to study how lexical tone systems diversify in Tai. This serves as a model for how we can improve theory and results by putting in the work to identify and aggregate lesser known resources, especially materials written in non-western languages and countries, wherever such underutilized bodies of work may exist.

2.2 Grey literature in linguistics

Linguistic analyses rely on access to data. However, there is no strong general convention in linguistics of making an original dataset available along with an analysis. However, the existence of the Austin Principles for Data Citation (Berez-Kroeker et al. 2018) point toward community recognition of this traditional lack of transparency. An improved culture of data sharing is needed to address the portions of the replication crisis (Schooler 2014) that apply to linguistics. If this is true now, it is even more true of 'legacy' data or archival data, as the literature surveyed here represents, which is made even more difficult by existing only in analog form on bookshelves, or even if digital, in paywalled, closed-access, or even paper-only publications.

It is fruitful, then, to apply the notion of 'grey literature' to linguistics. The theses uncovered in my library research for this dissertation can be thought of as a type of grey literature. Precisely what falls under the label of grey literature is a matter of debate. It was defined at the 12th International Conference on Grey Literature in 2010 as follows (University 2019):

> Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by libraries and institutional repositories, but not controlled by commercial publishers; i.e. where publishing is not the primary activity of the producing body.

Here I adopt the more fluid notions of *discoverability* and *accessibility* as metrics for greyness. In other words, linguistic grey literature is research output that falls outside of the traditional channels of publishing and dissemination. This includes dissertations and theses, conference handouts, working papers, field notes, government-sponsored language documentation, and organization-internal reports, such as surveys created by groups like SIL. I also extend this meaning to include traditional academic material produced in non-Anglophone or non-European countries, especially those not published in a national European language.

Linguistic grey literature may very well be publicly available online in electronic format, but of course this is not necessarily so, especially for older works. Further, even if works are available online somewhere, they are often not indexed in indices like Google Scholar. And of course, something that is available online today must not be assumed to available indefinitely. The result of all of this is that linguistic grey literature is difficult either to discover or to access, and often both, and thus large swaths of work can be considered functionally nonexistent for many scholars.

Grey data in linguistics is of course data that is sourced from, or that contributes to, grey literature. In language documentation, the notion of grey data is important because of the pressing nature of language endangerment and extinction. Time and resources to document rapidly vanishing languages are obviously limited, and so a more complete accounting of what work already exists is key to maximally effective use past, present, and future resources.

2.2.1 Discoverability

The discoverability of grey literature typically relies on aggregators, which vary from academic sites such as Glottolog (Hammarström et al. 2019b), to commercial sites such as Google Scholar and Academia.edu. More generally, search engines that index websites help to close the discoverability gap.

In Thailand, aggregators have been around for many years, run by university library consortia like ThaiLIS (Thai Library Integrated System) or academically-focused governmental bodies like the National Research Council. The vast majority of theses uncovered for this dissertation have apparently never been cited outside of Thai academia, and sometimes required great going to great lengths and many trips to university stacks to find certain items. Other times, when they were in national aggregators, discoverability was more amenable. In any case, it required a massive time investment over the course of several years, and even knowing where to look presented numerous challenges.

2.2.2 Accessibility

Access to grey literature in Thailand has benefited from a general trend that has taken shape over the last several years toward digitization of academic output, especially graduate theses and dissertations. Many universities now make some or all of their catalogs of dissertations available as PDF files, although downloads are often only available from campus-internal IP addresses, or in the case of the ThaiLIS (Thai Library Integrated System), by creating an account that requires a national identity card number, which only Thai citizens have. Fortunately, there is some redundancy, where resources in a more restricted source may be openly available directly from a university archive. Such restrictions are of course also seen in Anglophone linguistics, such as with the more egregious commercial services, like Academia.edu, which requires a logged in account to download material, and even then bombards the user with attempts to further engage them on the site.

At the time of my surveying, many other Tai language documentation works were not yet available in any digital form. This required many in-person trips to academic libraries at various Thai universities to identify existing resources and fill gaps where possible. Despite various trips over multiple years, typically before or after fieldwork in Myanmar, there are a few dozen works still not in hand.

Limitations notwithstanding, the overall trend is very positive and moving towards greater openness. This naturally leads to additional points about greyness that are worth taking to their conclusion, especially with respect to academic publishing.

2.2.3 Commercial academic publishing as grey literature

This brings up some additional points about greyness that are worth taking to their conclusion. Using the criteria of discoverability and accessibility discussed above, it should be clear that 'greyness' is a relative notion. Thus, while academic output in Thailand is little known to Anglophone linguists and can thus be considered grey literature, the academic output of linguists working in regions like Europe and the United States may also be undiscoverable or inaccessible to many scholars in Thailand. The approach of many academic publishers is copyright-driven, producing artificial scarcity to drive profits. The subscription costs of academic journal aggregators can be exorbitant and prohibitive, resulting in an imbalance of access to the current body of literature. From the perspective of developing nations, especially at any university except the most prestigious ones within those nations, it may indeed be the case that the majority of linguistics literature *ever produced* can be considered grey literature, undiscoverable or inaccessible to our peers in other nations.

The dawn of the digital age brought with it the promise of radically open access to knowledge, and in many ways this promise has been realized. Academic publishing as a rule is one notable exception to this ideal, with the proliferation of paywalls, use of political lobbying, and other tactics intended to prevent free distribution of scientific output. Lobbying by academic publishers like Elsevier contributed to proposed legislation in the form of the Research Works Act of 2011. This bill sought to outlaw open-access requirements for federally funded research in the United States. Elsevier and others withdrew their support after a fierce backlash (Elsevier 2012).

In extreme cases, the commercialization of academic work has had tragic outcomes, as in the death of Aaron Swartz. After Swartz used an internet connection at MIT to bulk-download 4.8 million academic articles from journal aggregator JSTOR, with the possible intent to make them freely available (but without having taken any steps to do so), Swartz was aggressively prosecuted by federal law enforcement. Despite JSTOR declining to pursue civil litigation against him, Swartz faced 13 felony charges, 35 years in prison and$1 million in fines. Swartz died by suicide shortly thereafter. Swartz's father publicly stated that MIT played a "central role" in his son's death, and the university's own internal report found that the university "didn't do anything wrong; but didn't do [itself] proud" by deciding not to intervene on Swartz's behalf in this "ruinous collision of hacker ethics, open-source ideals, questionable laws, and aggressive prosecutions," ultimately calling it a textbook example of the type case where MIT could have done good if it had chosen to Massachusetts Institute of Technology (2013). The tragic silver lining was that JSTOR responded to Swartz's death by publicly releasing 4.5 million of the 4.8 million files that were at issue to begin with.

The Research Works Act and the Swartz tragedy serve as stark reminders of the problematic and at times harmful nature of academic publishing driven by profit motives, the effects of which are disproportionately born by our colleagues in in developing nations.

2.3 Categories of documentation

In this section I provide an overview of some major categories of the Tai language documentation materials surveyed, with examples. The literature can be divided into a number of categories: tone documentation (which could be further subdivided into tone surveys, single language studies, and phonological sketches), lexical documentation (comprising "word geography" surveys and multi-generational lexical studies), and areal dialect surveys. Full bibliographic details for these works can be found in the references section at the end of the dissertation.

2.3.1 Tone documentation

This category consists of works whose primary aim is to document some aspect of lexical tone, including surveys of dialectal tone variation, or phonetic studies of particular lects. Examples of some works and the areas they document are provided here.

Author	Year	Language	Location
Koowatthanasiri	1981	Nyo	Sakon Nakhon; Nakhon Phanom
Ratanadilok Na Phuket	1983	Central Thai	Ratchaburi
Sritararat	1983	Phu Thai	3 provinces
Tienmee	1983	Khorat Thai	Nakhon Ratchasima
Kopprayun	1986	Tai Yoy	
Tanlaput	1988	Northern Thai	Lampang
Chinchest	1989	Lao Ngaew	Singburi
Aruneeung	1990	Central Thai	Bangkok
Tingsabadh	1990	Central Thai	Suphanburi
Panroj	1991	Central Thai	Bangkok
Kobsirikarn	1992	Central Thai	Suphanburi
Nualjansaeng	1992	Central Thai	Nakhon Pathom
Banditkul	1993	Central Thai	Prachuap Khiri Khan
Pornsib	1994	Central Thai	Phetchaburi
Krisnapan	1995	Central Thai	Phetchaburi
Sumransook	1995	Central Thai	Chonburi
Komontha	1996	Khorat Thai	Nakhon Ratchasima
Worawong	2000	Central Thai	Kanchanaburi
Khamrueangsi	2002	Nyo	7 provinces
Khemkhaeng	2002	Nyo	Mahasarakham
Khotchanthuek	2002	Multilingual	Nakhon Ratchasima

2.3.2 Lexical documentation

This category includes works that study the lexicon, include lexical isoglosses, lexical shift, and generational change in lexical usage. A sample is provided below, along with the region associated with each province (N = Northern Thailand, C = Central Thailand, NE = Northeast Thailand, and S = Southern Thailand).

"Word geography" surveys

Author	Year	Province	Region
Weesakul	1983	Sukhothai	N
Phanuphong	1984	Nakhon Ratchasima	NE
Ache	1986	Surat Thani	S
Peamphermphoon	1986	Buriram	NE
Maliwan	1987	Saraburi	C
Nakpuntawong	1987	Uttaradit	N
Sukpreedee	1988	Rayong; Chanthaburi; Trat	C
Boonkao	1989	Mahasarakham	NE
Panarat	1990	Lopburi	N
Sombatmaungkan	1990	Sakon Nakhon	NE
Chulkeree	1991	Phichit	N
Sawangwan	1991	Chaiyaphum	N
Thepsakunrat	1991	Songkhla	S
Burusphat	1992	Phetchabun	N
Jaipakdee	1992	Nakon Si Thammarat	S
Suwannaraj	1993	Ubon Ratchathani	NE
Thumsaro	1993	Pattani	S
Thikhachunhathian	1994	Loei	NE
Chaisakulsirin	1995	Lopburi	N
Phumcharoen	1995	Chiang Mai	N
Somboonsak	1996	Prachinburi	C

Three-generation lexical surveys

Name	Year	Language	Location
Tanyong	1983	Phuan	Lopburi; Singburi
Buranasing	1988	Black Tai	Suphanburi
Saraporn	1988	Northern Thai	Ratchaburi
Liamprawat, Watthanaprasoet	1996	Lao	Tha Chin River basin
Patpong	1997	Northern Thai	Sukhothai
Saengsrichan	1998	Tai Lue	Phayao
Jitbanjong	2002	Saek	Nakhon Phanom
Suwanmusik	2004	Southern Thai	Koh Samui, Surat Thani
Sornjitti	2007	Southern Thai	Chumphon
Plodkaew	2008	Southern Thai	Nakhon Si Thammarat
Thongchalerm	2008	Northeastern Thai	Ubon Ratchathani
Moontuy	2010	Yong	Chiang Mai
Jidlang	2012	Southern Thai	Trang
Tebpawan	2012	Southern Thai	Phangnga

2.3.3 General dialect surveys

Finally, the other major category is dialect overviews of certain geographic areas, which tend to be titled "Current Thai Dialects of (Location)".

"Current Thai dialects" surveys

Author	Year	Province
Thongchuay	1983	Kelantan, Kedah and Perlis
Nuansanong	1984	Phuket
Sawangchit	1986	Yala
Kaenkrachang	1987	Chumphon
Nuansanong	1987	Nakhon Si Thammarat
Wetchasit	1987	Narathiwat
Khwanritti	1987	Songkhla
Phitsaphak	1988	Krabi
Sikhwan	1988	Pattani
Sirinuphong	1988	Phangnga
Keochana	1988	Phatthalung
Daengwan	1988	Ranong
Thongphenchan	1988	Trang
Manomaya	1989	Surat Thani

2.4 Quasi-longitudinal analysis of tone systems

One result of improved access to this large body of documentary work is the possibility of "quasi-longitudinal" study of the tones of particular languages, based on the work of multiple authors. The point of such a study would be to trace tone change in a language or dialect as if it had been studied over period of time under a single project. Thus I define quasi-longitudinal analysis as follows: the use of two previous studies sampled from the same language community, which takes the language in one study to be a direct ancestor of the language in the other study, rather than a temporally displaced sister language of it. More than two studies should be possible, through the transitive property, though quasi-longitudinality should ideally be determined pairwise.

In such a case the speakers in the two studies will not be the same (as if they were then it would be true longitudinality). The benefit of this notion for diachronic analysis should be clear: genuinely longitudinal studies are very rare, but in areas that enjoy a relatively rich local language documentation culture, as Thailand does, studies that gathered data from speakers in the same approximate area should be suitable for drawing generalizations about multiple stages of a single lect. Below I argue for geographical proximity in the Tai dialectology context as the ideal way of determining quasi-longitudinality, and argue against

Exactly what level of proximity is required to make a set of studies suitable for quasi-longitudinal analysis requires some consideration, which the remainder of this chapter addresses.

2.4.1 Geography and community size as metrics for quasi-longitudinality

The granularity of dialectology surveys varies widely. Focusing on the large body of documentation fieldwork by researchers in Thailand, which is little known and seldom cited in the English-language linguistics literature, it is thus helpful to examine some facts about local administrative units in the country in order to determine what constitutes reasonable quasi-longitudinality. Note that this is not an argument about classifying doculects as languages or dialects, but rather about whether, for the purposes of drawing inferences about sound change, we can reasonably treat two studies as being samples of the same language variety at two different points in time. If two studies purporting to document the same language, written decades apart, come from the same village, it is likely to represent the same community. If they overlap only at the province level, however, it could be problematic to assume it is the same lect.

At present there are 77 province-level units, comprising 76 provinces /tɕaŋwat/ and the

Bangkok Metropolitan Area.[1] The province-level units comprise 878 districts /amphɤː/, plus another 50 districts /kheːt/ in Bangkok proper (Ministry of Education of Thailand 2016). District populations range from 2,000 people to over 500,000 people. Each district is further divided into subdistricts /tambon/, with more than 7,200 subdistricts nationwide. Below subdistricts, the lowest administrative unit is the village /mu: ba:n/, of which there are roughly 75,000 nationwide. Thus, each province has an average of 12 districts, each district has an average of 8 subdistricts, and each subdistrict has an average of 10 villages. According to the 1990 census, there are an average of 746 people in each village (National Statistical Office of Thailand 2019).

Given the widespread travel and telecommunications infrastructure that Thailand has, I argue that if fieldworkers document language of different villages within the same subdistrict, provided they identify those doculects as the same language, we can reasonably treat these as samples of the same language or dialect. Given the variable size of district and subdistricts, in many most cases the same will be true for documentation that takes place within the same district. However, I recommend caution in dealing with data from different documentation projects that align only at the province level or higher. I make this recommendation for a few reasons.

First, the notion of "dialect" in Thailand is unavoidably influenced by national and regional geopolitics. The popular conventional wisdom in the country is that Thailand has four primary dialects corresponding to the four major regions: Central Thai, Northern Thai, Northeastern Thai, and Southern Thai. Treating them as dialects of a single language is a modern oversimplification. Even where a specific province has a well-known dialect, the variety spoken in the urban area of the provincial capital may be taken as representative of the entire province, or in the case of Bangkok, the entire region. The result is that

[1] This number has grown from around 70 in the mid-20th century, which is when native documentation of Tai dialects began in earnest, as various provinces have been split due to population growth. The only instance in that time period of provinces being combined was in 1972, when Phra Nakhon and Thonburi provinces jointly became the Bangkok Metropolitan Area.

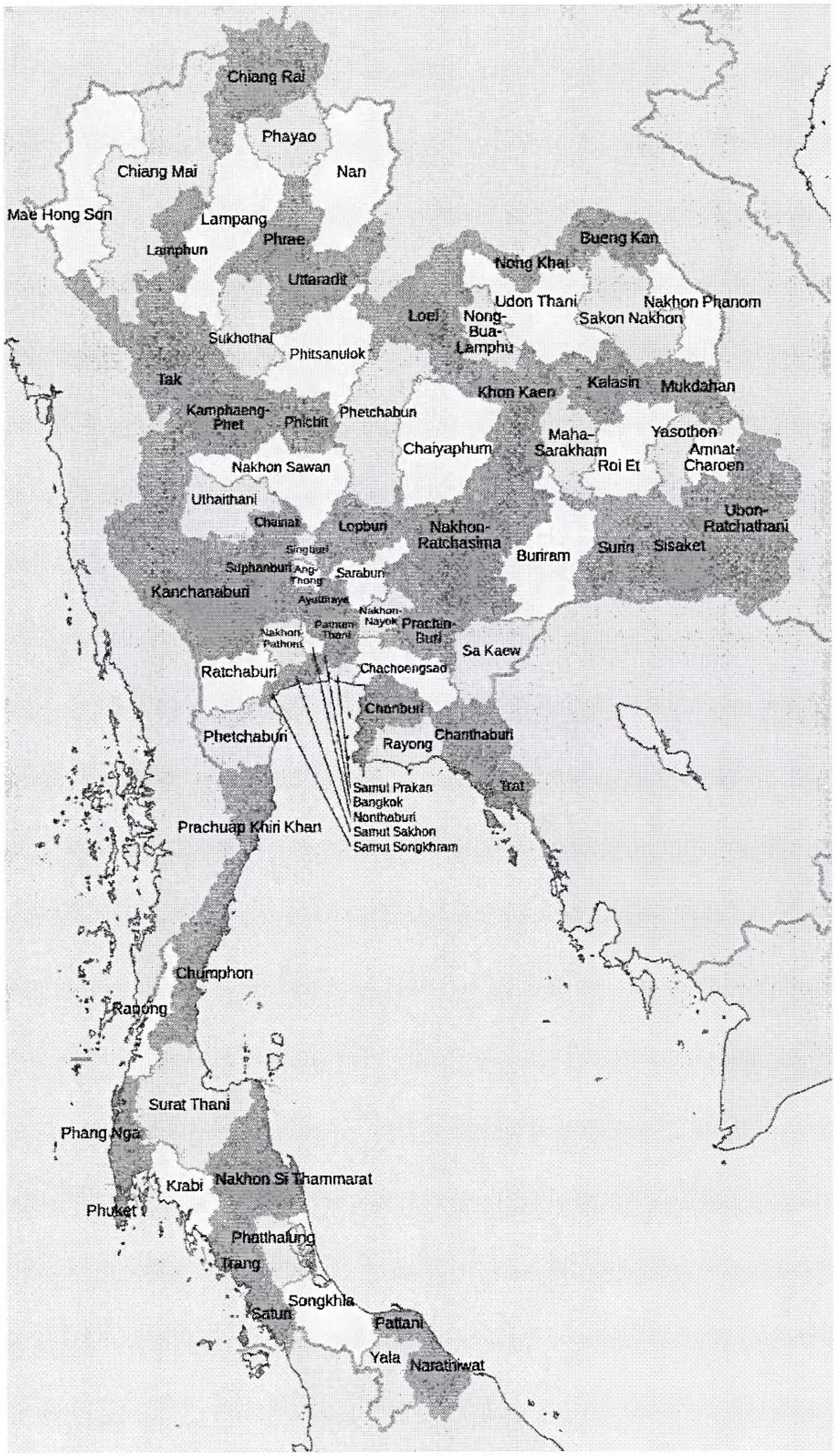


Figure 2.1: Province map of Thailand (Source: Wikimedia, CC BY-SA 3.0 license).

linguistic facts may be treated as monolithic at the province or regional level, when the reality is of course that language varieties neither expand to fill administrative boundaries, nor are they confined within them. But geopolitical notions of language diversity often filter into research design and results reporting in the actual dialectology fieldwork.

Second, we have many surveys that show important variation between data collection points within the same province. As a result of the issues discussed in the previous paragraph, two studies that identify a Tai language primarily based on the province may not be carefully distinguishing doculects with important differences. So we would not necessarily want to claim that Chiang Mai Thai from Suntharawakun (1962) and Hudak (2008) are the same lect.

Third, there is significant room for confusion with language endonyms and exonyms. For a very large number of Tai languages, the group endonym is cognate with "Tai" or "Thai." For example, the Tai group known variably as Song, Lao Song, and Thai Song (e.g. Saeng-ngam 2006) are also known as Black Tai in some sources. However, the connection to the group documented in northern Vietnam by Gedney and others (Hudak 2008) is of unclear relation, and even if there were some historical migratory or genetic connection, that will not necessarily make for similar doculects.

This of course does not apply to cases of Tai languages that are the result of much later migration events than those that led to the major dialect divisions. For instance, we would not want to confuse Yong [yno] of Lamphun province (Pankhuenkhat 1978; Neamnark 1985; Soiyana 2009; Moontuy 2010) with the Lamphun dialect of Northern Thai [nod] (Chaisri 1984).[2]

Table 2.1 gives a sample of some purportedly province-level studies, and the variation in detail contained therein. Some, such as Suntharawakun (1962), equate a specific dialect with an entire province. Others, as in Withayasakpan (1979a), study dialectal variation

[2] In fact, there are entire dissertations dedicated to comparing the phonology of these two languages (Wangsai 2007) and their sociolinguistic situation (Panrerk 2004).

Author	Year	Province	Region	Sites	Speakers	Focus
Suntharawakun	1962	Chiang Mai	N	1	1	General
Ache	1986	Surat Thani	S	228	228	Lex
Peamphermphoon	1986	Buriram	NE	—	127	Lex
Phanuphong	1984	Nakhon Ratchasima	NE	25	25	Lex
Boonkao	1989	Mahasarakham	NE	99	—	Lex
Withayasakpan	1979	Rayong	C	5	5	Phon
Ngaorangsi	1982	Phitsanulok	N	93	93	Tone

Table 2.1: Province level studies with highly varying degrees of detail.

in the majority regional language of that province. Others still, such as Ngaorangsi et al. (1982) and Boonkao (1989), survey both majority and minority Tai languages spoken in a province, with less focus on dialectal variation.

A more complete profile of language documentation and dialectology work conducted in Thailand on the Tai languages within the last 50 years could itself be a book-length work. This chapter serves only as an introduction to the types of work that have been done, and to its sheer volume. In a small way it also bridges the gap between these works, seldom cited and little known outside of Thailand, and the larger linguistics community.

Chapter 3

Tone documentation conventions: Issues for synchrony and diachrony

3.1 Introduction

In this chapter I describe the development of conventions for documenting and describe tone in Tai languages, and then discuss problems that arise from competing conventions in differing regional traditions for documenting tone.[1] Overly rigid regional conventions of both types have likely resulted in systematic under-documentation or mis-documentation of the phonetic and phonemic detail of tone systems, which then filter into both our historical and synchronic theory via descriptively inaccurate or incomplete data.

The main example of this that this chapter covers is the treatment of syllable shape in tone diachrony, and the two competing documentation conventions for handling tones on 'checked' syllables (i.e. stop-coda syllables). I call these two documentation conventions the *subset convention* and the *disjunction convention.* In the subset convention, checked-syllable tones are always assumed to be allotones of the most phonetically similar smooth-syllable tone. That is, the checked-syllable tones are always a *subset* of the smooth-

[1] Although Tai is the focus of this dissertation, many of the conventions developed in parallel with knowledge of tone more generally, and thus some observations are applicable throughout the Kra-Dai family, beyond it to the rest of the Sinospheric *Tonbund* (see 5.4), and perhaps to tone languages universally.

syllable tones, and in nearly all cases a proper subset as well. The subset convention holds for most work on tonal languages in mainland Southeast Asia. The *disjunction convention* is common in the Sinological tradition, and in that system, the checked syllable tones are always treated as disjoint from the smooth syllable tones—two separate sets tones with distribution limited by syllable type.

A given convention may also be embedded in the teaching and spelling traditions of a language, making it easy to locate native speakers who can identify tones consistent with that convention, in a way that appears to be accessing their phonological competence, but is in fact biased by pedagogy and their own literacy. Interference from pedagogy and literacy creates a problem for many kinds of experimental work on tone (and is a question that needs more attention in experimental work in phonology generally).

Descriptive accuracy, but also comparability, are essential for development of our synchronic and diachronic theory alike. To date, study of tone diachrony has focused primarily on tonogenesis, and less on how tone systems change and diversify after tone is well established. Differing conventions for tone documentation obscure important comparability, over matters such as what constitutes a distinct tonal category, leading to differences as basic as how many tones we count a given language as having, and how we characterize the distribution of those tones. Our incomplete understanding of tone change, combined with descriptive inconsistencies, have contributed to an oversimplified view of complexity in tone systems, and to some problematic assumptions in prior work.

One example is an overly binary view of tonality, such as the tonal-atonal dichotomy for classifying languages. Brunelle & Kirby (2015) show that this is an inadequate way of looking at tonal diversity in Southeast Asia, where tone language "exhibit a wide range of diversity, from simple two-tone systems based exclusively on pitch to complex tone systems combining large numbers of contrastive pitch units and voice qualities" (2015: 104). Another example is the ungrounded distinction between "simple" and "complex" tone systems in the World Atlas of Language Structure (Dryer & Haspelmath 2013), where *simple* is defined as having two tonemes, and *complex* is everything three or above. In reality, the variable vectors of complexity in tone systems are much more than toneme count, including interactions with register, phonation, and stress. This simplification has contributed to the assumption adopted in some literature that tone systems gain or lose contrasts linearly, one at a time Collins (2016), or the implicit assumption that a language with n tonemes is less complex than a language with $n+1$ tonemes, as in Everett et al. (2015).

In the remainder of this chapter, in §3.2 I provide a history of tone documentation in Tai languages, followed by the development of the omnipresent 'tone box' in §3.3. In §3.4, I describe two related problems in tone language documentation, the Checked Tone Problem (3.4.1) and the Orphaned Tone Problem (3.4.2).

3.2 History of Tai tone documentation

In this section I provide a general background for how Tai tone has been documented from the inception of Thai writing more than 700 years ago, and culminating in the Gedney (1972) tone box being the primary tool for historical tone analysis for the last 40 years. This also serves as essential context for the problems described in §3.4.

Evidence on the tone systems of Tai languages comes from a variety of sources. Three general categories emerge: (1) tone marking in native writing systems; (2) tone in early linguistic descriptions by Westerners; and (3) modern linguistic fieldwork of both native and non-native linguists.

3.2.1 Tone in native orthography

Old Thai is the earliest known practical orthography that marked phonemic tone (Diller 1996: 241). As a result, the earliest record of tone in Tai languages is in epigraphic texts of the Sukhothai Kingdom (13th-15th centuries CE). Epigraphy is the study of ancient inscriptions—texts inscribed on durable materials, most often stone, metal, and ceramics—and Southeast Asia has a long epigraphic tradition. It begins in the 5th century CE, following the spread of Brahmic scripts into the region. The rise of writing is directly linked to the spread of Hinduism and later Buddhism, and the use of sacred texts in Sanskrit and Pali, their respective liturgical languages. Scripts were adapted for the local vernacular languages, including Khmer, Mon, Burmese, and Pyu.

Epigraphic texts of those linguistic traditions predate the arrival of the Tai diaspora in Southeast Asia, a date that is somewhat disputed (CITE for best current estimates), but scripts for Tai vernaculars began to appear some time after their arrival. Beginning in the 13th-14th century CE we find the earliest surviving texts in a variety of Tai languages, from kingdoms including Sukhothai, Ayutthaya, Lanna, and Lan Xang. A cluster of related scripts were adapted by different Tai polities from the surrounding writing traditions, which by then were well established. A version of the Khmer script used at Angkor was the basis for the Sukhothai, Fakkham, and Tai Noi scripts, the vernacular scripts used in the Sukhothai, Lanna, and Lan Xang kingdoms, respectively. Sukhothai script is the ancestor of modern Thai script, and Tai Noi developed into modern Lao script. An older stage of Mon is the model for a Lanna liturgical script, which later came to replace Fakkham as the vernacular script of Northern Thai, and is still in use as a secondary traditional script in that region. Some combination of Mon script and Burmese script, its close descendant, are also the model for Shan script of Myanmar, Ahom script of Northeast India, and old Khamti script of both Myanmar and India.

Elsewhere in Asia, tone is first mentioned as early as 489 CE in Chinese texts (Diller 1996: 232). A comprehensive dictionary from 601 CE was organized into tonal categories (Norman 1988: 24), and strokes indicating historical tone classes were used in early Chinese dictionaries, probably to assist with scholarly etymology in the face of changing phonology, although tone was never marked in the writing system as widely used (Diller 1996: 234).

The Tai languages were the first tonal languages to adopt Brahmic abugidas,[2] giving us the earliest use of phonemic tone marking, indicated as a diacritic above the consonant. This is an important milestone in the written use of language, but tone marking was sporadic in early texts and remained the exception for a few more centuries. In fact, most early Tai scripts had no indication of tone. This includes the majority of the hundreds of surviving inscriptions from early Tai kingdoms. The extinct Southwestern Tai language Ahom of the Brahmaputra valley in modern Northeast India had a rich written tradition for centuries that did not mark tone Morey (2005a). Shan did not indicate tone until less than a century ago. The same is true for Tai Khamti, as seen in Needham's (1894) grammar, which uses native script throughout with no tone marking. This makes the examples that do exist of early tone marking all the more important as a source of textual and archaeological evidence on how tone propagated through the Asian tonbund.

The earliest Tai texts that mark tone show a three-tone system: unmarked, and marked above the consonant with either a short vertical line or a small equilateral cross. These three correspond with columns A, B, and C, respectively, in the Gedney (1972) tone box (see §3.3). This represents the first known instance of phonemic marking of tone in written human language. The evidence from epigraphy is generally taken to indicate that these texts predate the Great Tone Split, in which the number of contrastive tones in Tai languages doubled from three to six.[3]

[2] The Brahmic abugidas are the set of writing systems that derive from Brahmi, a script dating to the 1st century BCE in India. An *abugida* is an alphasyllabary, intermediate between an alphabet and a syllabary, in which the consonant is primary, and vowels are considered diacritics of a consonant. As a result, vowel marks are not necessarily written linearly with their consonant. For example, in Thai script vowel graphemes variably appear before, after, above, or below the consonant they modify, and in some cases two or three of those positions at once.

[3] "The Great Tone Split" is a term coined by Brown (1975). It was a system-wide neutralization of voicing contrasts in onset consonants across much of East and Southeast Asia. It is hard to underestimate its scale, and yet it receives relatively little attention. Brown notes that it dwarfs other more familiar sound changes labeled 'great', like the Great Vowel Shift of English. Indeed, a better name is needed, as its scale is such that affected both tonal and non-tonal language families. While it led to tone splits in languages that were already tonal, it is the source of registrogenesis in many languages that were not, with the classic examples being Mon and Khmer. I introduce the term East Asian Voicing Shift in chapter 5.4 of this work in order to address this need for a better name.

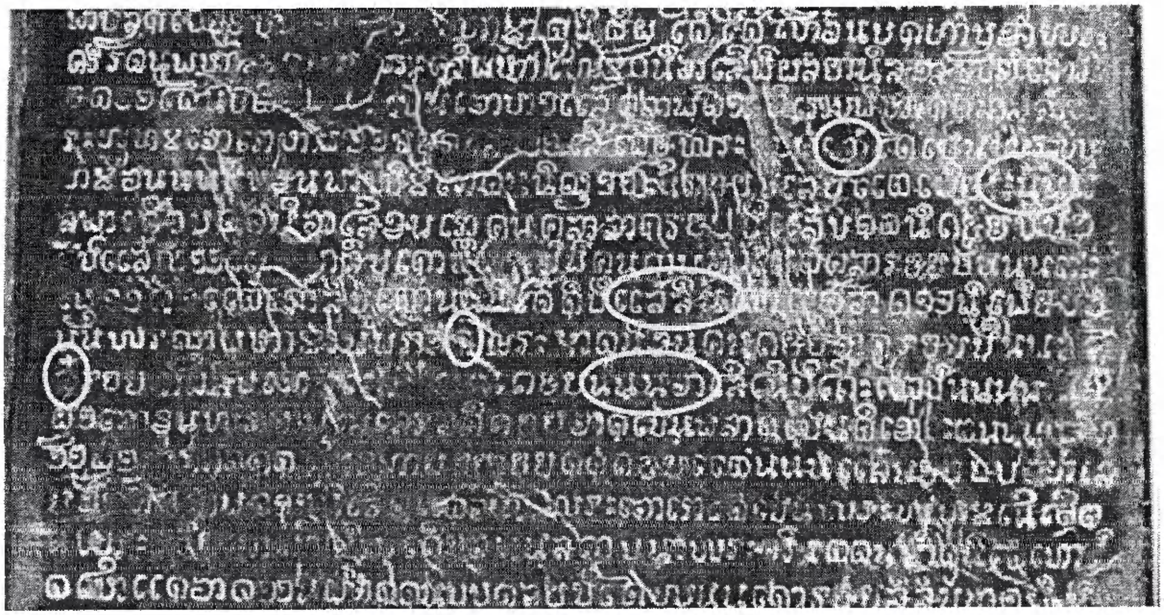


Figure 3.1: Nakhon Chum inscription (1357 CE), Kamphaeng Phet, Thailand. Circles indicate words bearing a phonemic tone mark.

The evidence from epigraphy that Sukhothai was a three-tone language has important implications for tone change and tone reconstruction. It requires us to reconcile this with the fact that the Great Tone Split appears to be exceptionless among the Tai languages, despite members of the family already having spread very far geographically by the time the Tai epigraphic record begins. I deal with this in detail in chapter 5, on the theoretical basis of the Tonal Comparative Method.

Despite the long use of tone marking and a very conservative orthography in Thai, centuries of sound change rendered the connection between tone marks and surface tones opaque long ago. As a result, traditional Thai pedagogy developed a convention for dealing with this no later than the 17th century (Pittayaporn 2016), and which continues to the present. Thai students are taught three consonant classes: high, mid and low. The surface tone of a syllable is determined by a combination of the (1) consonant class of the onset, (2) an optional tone mark, and (3) the syllable shape (termed *live* and *dead*, where dead syllables are those with stop codas, and all overs are live). Although the purpose of the traditional Thai consonant classes is not intentionally historical, as detailed in §3.2.3 the categories directly correspond to rows of the eventual Gedney tone box.

3.2.2 Tone in early Western descriptions

Until the 20th century, tone was not systematically documented in the descriptions of Tai languages written by Westerners. I use the category “Westerner” here to encompass missionaries, explorers, colonizers, and academics, categories which historically overlapped, and sometimes still do. Also, I use “systematic” here to mean an author indicating the tone of each lexical item, as opposed to merely a section describing what the basic tone shapes are in general terms. The earliest European-style grammars of Tai languages were Low (1828), Jones (1842), and Pallegoix (1850), all grammars of Thai; Cushing on Shan (1871; revised and expanded 1887); and Needham (1894) on Tai Khamti. As reported by (Enfield 2008: 7-12), no grammar of Lao appeared in a European language until the 20th century.

Low (1828), for example, focused on explaining the Thai writing system, but includes no romanization in the grammar at all, and thus no tone marking aside from the native orthography. Low is notable for being the oldest surviving example of printed Thai (Smyth 2001: 278, footnote 2). Low describes the sound of the tones in impressionistic terms: “natural,” “acute,” “grave,” etc (Low 1828: 14). He dedicates sections to describing each of the tone marks and its effects on the consonants, but he makes a number of analytical and typographic errors that indicate he did not fully understand the relationship between tone marks and the traditional Thai consonant classes. He remarks on the “inherently intonated power” of many of consonants (1828: 9), and includes a chart of consonants divided by their “inherent tones.”

Cushing also emphasizes the importance of Shan tones, but likewise includes no romanization: “The Shan is a tonal language. Accuracy in speaking it depends on an exact knowledge of the tones and the power of enunciating them. ... The precise extent and limitations of the tones can only be learned from the lips of a native teacher ... an elaborate system is more calculated to mislead the student than to assist him" (Cushing 1871: 8).

Just as with Cushing writing on Shan, in his grammar of Tai Khamti, Needham (1894) was dealing with a script that did not indicate tones. Unlike his predecessors, Needham did include systematic romanization, but did not mark tone. He notes the "finely modulated intonation," and gives an example six different tones that are all spelled the same.[4] "The character is not difficult, but the various tones met with in the language are very puzzling" (1894: ii). Needham managed to produce an otherwise important early grammar without a solid grasp of the tones.

One exception to this general lack of record of tone is the brief sketch of Tai Khamti by Robinson (1849: 342-349), which describes a four-tone system of Tai Khamti, and marks each word in the 282-word lexicon. (Morey (2005b) reconstructed the historical tone classes of the 1849 variety, as discussed in detail in chapter 6.) The early Thai dictionary compiled by Pallegoix (1854) also indicates tone regularly using diacritics reminiscent of modern Vietnamese *Quoc-ngu* script, and quite possibly modeled on that directly.

3.2.3 Modern tone documentation

Conventions for describing lexical tone in Tai languages have evolved along with linguistic understanding of tonal phenomena generally, and of the history of Tai tonal development specifically. (See 5.2 for more detail on past treatment of tonal evidence in historical lin-

[4] Needham appears to have copied portions from Robinson (1849), as this phrase is verbatim from that work, which appeared nearly 50 years earlier but is not cited anywhere by Needham. For comparison, the full contexts are:

> Robinson (1849: 312): "By its finely modulated intonations, sounds organically the same are often made to express totally different ideas. Thus, má, for instance (with the rising tone) signifies a dog; *m*á, (the Italic m denoting the falling tone) signifies to come; while the same syllable, with an abrupt termination, or a sudden cessation of the voice at the end of it, má̦, denotes a horse".
>
> Needham (1894: ii-iii).: "By finely modulated intonation sounds organically the same are often made to express totally different ideas; thus, to give a single illustration there are no less than six words written ၶႆ /khai/, but each one expresses a different meaning according to the tone in which it is uttered, namely, ၶႆ = ill; ၶႆ = sell; ၶႆ = buffalo; ၶႆ = egg; ၶႆ = go, depart; ၶႆ = tell, inform".

guistics.) Up through the end of the 19th century, colorful, impressionistic descriptions of tone were the norm. Bradley lamented the "irrational or even misleading nomenclature" (1911: 282) for Thai tones that had long prevailed. He used "Rousselot's apparatus" (Rousselot 1897) to make the first phonetic study of Thai tones (Bradley 1911).[5] Disagreements on which tone languages were related to which other languages lasted until the mid-20th century, and in some ways are not fully resolved today. It is fair to say that it was only after sufficient quantity and quality of field data on tonal languages accrued, and the basic mechanism of tonogenesis was fully clarified by Haudricourt (1954), that it became clearer what the major language families of Southeast Asia were.

That is not to say that early reconstructions of tone waited for questions of genetic relationship to be resolved, however. Reconstruction of the origin of Tai tones went hand-in-hand with segmental reconstruction (see chapter 5 for a theoretical treatment of this observation). Even before the mechanisms for tonogenesis were completely understood, Chinese philology served as an early jumping off point for Tai tone diachrony. Li recognized the parallels between Tai and Chinese tonal history, as evident in many of his publications (e.g. Li 1945, 1954, 1960). [6], and established the connection between Tai tones and the laryngeal configuration of onset consonants (Li 1943, 1954, 1966), building on earlier observations dating to Karlgren (1915) of the connection between tones and onset consonants in tone languages.

Li (1943) also introduced the labels A, B, C, and D for four Tai tone classes. The first

[5] Jongman (2013) succinctly explains the device Bradley would have used: "Rousselot applied the kymograph to the study of speech. The kymograph, invented in the 1840s by Ludwig, was originally used for measuring blood pressure and other physiological processes. For speech, the kymograph consisted of a rotating drum covered with paper coated with soot; speakers spoke into a rubber tube and the sound vibrations were captured by a stylus that registered the variations in air pressure, from which duration, intensity, and pitch could be measured."

[6] Li believed Tai and Chinese were ultimately part of the same language stock, and apparently never changed his mind throughout his career. This idea is no longer the dominant view, surviving mainly among linguists in China, as well as a minority of Thai scholars (e.g. Sodsongkrit 2009, 2010, 2012). As the idea of a Sino-Tai stock fell out of currency in the mid-20th century, study of tone diachrony predictably splintered more along family lines. See also §5.4 for more on the connections between the tone systems of different Asian language stocks.

three, A, B, and C, correspond to the three tones of Proto-Tai (and also to *unmarked*, *tone mark 1*, and *tone mark 2* in native Thai orthography, as discussed in 3.2.1). The fourth class, D, was for stop-final syllables, characterized by syllable shape rather than onset type. Later, in many Tai languages the D tone split based on the vowel length, usually signified as DL and DS (for long and short).

3.3 The Gedney tone box

As linguists began to uncover the connections between tones and their segmental forebears, the need arose to compactly represent a mapping from modern tonal categories to historical segmental categories. Linguists converged on a box divided into several cells, where each cell represents a subset of the lexicon that patterns together tonally on the surface, and also shares a historical conditioning environment. Due to the regularity of these correspondences, this approach proved extremely successful in Tai linguistics, and has remained the norm several decades, with the standard being the Gedney (1972) tone box, shown in Figure 3.2.

	Proto-Tai tonal categories				
Proto-Tai initials	A	B	C	D-short	D-long
Voiceless friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

Figure 3.2: Tone box for Tai historical analysis, adapted from Gedney (1972)

The Gedney tone box has a theoretical ceiling of 20 tonal categories, each representing a potential historical conditioning environment. This represents a partition of the native lexicon into 20 proper subsets, with each square of the grid standing for a set of words which should share the same lexical tone in any given Tai language. Each cell of the grid can then be populated with the surface tonemes of a given doculect (Good & Cysouw 2013). The tone box concisely shows how the tones of Proto-Tai have coalesced into a given doculect's current tone system. Thus while no single language actually makes even half of the possible tonal distinctions, every possible category on the grid is contrastive in at least one attested Tai doculect. An example of how this looks is given in Figure 3.3, with cells colored with arbitrary colors to highlight the different tonemes.

A	B	C	D-short	D-long
5	2	3	2	2
1	2	3	2	2
1	2	3	2	2
1	3	4	4	3

A	B	C	D-short	D-long
1	3	5	3	1
1	3	5	3	1
2	3	5	3	1
2	4	6	4	4

Figure 3.3: Tone boxes for Standard Thai and Yong (Soiyana 2009)

For each square in the grid, Gedney provided a checklist of cognate etyma that pattern together tonally (1972: 202-204). This modest checklist of 64 words was Gedney's distillation of his complete elicitation questionnaire of more than 1,000 Tai cognates, published in full after his death by Hudak (2004).[7] The number of Tai cognate sets and reconstructions has been expanded by Li (1977), Jonsson (1991), and Luo (1997), Hudak (2008), and others, to encompass thousands of proto-forms with accompanying proto-tones.[8] While there is of course still disagreement in many of the particulars of reconstruction, the fact that Gedney's method has held up so robustly, including elsewhere in Kra-Dai outside

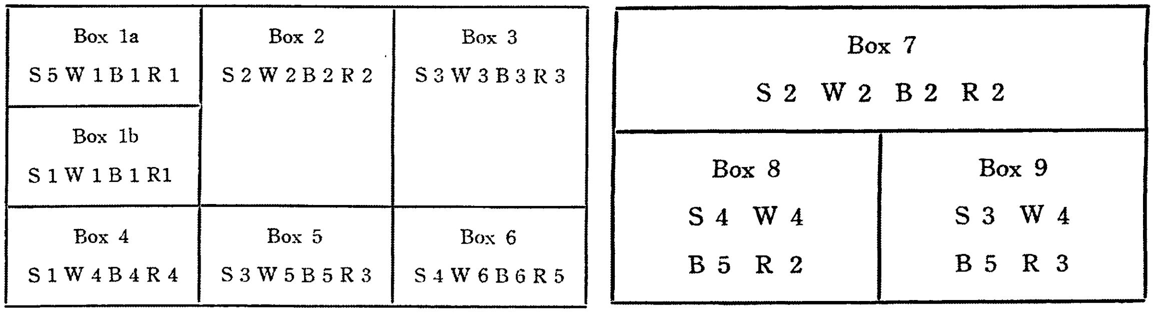
[7] Hudak also reveals much more about Gedney's extremely thorough documentation process, including what his students referred to as 'doing a Gedney', in which every possible syllable and tone combination was tested with native speakers to determine whether it existed as a lexeme in the doculect under study.

[8] Hundreds of these are widely acknowledged as probable Chinese loans, only they are old enough loans that they exhibit completely regular tonal and segmental descent. These ancient loans are also the primary reason behind the remnant factions of linguists, mostly within China, who classify Kra-Dai as a daughter or sister of Sino-Tibetan.

of the Tai branch, serves as confirmation of its continuing usefulness for Tai comparativists working on newly documented varieties, and points to the usefulness of tone for comparative work.

3.3.1 Progression of the tone box

Gedney formulated the tone box between Gedney (1964) and Gedney (1966). In his 1964 paper, Gedney describes in great detail his process for determining tonal correspondences using hundreds of slips of paper sorted into physical boxes, but without the end product of a familiar tone box, which first shows up in 1966. This early effort used data from Siamese, White Tai, Black Tai, and Red Tai (represented by S, W, B, and R, respectively, in Figure 3.4.



Box 1a S 5 W 1 B 1 R 1	Box 2 S 2 W 2 B 2 R 2	Box 3 S 3 W 3 B 3 R 3
Box 1b S 1 W 1 B 1 R1		
Box 4 S 1 W 4 B 4 R 4	Box 5 S 3 W 5 B 5 R 3	Box 6 S 4 W 6 B 6 R 5

Box 7 S 2 W 2 B 2 R 2	
Box 8 S 4 W 4 B 5 R 2	Box 9 S 3 W 4 B 5 R 3

Figure 3.4: Gedney's precursor to the tone box, illustrating how to establish historical tone classes, using data from four languages.

Despite the longstanding dominance of the Gedney tone box, we can observe a progression of increasing complexity, with a growing number of rows and columns, as fieldwork yielded more data, and additional conditioning environments for tonal splits were identified:

- 10 cells: 2 rows x 5 columns (Li 1954)
- 15 cells: 3 rows x 5 columns (Brown 1965)

- 20 cells: 4 rows x 5 columns (Egerod 1961; Gedney 1967, 1972)[9]

- 35 cells: 7 rows x 5 columns (Hanbo 2016)

It may be the case that some parts of this progression in complexity were come to independently, as evidenced by the competing notations for the same ideas. Instead of Li's alphabetic notation A B C D, Egerod (1961) preferred 0, 1, 2, 3 for the main four historical tones, based on the conventional numbering in Thai orthography: Egerod's 0 corresponds to syllables that bear no tone mark, 1 corresponds to the native tone mark *mai ek*, and 2 with the tone mark *mai tho*. Egerod labeled the stop-coda syllables not with a number but with G, 'because the final consonants in question are glottalized in most modern idioms' (Egerod 1961: 45).[10]

Brown (1965) may have followed Egerod or decided on a similar notation himself, but used the order 0 1 3 2 4 in the tone boxes he devised. The numbers 0 1 2 are identical to Egerod's notation, presumably for the same reason, with 3 and 4 correspond to DL and DS, respectively. The reason column 3 is placed before column 2 in Brown's notation is so that 1 and 3 are next to each other, corresponding to B and DL in the Li/Gedney notation. It has long been observed that these two historical categories frequently share the same surface tone.[11]

To make matters even more confusing, the columns known as B and C in Tai tonology correspond to the reversed names C and B in Vietnamese tonology. Court (1998) claims

[9] We might also add a stage of 24 cells, or 4 rows x 6 columns, based on a proposal by Court (1998), although that proposal was never formally published, and as a result is little known and failed to gain traction.

[10] While the preference of Egerod and Brown for a notation based on Thai tone mark numbering is understandable, it introduces confusion in its own way. As conservative as Thai orthography is, the modern spelling does not always reliably indicate either the historical tone or historical onset accurately. Some native words have undergone unetymological respellings, such as /kha:/ 'to kill', where a homophonous onset from a different consonant class has replaced the known historical one—in this case an onset typically reserved solely for Indic loans—resulting in a different tone mark needed to accurately reflect the spoken tone, thus doubly obscuring the word's native origin.

[11] Of the 362 doculects aggregated for study chapter 4.5, the B and DL tones share a surface tone in 249 of those, or 68.7% of the dataset.

Gedney	Egerod	Brown	Chinese
A1	H0	H0	1’
B1	H1	H1	5’
C1	H2	H2	3’
DL1	H:G	H3	9’
DS1	HG	H4	7’
A2	M0	M0	1
A3	G0		
B2	M1	M1	5
B3	G1		
C2	M2	M2	3
C3	G2		
DL2	M:G	M3	9
DL3	G:G		
DS2	MG	M4	7
DS3	GG		
A4	L0	L0	2
B4	L1	L1	6
C4	L2	L2	4
DL4	L:G	L3	10
DS4	LG	L4	8

Table 3.1: Conversion table between the historical tone categorization schemes of Gedney (1972), Egerod (1961), Brown (1965), and linguists in China.

that Li was the source of both conventions, despite their conflicting labels. See Table 3.1 for a conversion chart between the various historical tone class labeling conventions.

3.4 Problems in documentation conventions

3.4.1 The Checked Tone Problem

The Checked Tone Problem is a name I have given to the observation that the same types of tone systems are treated differently in different linguistic traditions. At the very least, this fact should be widely known, as it makes it impossible to have direct comparability of something as simple as the number of tones two languages have, a fact that forms thc basis of many studies (Dryer & Haspelmath 2013; Collins 2016; Everett et al. 2015, inter alia).

A more subtle problem is the fact that synchronic accounts inherit the assumptions about toneme identity that their data sources have, perhaps without realizing it. Consequently, theory built to account for variation will struggle to account for things that are in fact artifacts of differing conventions for documenting tone languages.

One example of this is synchronic accounts of segment-tone interaction (e.g. Yip 2002; Morén & Zsiga 2006). It is the norm in synchronic phonology to take tonemes to be phonologically atomic and composable with every syllable shape, and thus accounts in those frameworks must explain restrictions on tone distribution without recourse to diachronic origin. For example, of the five tones of Standard Thai, only three—low, high, and falling—occur on 'checked' syllables (those with stop codas). This restriction has been referred to as "puzzling" (Yip 2002: 23) and "previously unexplained" (Morén & Zsiga 2006: 116). Morén and Zsiga attribute it to a relationship between a glottal feature and low tones.

In fact the origins of the distribution of Thai tones have been known for many decades, and syllable shape is not a separate feature to tone, but completely fundamental to its origin, as amply demonstrated by the existence of the tone box. Apparent interactions between tone and coda consonants are artifacts of a documentation convention begun by Gedney.

There are two conventions for identifying checked-syllable tones in Asian Tonbund: (1) to assume every checked-syllable tone is an allotone of the most phonetically similar smooth-syllable tone, and (2) to assume that checked-syllable tones and smooth-syllable tones are completely disjoint from one another. I call these the *subset convention* and the *disjunction convention*, respectively.

The subset convention

The term *subset convention* describes the language documentation practice, chiefly used in Mainland Southeast Asia, of treating the tones on smooth syllables as the primary tone inventory, and tones of checked syllables as allotones of a smooth counterpart. The earliest direct statement of the reasoning for this method states:

> Each Tai dialect has made tonal splits conditioned by the phonetic nature of the original initial consonant … the exact pattern differing from one dialect to another. Each dialect has also made coalescences so that each ends up with a total of five, six, or seven tones, on the so-called free syllables, and a smaller number of tonal distinctions on checked syllables, where **complementation permits the analyst to identify tones with the phonetically most closely similar tones of free syllables**
>
> (Gedney 1966; emphasis added)

The following year, Gedney gave additional detail:

> There is **always a much smaller number of permitted tonal distinctions on checked syllables** … Descriptions sometimes differ as to whether the tones occurring on checked syllables are counted as extras or are identified on the basis of phonetic similarity with tones occurring on smooth syllables, with

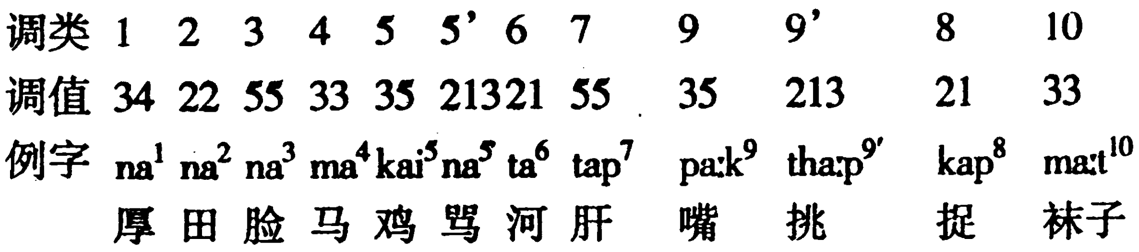
> which they are in **complementary distribution**; the latter practice is followed in numbering tones in the data cited here.
>
> (Gedney 1967; emphasis added)

What constitutes the most phonetically similar smooth-syllable tone is judged by the documenting linguist. While this practice dates to the era where such judgments would have been made strictly through audition by the linguist, it has continued through multiple generations of instrumental advancement. It is unclear when, or if, over the decades of documentation native speaker judgments have been sought as to whether two categories are truly the same phonological category in the grammar of the speaker, and to what extent this is typically a part of speaker awareness, or to what extent learned through pedagogy on tone identity and literacy. This presents a major documentary shortcoming for tonal languages.

Furthermore, Gedney's justification for the convention is faulty. He cites complementary distribution, this does not actually provide a principled way of assigning allotones. But since the two types of tones occur in totally different syllable shapes, every single smooth-syllable tone is in complementary distribution with every single checked-syllable tone. There is no clear justification for assigning a given checked tone as an allotone of a given smooth tone. The practice of assigning them to the most phonetically similar is logical, but we must recognize it as an outdated practice that at times misrepresents both the historical development and the cognitive phonemic reality. This then filters down into synchronic analyses when such work is unaware of the assumptions they may inherit from language documentation conventions.

The disjunction convention

The term *disjunction convention* is another practice predominant in language documentation in and around China. In this convention, checked-syllable tones are treated as disjoint from the smooth-syllable toneme inventory. An example from Zhang (1999), which documents Zhuang languages of the Northern Tai subgroup of the Kra-Dai family, is given in Figure 3.5. There are typically 6 tonemes described for smooth syllables, numbered 1 through 6, and up to four more for checked syllables, number starting from 7 and up through 10. At present I do not know the age of this practice, but it is at least several decades old, if not much older.



调类	1	2	3	4	5	5’	6	7	9	9’	8	10
调值	34	22	55	33	35	213	21	55	35	213	21	33
例字	na^1	na^2	na^3	ma^4	kai^5	na^5	ta^6	tap^7	paːk^9	thaːp$^{9'}$	kap^8	maːt^{10}
	厚	田	脸	马	鸡	骂	河	肝	嘴	挑	捉	袜子

Figure 3.5: Tones of a Zhuang variety (Zhang 1999:128)

The subset convention is not alone in being problematic. Both conventions are too rigid, resulting in tone ‘stuffing’. When strictly followed, will lead to systematic under- or mis-documentation of tonal detail. For the disjunction convention, linguists will under-document novel category mergers, since the convention requires them to have 6 smooth tones, and some additional number of checked tones. For the subset convention, whenever tone changes in checked tones, independent of the smooth tones, the linguist is still forced to ‘stuff’ checked tones into some smooth tone, even if there is no evidence for the cognitive phonemic reality of their unity as a tonal category.

3.4.2 The Orphaned Tone Problem

The Orphaned Tone Problem is an issue that arises within the subset convention. Data from Sopheap (2017) provides a clear-cut illustration of how tone change in one type of syllable, independent of the other syllable type, causes the probably miscategorization of allotones.

Sopheap (2017) is a study of Lao dialects spoken in Cambodia near the Cambodia-Laos border, a previously unstudied area. The author gathered data at 12 villages in three provinces, and presents tone boxes for them. Sopheap analyzes them by (1) total number of tonemes, 5, 6, or 7, and (2) splits and mergers in the A and B columns.

All 12 lects studied exhibit identical splitting of the C, DL, and DS columns: C1 = DL123, C234 = DL4 (the classic 'yin-yang' shape of these columns often associated with various Lao dialects (Akharawatthanakun 2003: 1). However, the symbolic tone numbers associated with these identical splits vary substantially depending on historical activity in the A and B columns. Consider the tone boxes of the 12 lects, given in Figure 3.6.

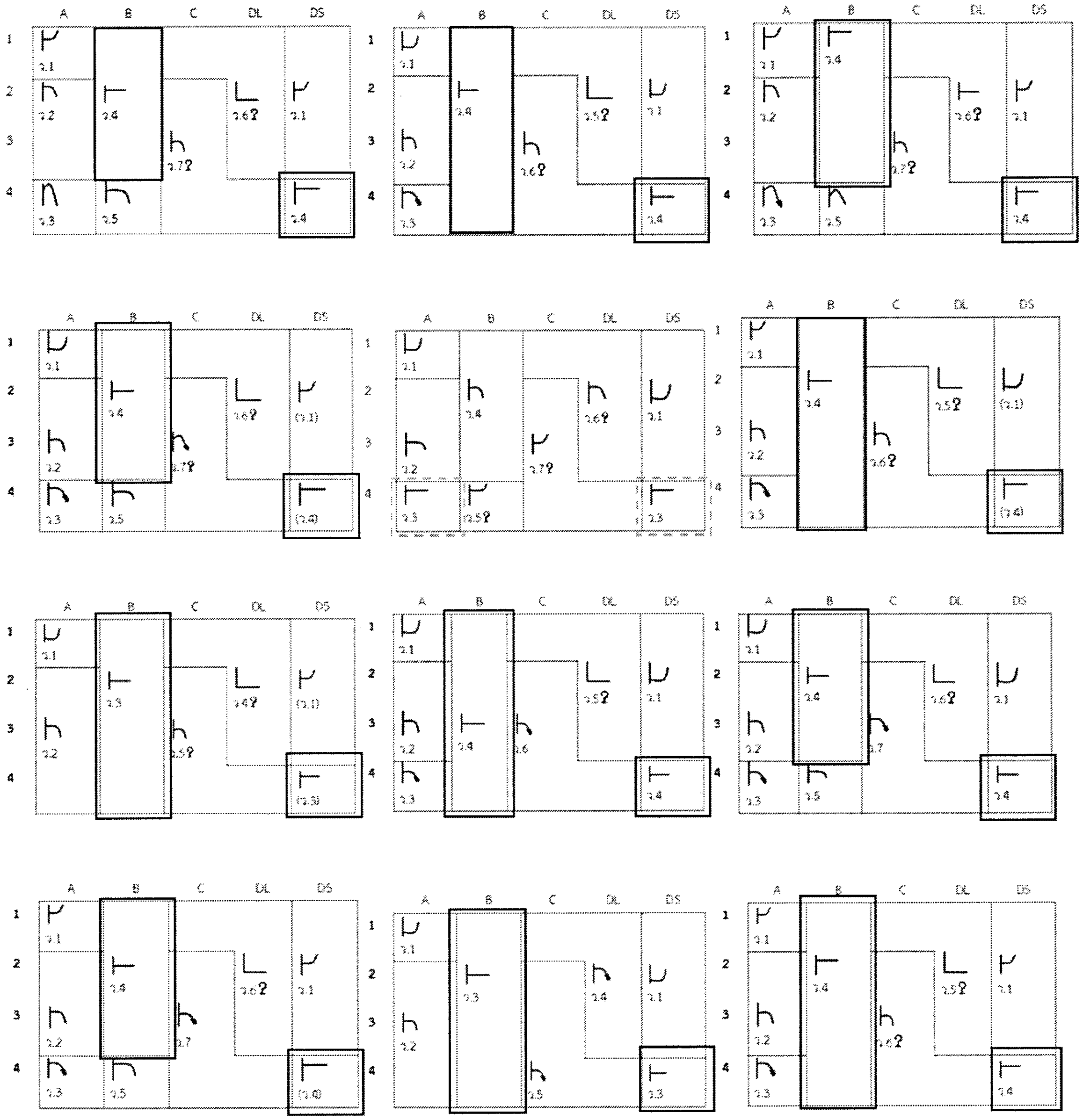


Figure 3.6: Gedney boxes from 12 lects in Sopheap (2017))

Cell DS4 of the tone box, the subset of the lexicon descended from checked syllables with a short vowel, is an allotone of tone 4 in 11 of the 12 lects. In the 12th lect, however, it is an allotone of tone 3. And yet in all 12 lects, the phonetic value of the surface tone in the DS4 cell is identical. The far more likely explanation than variable allotony, is that in one of the 12 lects, tone 4 simply changed its phonetic shape due to regular sound change, and thus 'orphaning' the tone it was previously most similar to. This alone is evidence for the non-allotonic relationship between the categories, because they can change independently. The linguist, bound the by subset convention, rather than label this as a seventh tone, chose to call it an allotone of a different toneme instead, despite the same linguist documenting all 12 lects.

Despite the phonetic variation between these closely related and closely situated doculects, there is no principled phonetic or historical reason to hypothesize that the DS4 toneme category has ever changed. It has the same historical conditioning environment across all lects, it is uniformly level across the 12 doculects, and inter-lectal variation varies at most one step on the Chao (1930) five-level pitch scale (44 vs. 33).

How widespread this type of issue might be is impossible to say. However, given that the two syllable shapes represent different conditioning environments, it is logical and expected that their tones must be able to change independently of one another, unless the cognitive reality of a toneme that crosses syllable shape categories in a given lect is established through perception and production studies. Such studies would be very welcome, and may even serve to refute some portions of this argument. As it stands, however, until those happen, this is a basic descriptive problem for tone documentation in Southeast Asia.

Chapter 4

Phylogenetic signal

4.1 Introduction

It should be reasonably intuitive to linguists that segmental phonology could contain useful information for determining the historical relatedness of languages. The longstanding historical Comparative Method (Weiss 2014), in use for two centuries, involves comparing lists of common words to identify cognate sets. Using these cognates we infer the version of past events that best explains our current data: we posit sound changes, classify languages into family trees, and reconstruct proto-sounds and proto-languages, all using primarily segmental phonological data. That said, key to the method is to look especially for *regular* sound correspondences, in order to make reliable inferences. Thus, while the manual method clearly relies on segmental data, it may not be intuitive to linguists that quantitative methods can replicate this process in a useful way when the input to these methods is automatically extracted from raw lexical data. Phonology has typically been the means to the end, the internalized knowledge of the linguist, which enables traditional analysis and coding for lexical cognacy in a dataset. Once coded for cognacy, however, the particular sounds involved in sound changes processes, indeed the phonology entirely, no longer factor directly into the analysis.

Less intuitive to linguists has been the idea that modern tone systems encode useful information about the past. Methods developed for segmental sound change do not directly transfer to tone systems because tones change differently from segments in ways not yet fully understood, thus making the appropriate object of comparison obscure. Combined with the fact that tone is a cross-family areal phenomenon in multiple parts of the world, some skepticism is appropriate.

In this chapter I present results from a set of computational phylogenetic tests which lay the empirical groundwork for the theory of the Tonal Comparative Method detailed in chapter 5. Specifically, I use tests for *phylogenetic signal* to show that phonological traits are inherited from ancestor languages, providing statistical confirmation of this key tenet of the traditional Comparative Method. Further, I show that not only is there strong phylogenetic signal in the segmental evidence, as we would expect given the widespread success of the CM, but that the tonal domain likewise contains ample phylogenetic signal. These findings strongly support the idea that contrary to some past claims and conventions in historical linguistics, tonal evidence is generally suitable for use in historical-comparative tasks like reconstruction and classification, with the caveat that tone must not be taken simply as modern surface categories, but rather linked back to the segmental origins that conditioned tone change. This chapter demonstrates that we can fruitfully use several phonological properties that are automatically derivable from lexical datasets—including phone and biphone distribution, phone and biphone frequency, and tone splits and mergers—in quantitative tasks such as tree inference and ancestral state reconstruction.

In the remainder of this chapter, §4.2 presents background on the types of traits that have been used in previous work in linguistic phylogenetics, while §4.3 describes tests for phylogenetic signal in linguistic datasets. In the following two sections I then present the results of two studies of phylogenetic signal in phonology: first in segmental properties of Tai lexicons in §4.4, followed by §4.5 with results of for traits extracted from Tai tone boxes (see §1.3 and §3.3 for background on tone boxes). Finally, I conclude the chapter in §4.6 by discussing the implications of these results for both the Comparative Method generally and for the Tonal Comparative Method specifically.

4.2 Phylogenetic signal in linguistic datasets

Mesoudi (2011: 26) summarizes Darwin's theory of evolution as comprising three preconditions: *variation*, *competition*, and *inheritance*. Over the course of that book, he goes through some of the "voluminous evidence" (2011: 32) showing that cultural traits evolve under the same three preconditions, although by quite different mechanisms. For a linguistic trait to exhibit evolutionary development, it must vary in its expression between individuals in the population, there must be selection for some linguistic variant over others due to competition, and the trait must be heritable from one generation to the next. Language is an area where we strongly see these traits, and thus we have evolution. We can identify the three preconditions of evolution in many components of human language, and thus scholars interested in evolutionary linguistics have applied these methods to many kinds of linguistic data.

While much work in linguistic phylogenetics has focused on lexical data, which is amply available for a large portion of the world's languages, linguists have shown consistent interest in applying the tools and techniques to many other data types as well. However, just as in evolutionary biology and the many other fields that have begun to work within an evolutionary framework, linguists must learn to use the available tools appropriately to produce valid results. Just as the Comparative Method requires identifying regular patterns of sound change and ignoring loan words and other confounds, it is critical in evolutionary linguistics to know that our conclusions are based on *homology*, or similarity due to divergent evolution from a shared ancestor, and avoid being confounded by *homoplasy*, or convergent evolution from different ancestors (Mesoudi 2011: 88). Using any linguistic dataset for historical analysis requires careful differentiation between similarity due to shared inheritance, and similarity due to chance resemblance.

To solve this problem, evolutionary biologists develop statistical tests to assess the degree of *phylogenetic signal* that a particular trait or set of traits has. Phylogenetic signal can be defined informally as the degree of similarity that is due to having a common ancestor. Slightly more technically, phylogenetic signal is a measure of the statistical dependency between traits in a dataset that is due to phylogenetic relationships (Revell et al. 2008). Just as with biological data, these tests must be applied to linguistic data to ensure that quantitative methods are producing reliable results. Different types of tests are available for different types of data, and ideally multiple tests should be used to ensure an accurate picture of the degree of signal. The tests used in the two studies in this chapter are described in the next section, §4.3.

At this point, a brief overview of the major categories of linguistic data used in linguistic phylogenetics, and the studies that employ them, is helpful in order to give background context to the results that are discussed in §4.4 and §4.5. Not all studies use one particular category of traits exclusively, but that has been the predominant trend. One reason for using only a single category of traits is that it facilitates larger amounts of comparable data, but whether this consistently produces the best results is questionable.

4.2.1 Lexical traits

At present, it is still true to say that the majority of computational phylogenetic studies in linguistics are done using lexical data, in the form of cognate sets coded as binary traits. One reason for using this type of data that is typically given is that it mirrors the use of cognate sets in the Comparative Method. This parallel is shallow, however, as the approach relies on cognate presence or absence, rather than the regular sound correspondences of the CM.

The most common goal of the lexical approach has been language classification, an activity also variously known as subgrouping or cladistics. This typically takes the form of tree or network construction. Relatively ample lexical data is available for a large number of the world's languages, but time-consuming cognate coding is still required, presenting a serious labor bottleneck.

Further, cognate coding is also susceptible to a potentially high degree of uncertainty of cognate judgments, simply because linguists regularly disagree about cognacy. Purpose-built cognate coding for a phylogenetic study may rely on plausible superficial similarity, and thus fail to filter out intra-family borrowing that may be caught by more careful manual checking for sound correspondences before judging cognacy. Practices can certainly be expected to improve iteratively, and software tools to aid in cognate detection are one major way that will happen.

A large and growing number of language families have received attention with the lexical approach, including Arawakan (Walker & Ribeiro 2011; Stark 2018), Aslian (Dunn et al. 2011a), Austroasiatic (Sidwell 2014), Austronesian (Gray & Jordan 2000; Gray et al. 2009; Greenhill et al. 2009), Bantu (Holden & Gray 2006), Chapacuran (Birchall et al. 2016), Central Solomons Papuan (Dunn & Terrill 2012), Dravidian (Kolipakam et al. 2018), Indo-European (e.g. Taylor et al. 1995; Rexová et al. 2003; Pagel 2009; Bouckaert et al. 2012; Chang et al. 2015), Japonic (Lee & Hasegawa 2011), Pama-Nyungan (Bowern & Atkinson 2012; Bouckaert et al. 2018), Semitic (Kitchen et al. 2009), Sino-Tibetan (Sagart et al. 2019; Zhang et al. 2019), Tasmanian (Bowern 2012), Tupian (Michael et al. 2015a), Uralic (Syrjänen et al. 2013; Honkola et al. 2013), and Uto-Aztecan (Dunn et al. 2011b; Wheeler & Whiteley 2015).

4.2.2 Syntactic traits

Syntactic data has been used to study Indo-European (e.g. Longobardi & Guardiano 2009; Longobardi et al. 2013). Dunn et al. (2011b) used trees inferred from lexical data to test syntactic word-order universals in Austronesian, Bantu, Indo-European, and Uto-Aztecan. Bowern (2018) explains the reasons we might expect syntactic traits to perform poorly in phylogenetic tasks, due to parametric traits having a small number of typological possibilities in human language. This limited set of possibilities makes parallel innovation both extremely common and difficult to distinguish from shared innovation. Problematic binary traits might include presence/absence of particular noun cases, constituent ordering, or passivization strategies (Bowern 2018: 288), all of which are parametric syntactic traits that we would expect to be among the easiest to aggregate and thus most likely to be used for quantitative studies.

4.2.3 Typological traits

The case for the use of typological traits, a mixture of features from various linguistic categories, is made in Wichmann & Saunders (2007). In particular, they argue that typological traits provide the only path for phylogenetic analysis of more ancient language families, where cognacy cannot be reliably determined, and thus the Comparative Method alone does not provide traction 2007: 378. Among the studies that have used typological traits include Sicoli & Holton (2014), for Dené–Yeniseian.

4.2.4 Phonological traits

A small number of studies have used phonological data to create a trait set for quantitative subgrouping. These include projects on languages from Pama-Nyungan (Macklin-Cordes 2015; Macklin-Cordes & Round 2015), Tukanoan (Chacon & List 2015), and Turkic (Hruschka et al. 2015). For Tukanoan, Chacon & List (2015) coded shared innovations, sound changes previously identified using the CM, as binary traits. For Turkic, Hruschka et al. (2015) began with lexical cognate sets and coded them as segmental sequences, making the comparison to gene sequencing, in order to identify regular correspondences, and thus sound changes. Macklin-Cordes (2015) used a series of tests for phylogenetic signal on Ngumpin-Yapa languages (from the Pama-Nyungan family) to show that fine-grained phonotactic information encodes useful historical signal without the need for manually coding cognates. (Study A in §4.4 follows a similar approach, and extends its findings.)

A different approach that might be classed under the phonological category is the Automated Similarity Judgment Program (ASJP; Brown et al. (2008)). The ASJP framework has been used for both language classification, as in Brown et al. (2008), and dating of language families, as in Holman et al. (2011). Unlike the studies mentioned above, cognate coding is not used. The ASJP data consists of a set of basic vocabulary from the supposedly most stable concepts, aggregated from as many languages as possible. The original study reported results using 245 (Brown et al. 2008) languages, but that grew to 4,817 in Holman et al. (2011), and as of this writing currently at 7,655 (Wichmann et al. 2018). ASJP is also notable for its small number of basic vocabulary items sampled per language, originally 100 but reduced to 40 after tests showed no difference in accuracy (Holman et al. 2008).

4.3 Methods

This section provides brief descriptions of the methods for measuring phylogenetic signal that are used in this study: the *D statistic*, for binary data, and *Blomberg's K*, for continuous data. Also used is *NeighborNet*, a common (and commonly misunderstood) tool for measuring the degree of connectedness in a dataset. A NeighborNet analysis yields two additional measures of treelike signal: the *delta-score* and the *Q-residual.*

4.3.1 *D* statistic

First developed by Fritz & Purvis (2010), this test measures phylogenetic signal in binary traits. The formula for the *D* statistic is as follows (Fritz & Purvis 2010: 1044):

$$D = [\Sigma d_{obs} - \text{mean}(\Sigma d_b)]/[\text{mean}(\Sigma d_r) - \text{mean}(\Sigma d_b)]$$

In this formula, Σd_{obs} is the observed sum of differences between sister node values, Σd_b is the expected distribution of sums under Brownian evolution, and Σd_r is the expected distribution of sums for randomly shuffled distribution of the trait. Thus, D is calculated by summing the observed differences and subtracting the mean expected Brownian distribution from that. This is then divided by the difference between the mean random distribution and the mean Brownian distribution. The resulting D statistic is tested for statistical significance against two null hypotheses: that of the expected distribution under Brownian evolution ($D = 0$), and that of random distribution of the trait ($D = 1$). And while 0 and 1 are the scores of the two null hypotheses, these are not the bounds of the D statistic. If a trait across a phylogeny is clumped even more conservatively than the Brownian expectation, the value will be less then 0, and traits distributed even more evenly than the random expectation will exceed 1. The number of permutations used to calculate both the Brownian and random expectations is also configurable, with 1000 permutations being the recommended by the authors (Fritz & Purvis 2010: 1045).

4.3.2 Blomberg's *K*

While both D statistic and NeighborNet test binary data, the K statistic is a test for phylogenetic signal in continuous data proposed by Blomberg et al. (2003), based on variances of *phylogenetically independent contrasts* (PIC). The PICs for a given trait are calculated by pairwise comparison of all values of the phylogeny tips for a given trait. The contrast of two tips is divided by the square root of the branch length distance that separates those tips (Felsenstein 1985: 8). These are then compared against the expected distribution of the trait under a Brownian model of evolution.

4.3.3 NeighborNet

A NeighborNet analysis (Bryant & Moulton 2004) is a graph of the connectedness of data in the dataset, though it does not distinguish between horizontal (contact) and vertical (genetic) signal. It does allow us to derive two additional statistics, however: the delta-score, introduced by Holland et al. (2002), and the mean Q-residual (Gray et al. 2010).

These tell us about how "treelike" a given dataset is, and are calculated by considering all of the taxa in a dataset in groups of four. Each combination of four taxa is considered by its possible pair-wise combinations (i.e. with four taxa a, b, c and d, we can combine them as ab + cd, ac + bd, and ad + bc). The distances between the taxa is then summed according to these combinations and ordered largest to smallest. If we label these summed combinations as $\Sigma 1$, $\Sigma 2$, and $\Sigma 3$, where $\Sigma 1$ represents the largest sum and $\Sigma 3$ the smallest, then we can express the delta-score and the *Q*-residual as follows (Gray et al. 2010: 3926):

delta-score: $(\Sigma 1 - \Sigma 2)/(\Sigma 1 - \Sigma 3)$

Q-residual: $(\Sigma 1 - \Sigma 2)2$

If the data is perfectly tree-like, both statistics will be equal to zero. Gray et al. found that *Q*-residual obscures less of the signal than the delta-score (2010: 3926).

4.4 Study A: Phylogenetic signal in segmental phonology

The first of the two studies examines whether features of phoneme inventories and phonotactic profiles that are automatically extractable from a set of lexicons themselves contain phylogenetic signal, without the need for being organized into cognate sets by a linguist. This study adds to the small but growing body of work on the use of phonological traits in computational phylogenetics for linguistics, two recent examples being Macklin-Cordes (2015) and Macklin-Cordes & Round (2015). These two works explore phylogenetic signal contained in the phonotactics of the Ngumpin-Yapa languages, a 10-language subgroup of the Pama-Nyungan language family, spoken in the Pilbara region of Western Australia. The present study confirms and extends the findings of that work, using a set of data from 20 lects of the Tai branch of the Kra-Dai language family. The confirmed finding is a strong phylogenetic signal in the more high-resolution continuous traits drawn from phoneme frequency and biphone transition probabilities. The additional novel finding from this study is that relatively strong phylogenetic signal exists in even the more coarse-grained binary traits of phoneme presence/absence and biphone presence/absence, which previous work was unable to do.

Study A uses the three statistical tests for phylogenetic signal described in §4.3: *D* statistic, Blomberg's *K*, and NeighborNet.

4.4.1 Data

The data for Study A comes from Hudak (2008). Compiled by Hudak from the extensive mid-20th century fieldwork of William J. Gedney, this source consists of posited 1,159 cognate sets covering 19 languages from the Tai subgroup of the Kra-Dai family.[1] One of the languages, Saek [skb], is also subdivided in the dataset into younger generation Saek and older generation Saek, for a total of 20 lects. (See Table 4.1 for list of lect names and their ISO 639-3 codes.)

The total Tai dataset from Hudak is 14,609 lexical items, giving an average lexicon size of about 750 items. The fact that the dataset consists entirely of posited cognate sets, as opposed to raw lexicons, could be argued to be sampling bias that will predispose the dataset to a positive result. However, given that the methods under investigation are still in need of validation using phonological traits to begin with, a suitably noise-free dataset gives the best chance at detecting phylogenetic signal. This dataset essentially puts the tests used in this study on equal footing with a linguist using the traditional Comparative Method: building a historical analysis around a group of cognate sets. And indeed, if signal cannot be detected in this subset of the lexicon, then it is quite unlikely that the results would prove fruitful on full lexicons, either.

Data that can be extracted from the raw lexical material falls into two broad types:

[1] The definition of 'cognate' used by Gedney and Hudak requires remark: they have included all forms believed to be modern reflexes of a particular form in Proto-Tai, and not just modern forms that coincide precisely in modern lexical meaning. This usage aligns with what Michael et al. (2015b) term 'quasi-cognates', who argue that the norm in linguistic phylogenetics has been to require identical semantics. Since semantic shift and sound change can occur independently, this definition rules out many historically valid cognates. Somewhat confusingly, what Michael et al. call 'quasi-cognates' is precisely how historical linguists have long used the term 'cognate'.

binary data on the presence or absence of phonological traits in the languages, and continuous data on the distribution or probability of those traits in their languages. Each of those two categories is further divided into phoneme data, which looks at individual segments, and biphone data, which looks at how those segments combine. These traits were extracted from lexicons using Python scripts that I modified from ones developed for Gasser & Bowern (2014), in conjunction with my own Python and R scripts.[2]

Lect	ISO code	Lect	ISO code
Black Tai	[blt]	Tai of Lungchow	[zzj]
Lao dialect of Nong Khai	[tts]	Tai of Lungming	[zzj]
Lue of Chieng Hung	[khb]	Tai of Ning Ming	[zzj]
Lue of Muong Yong	[khb]	Tai of Piang Siang	[zzj]
Saek (Old Generation)	[skb]	Tai of Po-ai	[zgn]
Saek (Young Generation)	[skb]	Tai of Western Nung	[nut]
Shan	[shn]	Tai of Wuming	[zyb]
Tai of Bac Va	[nut]	Thai	[tha]
Tai of Chiang Mai	[nod]	White Tai	[twh]
Tai of Lei Ping	[zzj]	Yay	[pcc]

Table 4.1: Tai lects used in this study, from Hudak (2008), and their ISO codes.

Binary data

In order to test just how fine-grained the phonological traits need to be in order to detect phylogenetic signal, two type of binary data were extracted from the Tai lexicons. The first is the presence or absence of each phoneme within each language. Thus all phonemes attested anywhere in some lexicon form a set of traits, and then each phoneme is coded as a one or a zero, present or absent respectively, for each language. The hypothesis underlying testing this data type is fairly intuitive: the more similar the phoneme inventories of two languages are, the more closely related they are expected to be. Of the 54 phonemes in the Tai data, 13 phonemes were present in every lect, while the other 41

[2] Thanks to Aidan Kaplan, who wrote the original version of these scripts as part of a course final project.

showed varying degrees of variation. When necessary for the nature of one of the phylogenetic tests, phonemes found in every language were pruned from the dataset, as there is no phylogenetic signal in a trait that exhibits no variation (since it fails one of the three basic preconditions). The binary phoneme data is summarized in Table 4.2.

Lects	Phonemes	
	Total	w/Variation
20	54	41

Table 4.2: Variation in phoneme data (binary).

The second type of binary data is on the presence or absence of biphones found in a given language, i.e. whether two segments appear next to each other in a particular sequence. While the binary phoneme data simply tells us whether, say, /b/ appears in a lexicon at all, the biphone data tells us whether or not /b/ appears next to each other possible segment, as well as whether it appears word-initially or word-finally. This is thus a rudimentary representation of the phonotactics of the language, and the hypothesis underlying it is similarly intuitive to that of the binary phoneme data: the more close related two languages are, we would expect them to not only have similar phoneme inventories, but for the distribution of those phonemes to be more similar among more closely related languages. Binary biphone data was only generated for attested biphones, as the probability of any unattested biphone is always zero. This data is summarized in Table 4.3.

Lects	Biphones	
	Total	w/Variation
20	555	526

Table 4.3: Variation in biphone data (binary).

Continuous data

In addition to the coarse-grained binary data, two types of more fine-grained continuous data were also extracted from the lexicons. The first type of continuous data is frequency data for the phonemes found in each language. The intuition underlying this type of data is that two identical languages would share both the same phonology and the same lexicon, and thus identical phoneme frequency. Therefore, the closer two languages are in both phonemes and the distribution of those phonemes across their lexicons, the closer those languages are likely to be to each other genetically. Distribution of a phoneme across the dataset would be expected to be more phylogenetically informative than simple binary data, since two languages may share some phoneme, but that phoneme may be a core phoneme with high functional load in one language and low-frequency or marginal in the other language.

Phoneme frequencies are calculated here language by language. As discussed in Gasser & Bowern (2014), there are two methods to measure phoneme frequency in a lexicon: (a) the quotient of the occurrences of a phoneme in a language and the total number of segments in that language, and (b) the quotient of the number of lexical items that a phoneme occurs in and the total number of lexical items in that language. Given the variable length of items in the lexicon and the potential for multiple instances of a phoneme within a lexical item, the first method is used here.

The second type of continuous data is biphone probabilities. As in prior work by Macklin-Cordes (2015: 34-35), these are modeled as Markov chain transition probabilities. A Markov chain model is a matrix of transition probabilities between each possible state, which sum to 1. For the purposes of modeling biphone phonotactics, we only need to be concerned with one-step transition probabilities. That is to say, we are concerned only with the probability of a transition to some phoneme, given the current phoneme. This gives us the following formula:

$P_{ij} = P(x_{n+1} = i \,|\, x_n = j)$

The probability P_{ij} represents the probability that the process will make a transition to state *i*, given that currently the process is in state *j*.

(Ching & Ng 2006: 3-5)

Markov chain transition probabilities provide a more robust representation of the phonotactics of a language than either phoneme probability, which does not consider the environment in which phonemes appear, or simple binary presence/absence information about the transitions, which does not take into account their likelihood of appearing in a given lexicon. Thus, the hypothesis for this type of data is that the closer two languages are in sharing a profile of Markov chain transition probabilities, the genetically closer those two languages are.

4.4.2 Results

D test

Study A calculated *D* statistics using a modified version of the `phylo.d` function of the R package `caper` (Orme et al. 2012), with 10,000 permutations, and a traditional tree of Tai lects adapted from Hudak (2008) (see Figure 4.5). *D* statistics were calculated for two types of binary Tai data: phonemes and biphones. The density plot of *D* values for binary phoneme data is presented in Figure 4.1.

On binary phoneme data, Macklin-Cordes writes that it is "unsurprising" that a list of present and absent segments is not phylogenetically informative, given the homogeneity of the segmental inventories in Ngumpin-Yapa languages. Only three segments showed any variation, and thus were testable, to begin with 2015: 69. His *D* test results are given in Table 4.4.

The difference between the Tai data and the Ngumpin-Yapa data is striking. All three *D* scores from Ngumpin-Yapa are well above 1, the threshold indicating clustering charac-

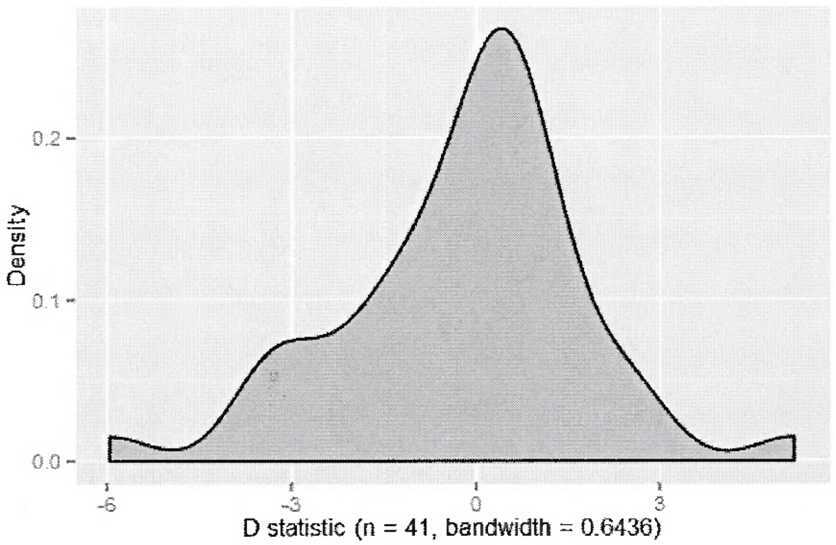


Figure 4.1: Density plot of *D* values for Tai phonemes (binary).

	D	$p_{(D=0)}$	$p_{(D=1)}$
iː	2.659	0.113	.735
uː	2.218	0.106	0.734
rR˜	1.994	0.465	0.398
Mean *D*	2.29		
SD	1.98		

Table 4.4: Ngumpin-Yapa phoneme *D* scores and *p* values (Macklin-Cordes 2015: 69).

teristic of random distribution, yielding a mean *D* of 2.29. Contrast this with the *D* scores of the 41 varying phonemes in the Tai data, given in Table 4.5.[3]

As Table 4.5 shows, approximately half of the phonemes have *D* statistics below zero, indicating strong signal. A minority are near or above 1. Even the mean *D* score for the Tai data is also negative, at -0.119, though with a large standard deviation. The dataset on average indicates slightly stronger phylogenetic signal even than the null hypothesis that the trait distribution is the result of Brownian evolution. This shows that there is a strong phylogenetic signal in the Tai binary phoneme data alone, the most coarse-grained of the four data types under investigation in this study.

[3] All *D* statistics presented in this chapter have been rounded to the third decimal place.

	D	$p_{(D=0)}$	$p_{(D=1)}$
ɣ	-6.027	0	0.950
ʔy	-3.368	0.103	0.663
ʔd	-3.349	0.099	0.671
ɤ	-3.130	0	0.951
ʔb	-3.019	0.097	0.659
θ	-2.385	0.010	0.858
th	-2.311	0.010	0.856
kh	-1.601	0	0.966
ă	-1.546	0.026	0.781
r	-1.327	0.004	0.859
ɔ	-1.226	0.002	0.861
ph	-1.133	0.006	0.861
ɬ	-0.734	0.002	0.838
ɰ	-0.721	0.002	0.848
ɯ	-0.356	0.041	0.651
u	-0.309	0.038	0.643
f	-0.240	0.065	0.579
x	-0.234	0.017	0.703
v	-0.105	0.045	0.618
ʔw	0.185	0.118	0.543
ɛ	0.189	0.046	0.499
š	0.373	0.152	0.399
i:	0.392	0.170	0.382
e:	0.397	0.166	0.384
u:	0.402	0.169	0.374
s	0.504	0.141	0.472
ch	0.505	0.141	0.358
ɔ:	0.549	0.132	0.326
ɲ	0.637	0.145	0.304
ɛ:	0.734	0.190	0.267
ɤ:	0.777	0.220	0.241
o:	0.986	0.428	0.151
ɯ:	0.998	0.425	0.149
b	1.087	0.407	0.181
e	1.420	0.497	0.246
d	1.496	0.616	0.093
c	1.518	0.749	0.042
k	2.461	0.350	0.229
o	2.475	0.353	0.229
ð	2.507	0.352	0.230
sh	4.944	0.705	0
Mean D	-0.119		
SD	1.98		

Table 4.5: Tai phoneme D statistics and the p values of the two null hypotheses, in descending order by D.

The way to interpret the p values of the null hypotheses in the table is to take them as the result of testing of whether the score is significantly different from the null hypothesis. So, in the first row of Table 4.5, the voiced velar fricative has a D statistic of -6.027, which is not significantly different from 0 ($p = 0$), the null hypothesis of Brownian evolution, and is significantly different from 1 ($p = 0.950$), the null hypothesis of a randomly distributed trait.

Now let us compare the above results with the D statistic for the Tai biphone transitions. A density plot of these values is given in Figure 4.2.

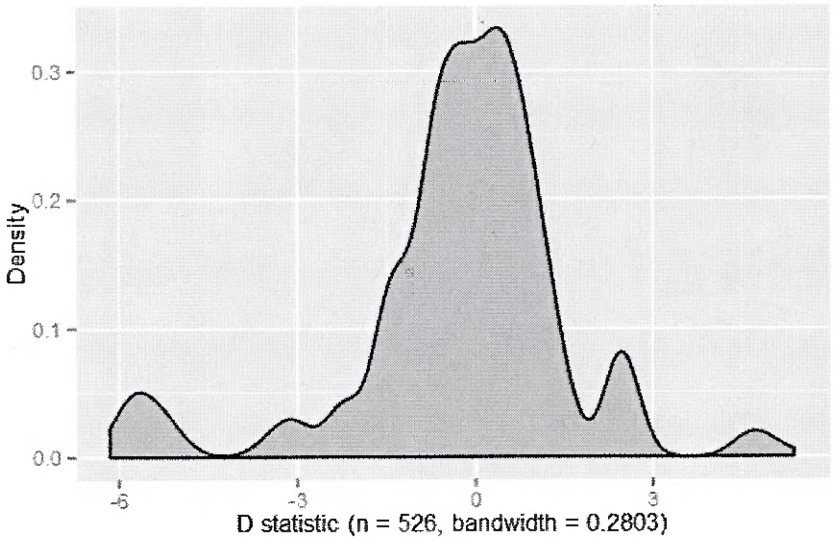


Figure 4.2: Density plot of D values for Tai biphones (binary).

Consider this against a comparable plot for Ngumpin-Yapa from (Macklin-Cordes 2015: 80) in Figure 4.3.

Macklin-Cordes describes this plot as irregular with heavy outliers. The mean D for the biphone dataset was 0.79 with standard deviation of 2.8. He concludes that the binary dataset has no significant level of detectable genetic signal (2015:79).

To provide a more granular look at the data from the Tai side, the twenty largest and smallest D scores, along with the mean and SD, are given in Table 4.6.

From Table 4.6 we can again see that as with the binary phoneme data, the binary

	D	$p_{(D=0)}$	$p_{(D=1)}$
rɤ:	-6.155	0	0.952
hă	-5.977	0	0.953
ɣa:	-5.956	0	0.948
re:	-5.926	0	0.949
ɛl	-5.919	0	0.833
wɤ:	-5.842	0	0.952
rɯ:	-5.816	0	0.947
#ɣ	-5.812	0	0.949
u:l	-5.675	0	0.837
ă#	-5.672	0	0.948
ɣɔ:	-5.654	0	0.947
ɲi:	-5.636	0	0.947
xe:	-5.594	0.007	0.949
a:l	-5.579	0	0.831
xo:	-5.566	0.007	0.952
xi:	-5.546	0.006	0.949
il	-5.495	0	0.833
ɔ:l	-5.399	0	0.834
al	-5.387	0	0.836
sɤ:	-5.379	0.006	0.947
…	…	…	…
ɤɰ	2.542	0.350	0.233
ʔa:	2.553	0.356	0.233
iŋ	2.560	0.350	0.236
at	2.568	0.354	0.233
ya:	2.584	0.354	0.227
#y	2.586	0.349	0.235
#ð	2.623	0.353	0.229
pa	2.683	0.362	0.228
#t	2.695	0.361	0.231
#k	2.743	0.360	0.231
oʔ	4.466	0.696	0
aʔ	4.554	0.702	0
ɤp	4.572	0.695	0
eʔ	4.638	0.694	0
#sh	4.673	0.702	0
shɯ	4.815	0.699	0
shɤ	4.876	0.700	0
shi	4.899	0.705	0
ɤʔ	4.905	0.700	0
bɤ	4.934	0.702	0
Mean D	-0.239		
SD	1.86		

Table 4.6: Tai phoneme D statistics and p values of the two null hypotheses, in descending order by D.

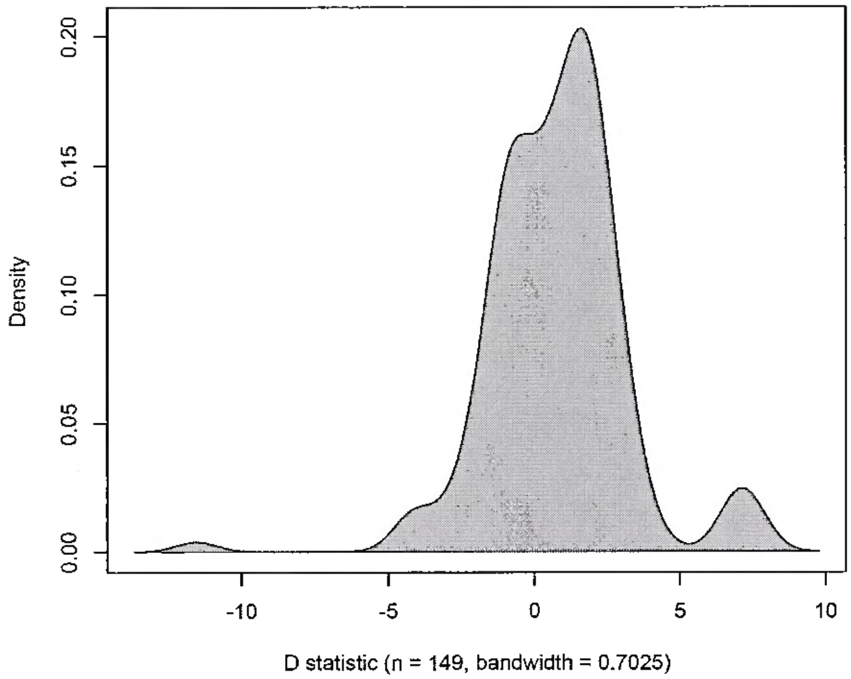


Figure 4.3: Density plot of *D* values for Ngumpin-Yapa biphones (binary).

biphone data has a mean *D* value less than zero and (including the elided rows of *D* scores) that fully 267 of the 526 observed biphones are below zero, with many more above zero but very close to it. This test indicates strong phylogenetic signal in the binary biphone data. This points to the likelihood that there was simply insufficient phonemic variation in the languages studied by Macklin-Cordes, rather than a general lack of useful signal in this type of data.

NeighborNet

In the next test, three types of NeighborNet graphs were produced in SplitsTree 4 (Huson & Bryant 2006) from files in the .nexus format generated from the Tai data, all representing sets of binary traits. These three NeighborNet graphs are for Tai phonemes, Tai biphones, and a traditional Tai cognate analysis, as the Hudak data is already organized into proposed cognate sets. The delta-scores and *Q*-residuals for each dataset are presented in Table 4.7.

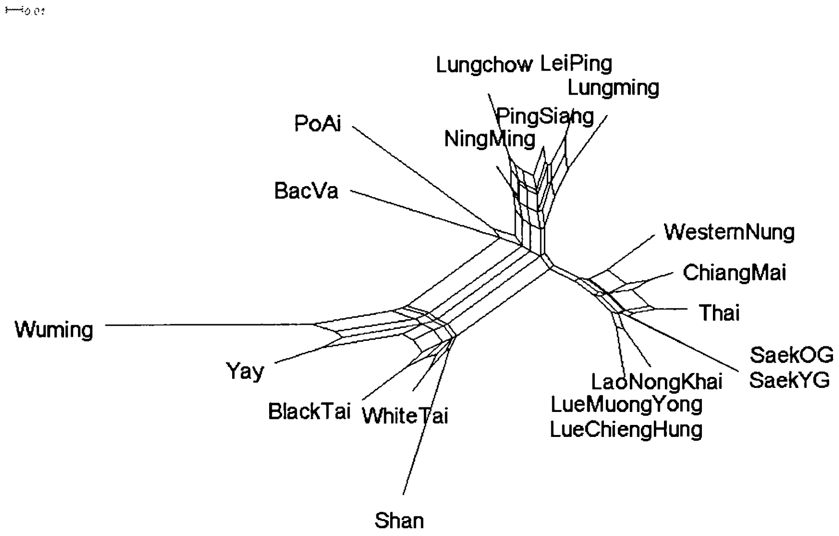


Figure 4.4: NeighborNet graph of Tai phonemes (binary).

Dataset	Delta	*Q*-residual
Tai phonemes	0.3115	0.03942
Tai biphones	0.2988	0.02615
Tai cognates	0.2808	0.04088

Table 4.7: NeighborNet measures for treelikeness in the Tai dataset.

The proper interpretation of delta-scores and *Q*-residuals is still somewhat of an open question, but as Macklin-Cordes (2015: 82) points out, for comparison, Gray et al. describe a dataset with delta-score of 0.29 and *Q*-residual of 0.05 as "moderately tree-like", and delta-score of 0.41 and *Q*-residual of 0.02 as "strikingly non-tree like" (2010: 3926-3927). Given these characterizations, it seems that all three of the NeighborNet graphs produced here for Tai data fall close to the "moderately tree-like" category. The first of the NeighborNet graphs is given in Figure 4.4.

For comparison, a traditional Tai family tree, modified from Hudak (2008), is given in Figure 4.5.[4] Figure 4.4 above picks out some clusters that closely match the tree,

[4] While this tree certainly does not represent the state of the art in Tai subgrouping, something still very

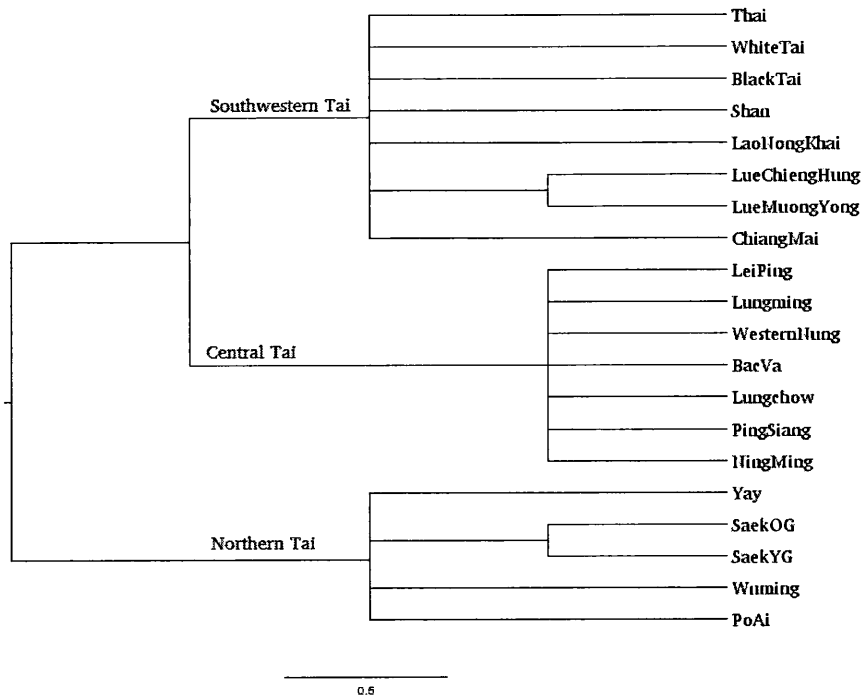


Figure 4.5: Family tree of Tai languages (adapted from Hudak (2008).

including five of the seven languages from the Central Tai branch in one cluster, five of the eight languages of Southwestern Tai mixed in another cluster with three Northern Tai languages. The other three Southwestern Tai languages, Black Tai, White Tai and Shan, are also close together, together with two other languages from Northern Tai. Assuming that the tree in Figure 5 is accurate, this would seem to indicate two subgroups each for Southwestern and Northern Tai, with horizontal transmission between the two groups.

Next, compare the tree in Figure 4.5 against the NeighborNet graph of Tai biphones in Figure 4.6. The NeighborNet has some clusters that more closely resemble the portions of the Tai tree, but in many ways picks out the same groupings as the binary phoneme data. For instance, Thai, Chiang Mai, and Lao Nong Khai cluster most closely together, which is expected given they are not only all from the Southwestern branch, but are all

much in flux, the three-branch Northern, Central, and Southwestern Tai tree, and variations on it, have been the most commonly cited classification for several decades. Pittayaporn (2009) provides the newest and most novel major subgrouping proposal, which should be compared against these results in future work, but contains primarily higher-level structure, without no resolution within the branch that corresponds to Southwestern Tai, for instance. A new lexical phylogenetic subgrouping is also currently in preparation by the author.

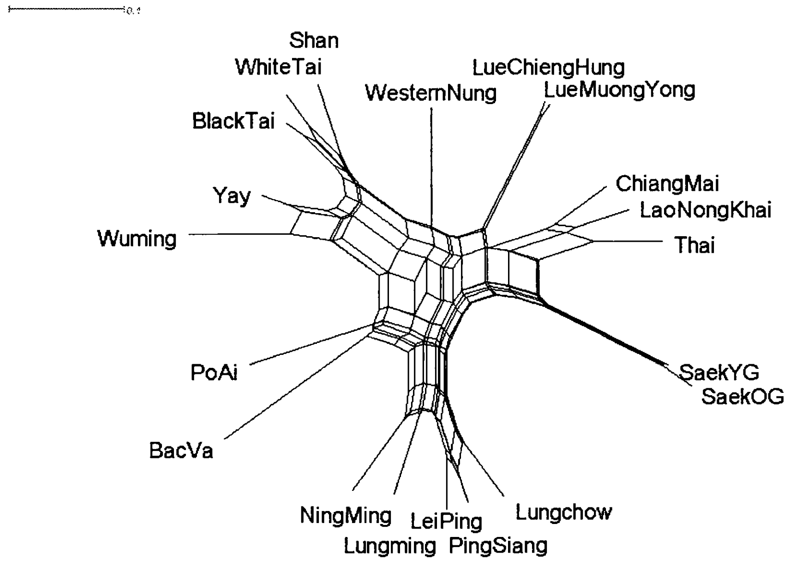


Figure 4.6: NeighborNet graph of Tai biphones (binary).

in intense longstanding contact with one another, due to all being spoken in Thailand and each having millions of speakers. The two Lue varieties, also Southwestern Tai, are not far from the Thai-Lao-Chiang Mai cluster, but are much more clearly grouped distinctly as a pair than in the previous graph. Once again five of the seven Central Tai languages are clearly clustered together at the bottom of the graph. And the same set of five languages spanning both Southwestern Tai (Shan, White Tai, Black Tai) and Northern Tai (Yay, Wuming) are also grouped closely together.

The third and final NeighborNet graph, derived from the lexical cognate data, is given in Figure 4.7. In this figure we see perhaps the best representation of at least the Southwestern Tai branch of our reference tree yet. Seven of the eight Southwestern Tai languages form an obvious cluster, with Chiang Mai being the surprising outlier. The fact that the Chiang Mai lect does not cluster with anything else may indicate lexical innovation in that language, but at the very least we can say that it is not misgrouped with anything else in the NeighborNet. Yet again, five of the seven Central Tai languages cluster together, but of the two missing, Western Nung and Bac Va, neither seems to cluster with any other language, either. The fact that we don't see Northern Tai languages grouping with South-

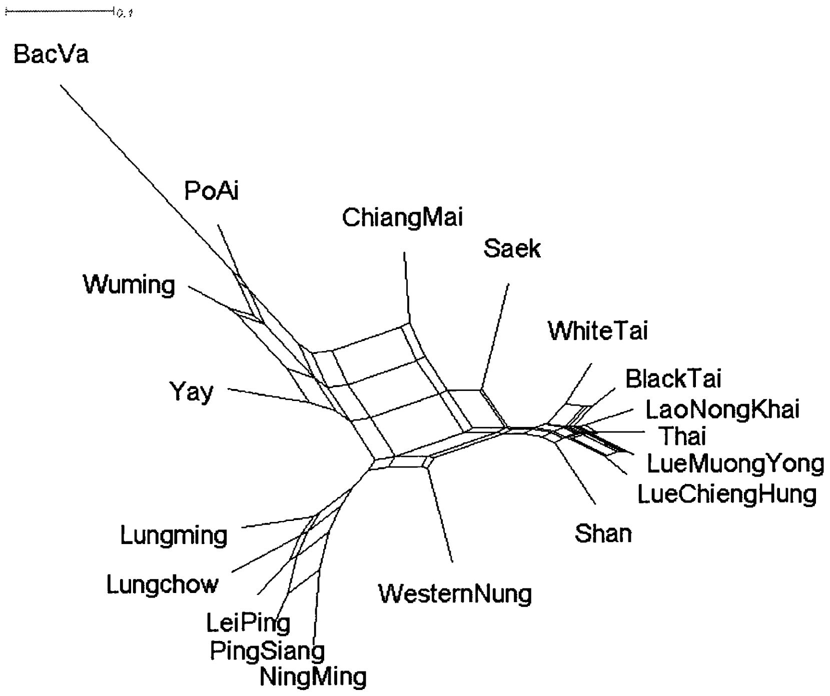


Figure 4.7: NeighborNet graph of Tai lexical cognates (binary).

western Tai languages in the way we saw before would seem to indicate that there has been horizontal phonological transfer between different Southwestern and Northern Tai languages, but often without lexical replacement.

Blomberg's *K*

The final test for phylogenetic signal deals not with binary data, as *D* and NeighborNet do, but with continuous data of both subtypes: phoneme segment probabilities and Markov chain transition probabilities. *K* statistics were calculated using the `multiPhylosignal` function of the R package `picante` (?)Kembel et al 2010).

The *K* test requires at least some variation in each trait examined, so traits which showed no variation across their respective datasets were dropped for this test. The mean *K* for each language family and data type are given in Table 4.8.

Density plots for the *K* values of these two datasets are presented in Figures 4.8 and 4.9.

Probabilities	
Phonemes	Biphones
0.71	0.68

Table 4.8: Mean K values for two types of Tai continuous data.

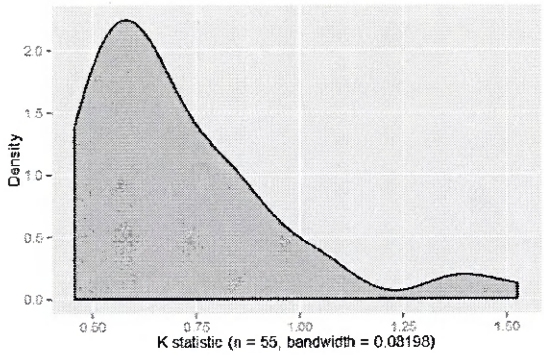


Figure 4.8: Density plot of K for Tai phonemes.

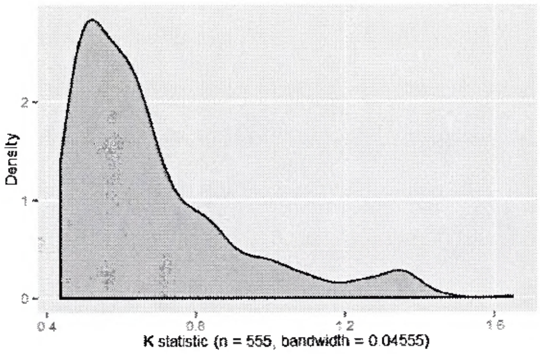


Figure 4.9: Density plot of K for Tai biphone transitions.

In testing with Blomberg's K, a score of zero indicates fully independent traits, while a score of one indicates trait distribution as expected under Brownian evolution. For comparison, Ngumpin-Yapa phone probabilities had a mean K of 0.9, and biphone transitions *0.87* (Macklin-Cordes 2015:. 89). The tests on the Tai data return a lower K on average, but still a strong indication of phylogenetic signal in the dataset overall.

4.4.3 Discussion

The preceding tests for phylogenetic signal in Tai phonological data produced a variety of positive results. In some areas, such as the D test on binary data, data from Ngumpin-Yapa proved too homogeneous to identify a phylogenetic signal, whereas with the Tai data the presence of signal was clear. In other areas, like the K test of continuous probability data, Ngumpin-Yapa produced average K values that indicated a stronger phylogenetic signal on average than the corresponding Tai data, but the set overall still demonstrated strong signal, and many individual Tai traits had very strong signal as well.

For NeighborNet tests, the two types of Tai phonological data were compared against data coded for lexical cognacy. Better performance of the cognate data is to be expected, given that that kind of data is the core of traditional language classification generally. However, there were some clusters, notably Central Tai and parts of Southwestern Tai, that were found across all three NeighborNet analyses, indicating that contra Macklin-Cordes' dismissal of binary phoneme and biphone data, given sufficient phonological variation in a set of related languages, some language clusters are recoverable from both coarse- and fine-grained binary phonological data.

Of course, since both types of data can be generated automatically from lexicons, it is hard to imagine choosing to use only phoneme presence/absence, when one could as easily generate Markov chain transition probabilities for biphones. The benefit of demonstrating signal in the low-resolution phoneme data is to further strengthen confidence in the reliability of the results obtained with high-resolution phonotactic data. To make a comparison, if the job is possible with a blunt instrument, then a finely honed one is all the more reliable.

Several additional tests of phylogenetic signal are available, depending on whether the data is binary or continuous, including Abouheif's *Cmean* (1999) and Pagel's lambda (1999), among other possibilities. In some ways these tests are still being refined themselves, especially for use in linguistics, and both their statistical power and the correct interpretation of their output is also under development. As such, comparing and contrasting results from multiple tests on novel datasets in the manner done in this study is an important part of refining the use of these methods.

Study A confirms the findings of Macklin-Cordes (2015) that phylogenetic signal is strong in the phonotactics of language. However, this study also affirms the presence of detectable signal in some areas where other studies were unable to do so.

By applying these methods to a Tai dataset, the results of these tests are made more robust, as they are shown to be useful for data from additional language families. As such, the results of the present study are of interest to linguists generally in the ongoing work of developing and testing phylogenetic methods of linguistic analysis. While the relative difficulty of using the traditional linguistic comparative method with Australian languages makes phylogenetic tools especially attractive and useful, the demonstrated results with the Tai data also shows the potential utility of these methods in other language families where traditional methods already have some traction. The Tai branch thus serves as models for the application of these tests to language families and geographical regions in need of improved language classification throughout the world.

4.5 Study B: Phylogenetic signal in tonal phonology

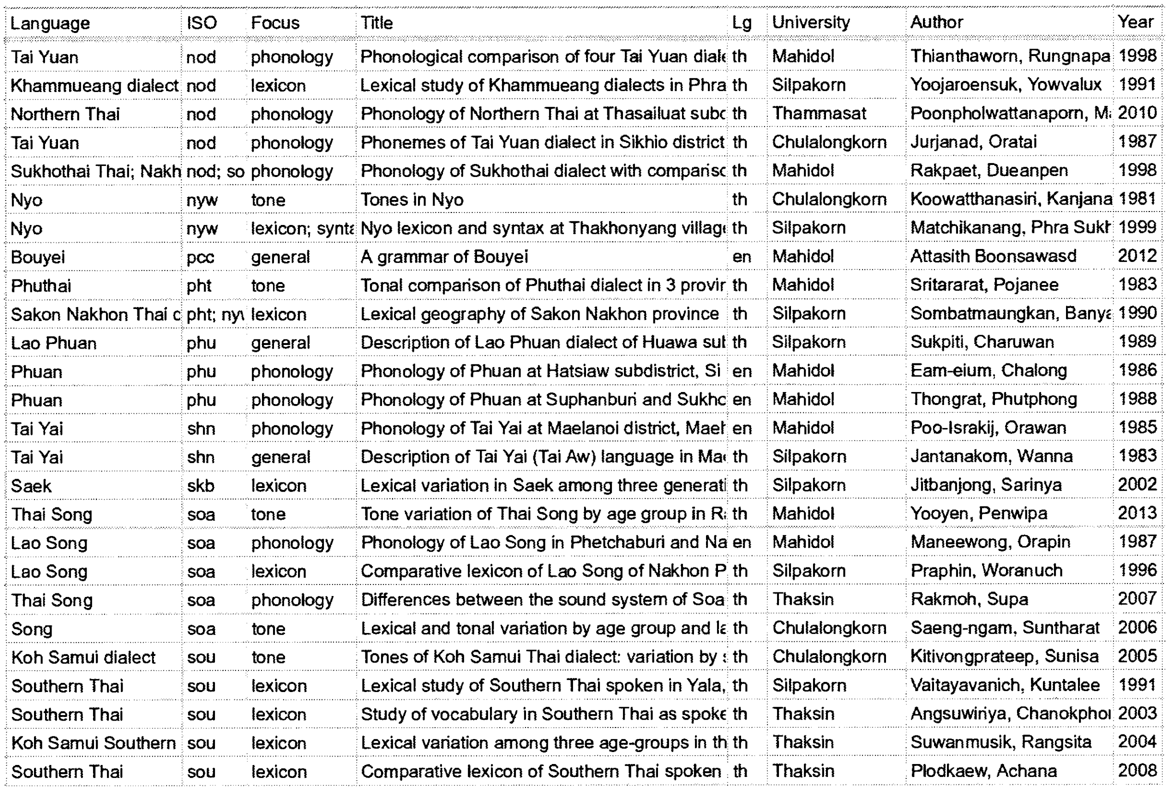
The second of the two studies examines the phylogenetic signal in tone phonology. As discussed in chapter 5, especially §5.6, the lack of a suitable object of comparison, equivalent to the segment in the traditional Comparative Method, has been a barrier to applying the Comparative Method in the tonal domain. We now know that tone typically arises as compensatory for some loss of complexity in the segmental domain (Krauss 1973). The missing link of comparability is to divide the lexicon not simply by modern toneme categories, but to connect modern tonemes to the segmental conditioning environments that gave rise to them. In other words, to divide not by modern tonemes but by historical tone classes. In this way, we have suitably comparable data that we can extract from a lexicon and code for quantitative use.

Key to identifying the suitable object of comparison for Tai tonal data was the creation of the Gedney tone box. (See Figure 4.11, and see §3.3 for more background.) The Gedney tone box compactly ties modern surface tones back to their segmental forebears. Once distilled into a tone box, the historical tone classes for any given Tai doculect can be easily aggregated in tabular format for database creation and quantitative analysis.

Study B uses binary traits extracted from those tone boxes and applies the *D* test for phylogenetic signal (see §4.3 for details on the *D* statistic).

4.5.1 Data

Data for Study B comes from a comprehensive review of resources on the tone systems of Tai doculects (see chapter 2), resulting in a dataset of tone systems from 362 Tai doculects (see tree in Figure 4.13). The survey encompasses both tonal and lexical data, from a wide variety of sources. One novel contribution of this survey is the compilation of hundreds of language documentation theses, mainly from universities throughout Thailand, which are little known and seldom cited in English-language linguistics. An example of metadata from these is given in Figure 4.10, and see also Appendix A.1 for a complete list of data sources for this study.



Language	ISO	Focus	Title	Lg	University	Author	Year
Tai Yuan	nod	phonology	Phonological comparison of four Tai Yuan diale	th	Mahidol	Thianthaworn, Rungnapa	1998
Khammueang dialect	nod	lexicon	Lexical study of Khammueang dialects in Phra	th	Silpakorn	Yoojaroensuk, Yowvalux	1991
Northern Thai	nod	phonology	Phonology of Northern Thai at Thasailuat subc	th	Thammasat	Poonpholwattanaporn, M:	2010
Tai Yuan	nod	phonology	Phonemes of Tai Yuan dialect in Sikhio district	th	Chulalongkorn	Jurjanad, Oratai	1987
Sukhothai Thai; Nakh	nod; so	phonology	Phonology of Sukhothai dialect with comparisc	th	Mahidol	Rakpaet, Dueanpen	1998
Nyo	nyw	tone	Tones in Nyo	th	Chulalongkorn	Koowatthanasiri, Kanjana	1981
Nyo	nyw	lexicon; synta	Nyo lexicon and syntax at Thakhonyang villag	th	Silpakorn	Matchikanang, Phra Sukh	1999
Bouyei	pcc	general	A grammar of Bouyei	en	Mahidol	Attasith Boonsawasd	2012
Phuthai	pht	tone	Tonal comparison of Phuthai dialect in 3 provir	th	Mahidol	Sritararat, Pojanee	1983
Sakon Nakhon Thai c	pht; nyv	lexicon	Lexical geography of Sakon Nakhon province	th	Silpakorn	Sombatmaungkan, Banya	1990
Lao Phuan	phu	general	Description of Lao Phuan dialect of Huawa sul	th	Silpakorn	Sukpiti, Charuwan	1989
Phuan	phu	phonology	Phonology of Phuan at Hatsiaw subdistrict, Si	en	Mahidol	Eam-eium, Chalong	1986
Phuan	phu	phonology	Phonology of Phuan at Suphanburi and Sukhc	en	Mahidol	Thongrat, Phutphong	1988
Tai Yai	shn	phonology	Phonology of Tai Yai at Maelanoi district, Maeh	en	Mahidol	Poo-Israkij, Orawan	1985
Tai Yai	shn	general	Description of Tai Yai (Tai Aw) language in Ma	th	Silpakorn	Jantanakom, Wanna	1983
Saek	skb	lexicon	Lexical variation in Saek among three generati	th	Silpakorn	Jitbanjong, Sarinya	2002
Thai Song	soa	tone	Tone variation of Thai Song by age group in R:	th	Mahidol	Yooyen, Penwipa	2013
Lao Song	soa	phonology	Phonology of Lao Song in Phetchaburi and Na	en	Mahidol	Maneewong, Orapin	1987
Lao Song	soa	lexicon	Comparative lexicon of Lao Song of Nakhon P	th	Silpakorn	Praphin, Woranuch	1996
Thai Song	soa	phonology	Differences between the sound system of Soa	th	Thaksin	Rakmoh, Supa	2007
Song	soa	tone	Lexical and tonal variation by age group and la	th	Chulalongkorn	Saeng-ngam, Suntharat	2006
Koh Samui dialect	sou	tone	Tones of Koh Samui Thai dialect: variation by :	th	Chulalongkorn	Kitivongprateep, Sunisa	2005
Southern Thai	sou	lexicon	Lexical study of Southern Thai spoken in Yala,	th	Silpakorn	Vaitayavanich, Kuntalee	1991
Southern Thai	sou	lexicon	Study of vocabulary in Southern Thai as spoke	th	Thaksin	Angsuwiriya, Chanokphoi	2003
Koh Samui Southern	sou	lexicon	Lexical variation among three age-groups in th	th	Thaksin	Suwanmusik, Rangsita	2004
Southern Thai	sou	lexicon	Comparative lexicon of Southern Thai spoken	th	Thaksin	Plodkaew, Achana	2008

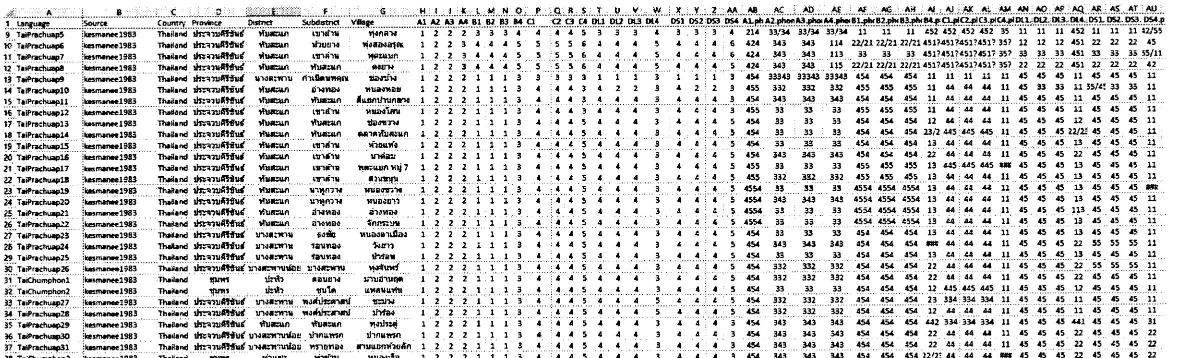
Figure 4.10: Example metadata from Tai language documentation theses.

While the linguistic analyses in these works is of highly variable quality, the data therein represent an enormous untapped resource for comparative Tai linguistics, largely unknown outside of Thailand. The type of data contained in these sources varies widely, since the range of topics is so broad: e.g. tonal studies, phonology sketches, dialectology surveys, and multi-generational studies of particular languages, to name a few. The common thread that runs through them all is the Gedney tone box, pictured in Figure 4.11.

	Proto-Tai tonal categories				
Proto-Tai initials	A	B	C	D-short	D-long
Voiceless friction *pʰ, *tʰ, *kʰ, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

Figure 4.11: Tone box for Tai historical analysis, adapted from Gedney (1972).

The Gedney box is created using a checklist of 60 lexical items that allow the linguist to quickly determine historical tone classes. Most sources have already distilled a tone box from the lexical data. Others sources without tone boxes contain lexical data from which I created a tone box. In addition to the many theses, dozens of other sources with tonal data were also compiled from the wider academic literature to fill out the dataset. See Figure 4.12 for an example of data extracted from these sources.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU
1	Language	Source	Country	Province	District	Subdistrict	Village	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	DL1	DL2	DL3	DL4	DS1	DS2	DS3	DS4	A1.ph	A2.phon	A3.phoi	A4.phoi	B1.phi	B2.phi	B3.phi	B4.p	C1.pl	C2.pl	C3.pl	C4.pl	DL1.	DL2.	DL3.	DL4.	DS1.	DS2.	DS3.	DS4.p
9	TaiPrachuap5	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	ทุ่งกลาง	1	2	2	2	3	3	3	4	4	4	4	5	3	3	3	4	3	3	3	4	214	33/34	33/34	33/34	11	11	11	452	452	452	452	35	11	11	11	452	11	11	11	42/55
10	TaiPrachuap6	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ห้วยยาง	ทุ่งสองอรุณ	1	2	2	3	4	4	4	5	5	5	5	6	4	4	4	5	4	4	4	6	424	343	343	114	22/21	22/21	22/21	451?	451?	451?	451?	35?	12	12	12	451	22	22	22	45
11	TaiPrachuap7	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	พุตะแบก	1	2	2	3	4	4	4	5	5	5	5	6	4	4	4	5	4	4	4	6	424	343	343	113	33	33	33	451?	451?	451?	451?	35?	33	33	33	451	33	33	35	55/11
12	TaiPrachuap8	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ทับสะแก	ดงยาง	1	2	2	3	4	4	4	5	5	5	5	6	4	4	4	5	4	4	4	5	424	343	343	115	22/21	22/21	22/21	451?	451?	451?	451?	35?	22	22	22	451	22	22	22	42
13	TaiPrachuap9	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	กำเนิดนพคุณ	ช่องม้าง	1	2	2	2	1	1	1	3	3	3	3	3	1	1	1	3	1	1	1	3	454	33343	33343	33343	454	454	454	11	11	11	11	11	45	45	45	11	45	45	45	11
14	TaiPrachuap10	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	อ่างทอง	หนองหอย	1	2	2	2	1	1	1	3	4	4	4	3	4	2	2	3	4	2	2	3	455	332	332	332	455	455	455	11	44	44	44	11	45	33	33	11	35/45	33	35	11
15	TaiPrachuap11	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ทับสะแก	สี่แยกบ้านกลาง	1	2	2	2	1	1	1	3	4	4	4	3	4	4	4	3	4	4	4	3	454	343	343	343	454	454	454	11	44	44	44	11	45	45	45	11	45	45	45	11
16	TaiPrachuap12	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	หนองโสน	1	2	2	2	1	1	1	3	4	4	4	3	4	4	4	3	4	4	4	3	455	33	33	33	455	455	455	11	44	44	44	11	45	45	45	11	45	45	45	11
17	TaiPrachuap13	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ทับสะแก	ช่องขวาง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	12	44	44	44	11	45	45	45	12	45	45	45	11
18	TaiPrachuap14	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ทับสะแก	ตลาดทับสะแก	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	23/2	445	445	445	11	45	45	45	22/23	45	45	45	11
19	TaiPrachuap15	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	ห้วยแห้ง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	13	44	44	44	11	45	45	45	13	45	45	45	11
20	TaiPrachuap16	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	นาต้อม	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	343	343	343	454	454	454	22	44	44	44	11	45	45	45	22	45	45	45	11
21	TaiPrachuap17	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	พุตะแบก หมู่ 7	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	455	33	33	33	455	455	455	13	445	445	445	###	45	45	45	13	45	45	45	11
22	TaiPrachuap18	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	เขาล้าน	สวนขนุน	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	455	332	332	332	455	455	455	13	44	44	44	11	45	45	45	13	45	45	45	11
23	TaiPrachuap19	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	นาหูกวาง	หนองขวาง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	4554	33	33	33	4554	4554	4554	13	44	44	44	11	45	45	45	13	45	45	45	###
24	TaiPrachuap20	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	นาหูกวาง	หนองยาว	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	4554	343	343	343	4554	4554	4554	13	44	44	44	11	45	45	45	13	45	45	45	11
25	TaiPrachuap21	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	อ่างทอง	อ่างทอง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	4554	33	33	33	4554	4554	4554	13	44	44	44	11	45	45	45	113	45	45	45	11
26	TaiPrachuap22	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	อ่างทอง	จักกระบน	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	4554	33	33	33	4554	4554	4554	13	44	44	44	11	45	45	45	13	45	45	45	11
27	TaiPrachuap23	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	ธงชัย	หนองตาเมือง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	13	44	44	44	11	45	45	45	13	45	45	45	11
28	TaiPrachuap24	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	รอนทอง	วังยาว	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	343	343	343	454	454	454	###	44	44	44	11	45	45	45	22	55	55	55	11
29	TaiPrachuap25	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	รอนทอง	ป่ารอน	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	13	44	44	44	11	45	45	45	13	45	45	45	11
30	TaiPrachuap26	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพานน้อย	บางสะพาน	ทุ่งจันทร์	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	332	332	332	454	454	454	22	44	44	44	11	45	45	45	22	55	55	55	11
31	TaiChumphon1	kesmanee1983	Thailand	ชุมพร	ปะทิว	ดอนยาง	มาบอำมฤต	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	332	332	332	454	454	454	22	44	44	44	11	45	45	45	22	45	45	45	11
32	TaiChumphon2	kesmanee1983	Thailand	ชุมพร	ปะทิว	ชุมโค	แหลมแท่น	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	5	454	33	33	33	454	454	454	12	445	445	445	11	45	45	45	12	45	45	45	11
33	TaiPrachuap27	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	พงศ์ประศาสน์	ชะม่วง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	5	4	4	4	5	454	332	332	332	454	454	454	23	334	334	334	11	45	45	45	11	45	45	45	11
34	TaiPrachuap28	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	พงศ์ประศาสน์	ป่าร้อง	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	5	4	4	4	5	454	332	332	332	454	454	454	12	44	44	44	11	45	45	45	11	45	45	45	11
35	TaiPrachuap29	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	ทับสะแก	ทุ่งประดู่	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	3	454	343	343	343	454	454	454	442	334	334	334	11	45	45	45	441	45	45	45	31
36	TaiPrachuap30	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพานน้อย	ปากแพรก	ปากแพรก	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	3	454	343	343	343	454	454	454	22	44	44	44	11	45	45	45	22	45	45	45	22
37	TaiPrachuap31	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพานน้อย	ทรายทอง	สามแยกห้วยสัก	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	3	454	343	343	343	454	454	454	22	44	44	44	11	45	45	45	22	45	45	45	22
38	TaiChumphon3	kesmanee1983	Thailand	ชุมพร	ท่าแซะ	ท่าข้าม	หนองเรือ	1	2	2	2	1	1	1	3	4	4	4	5	4	4	4	3	4	4	4	3	454	343	343	343	454	454	454	22/23	44	44	44	###	45	45	45	22	45	45	45	22
39	TaiPrachuap32	kesmanee1983	Thailand	ประจวบคีรีขันธ์	ทับสะแก	นาหูกวาง	ดอนทราย	1	2	2	2	1	1	1	3	4	4	4	5	4	3	3	3	4	3	3	5	4554	343	343	343	4554	4554	4554	13	44	44	44	11	45	13	13	13	45	22/22	22/23	###
40	TaiPrachuap33	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	ธงชัย	ทางสาย	1	2	2	2	1	1	1	3	4	4	4	5	4	3	3	3	4	3	3	5	454	343	343	343	454	454	454	13	44	44	44	11	45	13	13	13	45	13	13	11
41	TaiPrachuap34	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	กำเนิดนพคุณ	อ่าวยาง	1	2	2	2	1	1	1	3	4	4	4	5	4	3	3	3	4	3	3	5	454	33	33	33	454	454	454	13	44	44	44	11	45	113	113	113	45	13	13	11
42	TaiChumphon4	kesmanee1983	Thailand	ชุมพร	เมือง	ทุ่งคา	ดอนมะม่วง	1	2	2	2	1	1	1	3	4	4	4	5	4	3	3	3	4	3	3	5	454	343	343	343	454	454	454	13	44	44	44	22	45	13	13	13	45	13	13	11
43	TaiPrachuap36	kesmanee1983	Thailand	ประจวบคีรีขันธ์	บางสะพาน	ชัยเกษม	หนองหญ้าปล้อง	1	2	2	2	1	1	1	3	4	4	4	5	4	3	3	3	4	2	2	5	454	33	33	33	454	454	454	13	44	44	44	11	45	13	13	13	45	33	33	11
44	LaoNgaew_tanako	tanakorn1983						5	1	1	1	2	2	2	2	4	3	3	3	5	5	5	2	4	4	4	3																				
45	NakhonThai_tanak	tanakorn1983						5	1	1	1	2	2	2	2	4	3	3	3	4	4	4	3	4	4	4	3																				

Figure 4.12: Example tone data extracted from dialect surveys.

Each row of data in Figure 4.12 represents a doculect. The left third of the image is the identifying metadata, including source and data location. The middle third is the tone box flattened into a single row. The tone box has five columns, labeled A, B, C, DL and DS, representing proto-tones, and four rows, labeled 1-4, representing proto-onsets that conditioned tonal splits. The result is a set of labels A1, A2, et cetera, up through DS4. Since the tone box represents a mapping of modern tones to historical conditioning environments, each cell thus represents a subset of the native lexicon that patterns together tonally in modern lects. Each of those cells becomes a column in the flattened tone box, which is associated with a modern tone category, represented simply as symbolic numbers from 1 to *n*, where *n* is the total number of phonemic tones in the language.

The right third of the figure is the same tone box, only this time with the phonetic values of the tone categories, given in Chao (1930) tone numerals. The reason for representing the box in two different ways is the variation in the level of phonetic detail given in different sources. Often the Chao tone numerals can't simply be compared directly to determine categorical equivalency, because some authors record allotony in different environments, and so the Chao numerals for different allotones of the same toneme will differ in different cells of the tone box. Without the symbolic category numbers as a guide, it isn't always clear without careful reading of the text what level of phonetic granularity the author is using.

From tone boxes to binary traits

Once extracted from the 362 doculects, the historical tone classes from the tone boxes were converted to binary values suitable for phylogenetic analysis. Traits for Study B were created by simple pairwise comparison of every cell of the tone box against every other cell using an original script written in R (R Core Team 2019).

Traits were given names to indicate which cells of the tone box were being compared: A1 = A2, A1 = B1, and so forth. The trait was assigned the value 0 if the two cells have different modern tones in those cells of the tone box in that given lect, and assigned 1 if

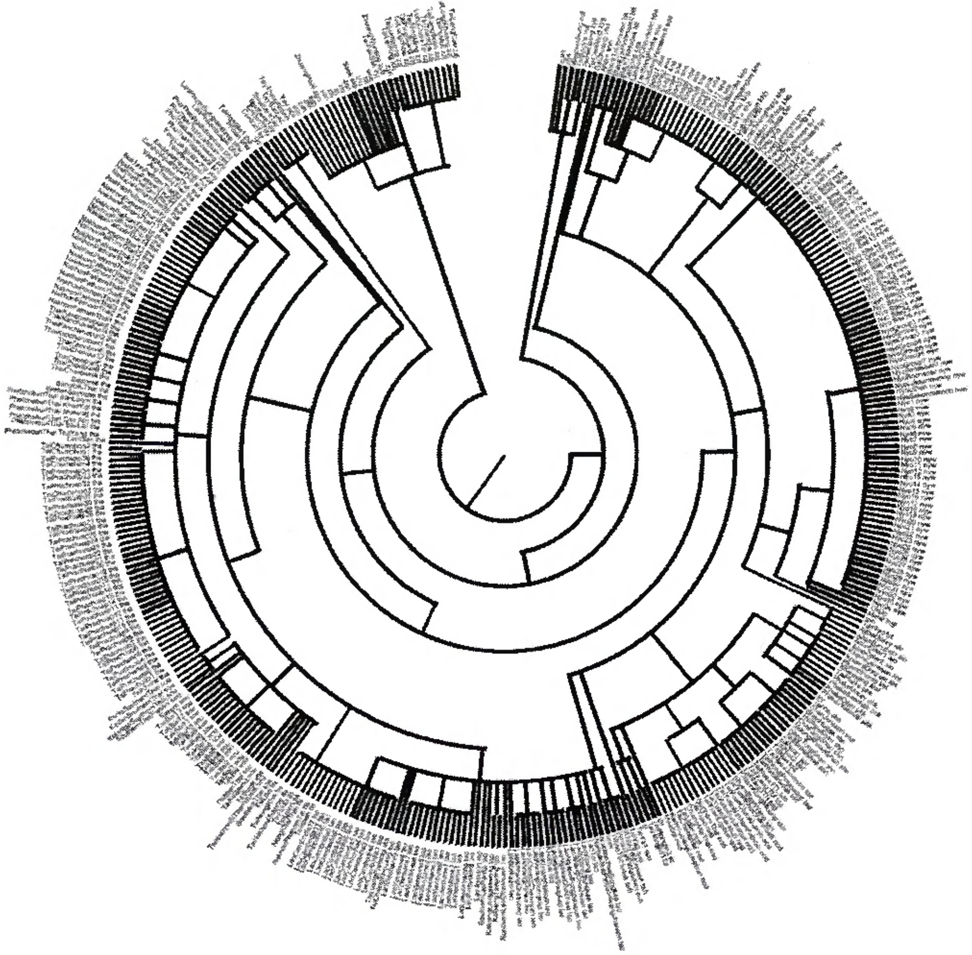


Figure 4.13: Radial tree diagram of 362 Tai doculects, mapped to entries on the Glottolog tree, and tips without associated tone boxes dropped.

the two cells have the same modern tone for those two categories.

Through this method of pairwise comparison, the traits actually use the known historical tone classes to encode sound change events in the history of each lect. Since the letter labels A, B, and C indicate the tones in Proto-Tai, if two cells belong to different columns, for instance the cells A1 and B1, this means they once had different proto-tones (different columns of the tone box), but shared a proto-onset class (same row of the tone box). If they share a modern surface tone, then there was must have been some merger across columns within that row.

This pairwise comparison must be done within each lect, and the result is a string of zeroes and ones that map between historical tone classes and surface tones in that lect, and encode tone splits and mergers. A sample of lects and some of their trait values is given in Table 4.9.

	Trait					
Doculect	A1 = A2	A1 = A3	A1 = A4	A1 = B1	A1 = B2	A1 = B3
Khamti_kht	0	0	0	1	1	1
Nyo_nyw	1	0	0	0	0	0
PhuThai_pht	1	0	0	0	0	0
Yooy_yoy	1	0	0	0	0	0
Lao_lao	0	0	0	0	0	0
Kaloeng_lao	1	0	0	0	0	0
Aiton_aio	1	0	1	1	1	1
Phake_phk	0	0	0	0	0	0
Khamyang_ksu	0	0	0	0	0	0

Table 4.9: Sample of binary traits for tone category equivalence.

This method of trait generation ignores the phonetic values of the modern tones, including allotony, and simply queries whether the description by the documenting linguist treats the two cells as part of the same surface toneme (regardless of the allotone).

If all 20 cells of the Gedney tone box are included, the result would be 190 binary traits: $(20 * 19)/2 = 190$. However, the DS and DL columns represent syllables with stop codas, which are problematic for the reasons discussed in chapter 3. Thus, only the A, B and C columns, the tones that appear on sonorant-final syllables, were included in this study. This represents 12 of the 20 cells of the tone box. Thus the number of traits is 66: $(12 * 11)/2 = 66$. Of those 66, three traits showed no variation in the data: A1 = C2, A1 = C3, and A1 = C4. A trait with no variation means that those two historical classes share the same surface tone in every doculect. Since there is no variation, no phylogenetic test can be run, and these traits were excluded, resulting in the final count of 63 binary traits.

4.5.2 Results

In Study B, the D statistic for each binary trait was calculated in R (R Core Team 2019) using a modified version of the `phylo.d` function of the package `caper` (Orme et al. 2012), set to 10,000 permutations, and provided a tree of Tai derived from Glottolog (Hammarström et al. 2019b) (see Figure 4.13). The resulting D statistics, and the p values of the two null hypotheses they are tested against, are given in Tables 4.10.

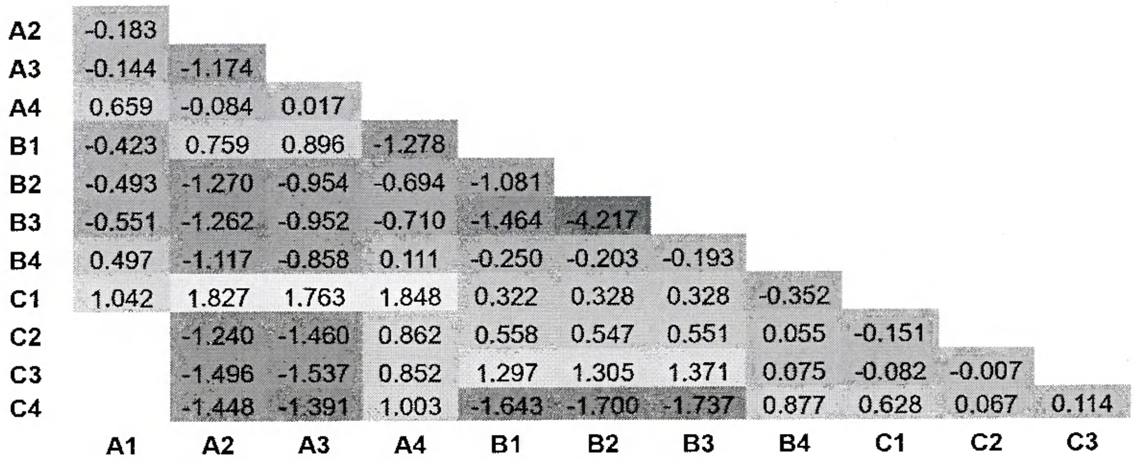
As mentioned in the results for Study A above, the way to interpret the p values is as a test of whether the D score is significantly different from the null hypothesis. So, in the first row of Table 4.10, the trait B2 = B3 has a D statistic of -4.217, which is not significantly different from 0 ($p = 0.031$), the null hypothesis of Brownian evolution, and it approaches significant difference from 1 ($p = 0.893$), the null hypothesis of a randomly distributed trait.

Of the 63 traits, the vast majority indicate strong phylogenetic signal individually, and the dataset as a whole likewise can be taken as strong evidence for the value for historical analysis of tone change events, i.e. splits and mergers. We are interested in the phylogenetic signal of the entire dataset in addition to the signal of individual traits, and indeed, we see that for the entire dataset the average D is -0.210, somewhat more phylogenetically clumped than a vanilla Brownian evolutionary model. We can thus conclude that there is strong phylogenetic signal in these tone changes as a whole.

A different visualization of the results is given in Figure 4.14, a distance table showing the D score for surface tone equivalence of any two cells from the A, B, and C columns of the tone box. The coloring is gradient between green and yellow, with the darkest green indicating the smallest D scores and strong phylogenetic signal, and yellow indicating the largest scores and no phylogenetic signal.

Trait	D	$p_{(D=0)}$	$p_{(D=1)}$
B2 = B3	-4.217	0.031	0.893
B3 = C4	-1.737	0.066	0.864
B2 = C4	-1.700	0.068	0.863
B1 = C4	-1.643	0.066	0.860
A3 = C3	-1.537	0.079	0.810
A2 = C3	-1.496	0.080	0.813
B1 = B3	-1.464	0	0.935
A3 = C2	-1.460	0.081	0.803
A2 = C4	-1.448	0.073	0.804
A3 = C4	-1.391	0.072	0.813
A4 = B1	-1.278	0	0.913
A2 = B2	-1.270	0	0.930
A2 = B3	-1.262	0	0.927
A2 = C2	-1.240	0.075	0.803
A2 = A3	-1.174	0	0.959
A2 = B4	-1.117	0	0.942
B1 = B2	-1.081	0	0.895
A3 = B2	-0.954	0	0.892
A3 = B3	-0.952	0	0.889
A3 = B4	-0.858	0	0.912
A4 = B3	-0.710	0	0.835
A4 = B2	-0.694	0	0.827
A1 = B3	-0.551	0	0.912
A1 = B2	-0.493	0	0.895
A1 = B1	-0.423	0	0.871
B4 = C1	-0.352	0	0.862
B1 = B4	-0.250	0	0.781
B2 = B4	-0.203	0	0.739
B3 = B4	-0.193	0	0.732
A1 = A2	-0.183	0	0.732
C1 = C2	-0.151	0	0.715
A1 = A3	-0.144	0	0.689
A2 = A4	-0.084	0	0.621
C1 = C3	-0.082	0	0.628
C2 = C3	-0.007	0.003	0.606
A3 = A4	0.017	0	0.489
B4 = C2	0.055	0	0.462
C2 = C4	0.067	0	0.446
B4 = C3	0.075	0	0.426
A4 = B4	0.111	0.007	0.566
C3 = C4	0.114	0	0.388
B1 = C1	0.322	0	0.238
B3 = C1	0.328	0	0.228
B2 = C1	0.328	0	0.236
A1 = B4	0.497	0.163	0.712
B2 = C2	0.547	0.043	0.480
B3 = C2	0.551	0.046	0.473
B1 = C2	0.558	0.044	0.470
C1 = C4	0.628	0.003	0.108
A1 = A4	0.659	0.029	0.244
A2 = B1	0.759	0.080	0.304
A4 = C3	0.852	0.197	0.391
A4 = C2	0.862	0.205	0.383
B4 = C4	0.877	0.051	0.001
A3 = B1	0.896	0.174	0.109
A4 = C4	1.003	0.360	0.114
A1 = C1	1.042	0.502	0.059
B1 = C3	1.297	0.584	0.310
B2 = C3	1.305	0.584	0.313
B3 = C3	1.371	0.589	0.314
A3 = C1	1.763	0.909	0.032
A2 = C1	1.827	0.912	0.031
A4 = C1	1.848	0.909	0.030
Mean D	-0.210		
SD	1.096		

Table 4.10: Tai tone D statistics and p values of the two null hypotheses, in descending order by D.



	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3
A2	-0.183										
A3	-0.144	-1.174									
A4	0.659	-0.084	0.017								
B1	-0.423	0.759	0.896	-1.278							
B2	-0.493	-1.270	-0.954	-0.694	-1.081						
B3	-0.551	-1.262	-0.952	-0.710	-1.464	-4.217					
B4	0.497	-1.117	-0.858	0.111	-0.250	-0.203	-0.193				
C1	1.042	1.827	1.763	1.848	0.322	0.328	0.328	-0.352			
C2		-1.240	-1.460	0.862	0.558	0.547	0.551	0.055	-0.151		
C3		-1.496	-1.537	0.852	1.297	1.305	1.371	0.075	-0.082	-0.007	
C4		-1.448	-1.391	1.003	-1.643	-1.700	-1.737	0.877	0.628	0.067	0.114

Figure 4.14: Table of *D* statistic results from pairwise cell comparison. Gradient shading indicates relative degree of phylogenetic signal, with dark green indicating strong signal, and yellow indicating no signal. Empty cells indicate traits that showed no variation in the data, and thus could not be tested.

4.5.3 Discussion

The results amply demonstrate phylogenetic signal in the mapping from historical tone categories to modern surface tones. However, it also became necessary to deal with the fact that generation of pairwise binary traits for every cell of the tone box is a very blunt instrument. When generated this way, the traits do not take into account whether the equivalence of two cells is meaningful due to being the same, or meaningful due to being different. That is, for some traits, the value of 0 is historically informative, if we have reason to expect them to be the same and they in fact differ (i.e. a tone split took place), and in other cases the value of 1 is informative, if we expect them to differ and they are the same (i.e. the categories merged).

To put it in more directly in the terminology of historical linguistics, it is not enough simply to say that two cells of the tone box share a surface tone. We must know whether the fact of their equivalence is likely to be a *retention* from some common ancestor or an *innovation*.

The 20-cell tone box represents the theoretical maximal set of possible tone splits.

However, no single language has anywhere near 20 tones. In fact, the majority of potential conditioning environments are not used by all lects. Early on in the history of study of Tai historical tone, before the genesis of the Gedney tone box, Brown (1965) claimed that Tai tones first split into every possible tone (though there only 15 cells in his version of the tone box, instead of 20), and then quickly collapsed into the 4-7 tones that most Tai languages have today. Gedney (1972) refutes this, as there is no empirical evidence to believe that this would be the case, and no principled theoretical reason to suppose it, either. Furthermore, given the timing of tone splits and the spread of the Tai diaspora, it is certainly the case that some tone change events must have occurred multiple times at different points in time and space, just as we see common types of segmental sound change.

The way to refine the tonal trait set, and categorize traits in a historically informed manner, is to consider them against what was likely the most recent common ancestor (MRCA) tone box, as identified using traditional historical analysis. It is uncontroversial that Proto-Tai had three tones on open and sonorant-coda syllables, conventionally named A, B and C. It is further uncontroversial that the first major tone change was a splitting of the 3-tone system into a 6-tone system as part of the Great Tone Split (Brown 1975), a change that swept across region as consonant voicing contrasts collapsed, affecting multiple language families. Thus the MRCA of all modern tone boxes is the 6-tone system, as given in Figure 4.15. The row labels are representative segments for the four consonant natural classes: voiceless aspirated, voiceless unaspirated, glottalized (or perhaps implosive), and voiced.

The MRCA tone box shed new light on the 12 cells used for trait generation in §4.5. For example, consider one of the 63 traits tested $A1 = A2$. The *D* score of that trait was -0.183, indicating strong phylogenetic signal. But without taking into account the insight from traditional methods, in the form of the MRCA tone box, we would not from the trait alone know whether sharing a tone in a modern lect (the value 1) indicates a retention or

	A	B	C
*p^h	1	3	5
*p			
*ʔb			
*b	2	4	6

	A	B	C
*p^h	A1	B1	C2
*p	A2	B2	C2
*ʔb	A3	B3	C3
*b	A4	B4	C4

Figure 4.15: The 6-tone MRCA of all Tai language tone systems, with symbolic category numbering (left) and Gedney cell labels (right).

innovation. With the MRCA tone box, however, it is possible to classify each trait based on whether any two tone box cells were the same tone after the Great Tone Split. Continuing with the same example trait, A1 = A2, when comparing with Figure 4.15, because A1 and A2 were both part of the as-yet unsplit tone A, when a value of 1 in a modern lect indicates a retention from the common ancestor. We can apply the same logic to every cell pairing.

Another insight arising from the MRCA tone box is the observation that comparing every cell pairwise results in duplicate observations. Consider column A, for example. There were only two tones after the Great Tone Split, A123 and A4. Thus the traits A1 = A2, A2 = A3, and A1 = A3 are three observations of the same fact: that voiceless onsets in the A column had not yet split, and if they share a surface tone in modern lects, it is likely to be a shared retention from the MRCA.

Both of these insights, combined with the *D* statistic results, allow us identify a refined trait set suitable for phylogenetic tasks that require more nuance than simply determining whether this category of data contains signal, namely tree inference and ancestral state reconstruction. A sampling of modified traits, and the type of tone change event that they represent, are given in Table 4.11. The table is not an exhaustive listing of the traits that will be useful for future quantitative tasks, but rather it represents the traits that had the strongest signal, while also collapsing any redundant observations.

The modified traits are named such that a value of 1 always represents an innovation,

Trait	Type
C1! = C23	split
A1! = A23	split
A23 = A4	merger
C123 = B4	merger
C23 = C4	merger
B123 = B4	merger
A1 = B123	merger
B123 = C1	merger

Table 4.11: Maximally informative tonal traits, and the type of tone change event from the MRCA that each represents.

and is phylogenetically informative. Since we expect the first three rows of each column to have been the same tone in the MRCA tone system, traits like C1! = C23 and A1! = A23 only represent innovations when they do not share a modern surface tone in some lect, i.e when a tone split took place. On the opposite side, traits such as A23 = 4, C123 = B4, C23 = C4, B123 = 4, A1 = B123, and B123 = C1 are likely to be shared retentions if the surface tones differ, and only represent innovations when a merger has taken place to give them the same surface tones.

Study B combines the statistical tools for testing phylogenetic signal with insights from traditional historical tone analysis to outline a path forward for incorporating tonal evidence into our quantitative analyses. The benefits of this hybrid approach could include resolving the internal classification problem that faces Tai. The high degree of cognacy and the relatively similar segmental inventories have contributed to the internal branching of Tai remaining an unsolved issue, which the use of tonal evidence can help to remedy.

4.6 Conclusion

The results of Study A (segmental traits) and Study B (tonal traits) are discussed in their respective sections, §4.4 and §4.5. When combined, the results of these two studies form the empirical foundation for a new type of quantitative historical analysis using phonological traits that can be derived directly, and usually automatically, from lexicons. Study A extends prior work by demonstrating that, given sufficient variation in the phonological data, these methods work at multiple levels of granularity, from simple binary phoneme presence/absence, to continuous data on the probability of biphone transitions in lexicons. Study B shows that the utility of phonological traits for quantitative study is not limited to the segmental domain, but that 'desegmental' tonal changes (i.e. those descended from segments) also contain strong signal and are suitable for similar reconstruction and classification tasks.

Neither of these studies justifies an argument to deprecate or replace the traditional Comparative Method. On the contrary, these results provide important statistical confirmation that it has been a legitimate method all along. It might be said that historical linguistics has already moved from an era where the norm was *teleo-reconstruction* (Benedict 1973), in which only a small sample of "well-chosen" languages are deemed necessary to reconstruct a proto-language, and into an era where more intermediate stages are reconstructed first wherever possible. The teleo-reconstruction era was in many cases the only option for reconstruction of a large family, due to incomplete knowledge (Matisoff 2003: 9), but it was also a philosophical and theoretical position struck by Benedict, Matisoff, and others.

Moving into the future, as lexical datasets become larger, and coverage of language families becomes ever more granular, our classifications and reconstructions will also become more granular. We need tools that allow us to form and test falsifiable hypotheses using all of the data available to us, and all of the linguistic traits available to us in the data. Expanding the use of phonological traits in our quantitative analysis moves us closer to that goal in an empirically and statistically grounded fashion. Furthermore, since tonogenesis resulted in a loss of information in the segmental inventory, and limiting prospects for reconstruction using the traditional Comparative Method, incorporating tonal evidence in our analyses will be a necessary route for overcoming this obstacle. The theory and method for this are laid out in chapter 5.

Chapter 5

The Tonal Comparative Method

5.1 Introduction

The Comparative Method (CM) is one of the primary tools of historical linguists for determining phylogenetic relationships between languages and reconstructing ancestor proto-languages. Key to its scientific validity is having generally reproducible principles for distinguishing innovations from retentions and chance resemblance (Weiss 2014). The CM traditionally focuses on segmental reconstruction, with lexical tone sometimes used to exemplify areas where the CM is not applicable at all (e.g. Meillet 1914; Campbell 2003). In this chapter I present the Tonal Comparative Method (TCM), a framework for incorporating evidence from lexical tone into our historical analyses, especially language classification, subgrouping, and reconstruction.

The idea of reconstructing tone, and the use of tonal evidence for language classification, is not new. However, the predominant conventional wisdom has long been that after the initial phonetic conditioning of tonogenesis, tones change in irregular ways, generally impervious to the CM. One result of this has been little progress toward a larger theory of tonal change, or pushing the limits of our knowledge of tone diachrony.

The Tonal Comparative Method is an extension of the logic of the traditional Comparative Method that demonstrates how to use evidence from tone in a way that is consistent with the first principles of the longstanding method. The Tai languages serve as a model for (1) a more generalized reasoning of why tonal evidence is not only possible to incorporate into a historical analysis, but ultimately will be a crucial element of the best historical analyses going into the future, and (2) how tonal evidence can resolve outstanding issues where predominantly segmental evidence has so far failed to do so. Exactly where the typological and temporal limits of the TCM lie will be left for future study, as there is certainly major variation in lexical tone worldwide. However, the aim of the present chapter is to explain the principles that unify successful tonal reconstruction in Tai and other families, and to convince the reader that we can expect the diachronic explanatory power of tone to extend well beyond the level achieved to date.

In the remainder of this chapter, §5.2 discusses the history of primarily skeptical thought toward the role of tonal evidence in historical linguistics; in §5.6, I discuss the obstacles preventing progress on this front; in §5.4 I comment on the cross-family tonal correspondences of the Sinospheric Tonbund, followed by §5.5 in which I lay out the theoretical basis for the TCM; in §5.6 and §5.7, I explain for two different stages of analysis, early and advanced, how a linguist working on a set of languages with a complex tonal situation can begin incorporate tone into diachronic analysis; and finally in §5.8, I conclude with a discussion of the limitations of the TCM.

5.2 Skepticism of tone in historical linguistics

Just as the idea of reconstructing tone is not new in linguistics, neither is the idea that tone is not diachronically informative. More than a century ago, Meillet (1914) cited tone as an example of a linguistic feature that cannot be used to establish the relatedness of languages:

> Le chinois et telle langue du Soudan, celle du Dahomey ou ewe, par exemple, peuvent se servir également de mots courts, en général monosyllabiques, faire

> varier la signification des mots en changeant l'intonation, fonder leur grammaire sur l'ordre des mots et sur l'emploi de mots accessoires; il n'en résulte pas que le chinois et l'ewe soient des langues parentes; car le détail concret des formes ne concorde pas; or, seule la concordance des procédés matériels d'expression est probante.

An English translation with modernized linguistic terminology (emphasis added) is provided in Campbell (2003: 279):

> Chinese and a language of Sudan or Dahomey such as Ewe, for example, may both use short and generally monosyllabic words, make contrastive use of tone, and base their grammar on word order and the use of auxiliary words, but it does not follow from this that Chinese and Ewe are related, since the concrete detail of their forms does not coincide; *only coincidence of the material means of expression is probative.*

This skepticism is not unwarranted. Lexical tone is an areal feature of multiple areas of the world, primarily East and Southeast Asia, Africa, and Mesoamerica, and has diffused areally across language family lines in those places. Early interest in the potential relatedness of tonal languages contributed to some of the dismissal that tone is met with in historical linguistics. Until the start of the 20th century it seemed self-evident that the tonal languages of Asia must all be related, due to shared features including tone. Indeed, it was this incorrect view that Meillet was responding to in 1914. We can point to now-defunct proposals such as Sapir's Sino-Dene hypothesis (Golla 1984: 374-382, Bengtson 1994), but certainly not all of the early ideas were wrong.

The academic genealogy of tone diachrony is not far removed from Sapir's early theory. Working with Sapir on Athabaskan fieldwork and diachrony was Li Fang-kuei, Sapir's first graduate student at Chicago. Li went on to become one of the preeminent figures in Tai historical linguistics. And while it is Haudricourt who is best remembered for his groundbreaking work on tonogenesis that showed parallel development of tones in Chinese, Tai, Hmong-Mien, and Vietnamese (1954), he stood on the shoulders of his predecessors and contemporaries both for their data and reconstructions.

After an initial period exuberance ending in the mid 20th century, matters of tone diachrony have largely been pursued for individual languages, clades, or families, perhaps due to the difficulty of establishing comparability of tones across family and regional boundaries. As a result, tone has largely been absent from the larger theorizing and debates on sound change processes. Advancements in knowledge of tone diachrony for individual language families has not been unified or generalized.

Matisoff (1973: 89) took a strong stance against the use of tone in classification, stating that for Tibeto-Burman, "tonal criteria are not even sufficient to establish genetic subgroupings for languages which are already known to be related." And even in individual language families that have made good progress on reconstructing tone, attempts to use tonal evidence for classification have not been well justified, leading to quite recent argumentation against tonal evidence by Pittayaporn (2013: 306):

> [past use of Tai tonal evidence] is not consistent with the shared-innovation method used in subgrouping, because many tonal changes may not in fact be shared innovations … A subgrouping proposal for Southwestern Tai should primarily use as criteria consonantal and vocalic changes that can be shown empirically to have occurred relatively early.

And yet it has not all been pessimism in the linguistics literature. Janda & Joseph (2003: 117) show a cautious optimism, though they do not appear to expect progress to be coming any time soon:

> …prosodic change seems fully tractable in terms of analytical methods ... time-tried for other aspects of phonological change … on the other hand, there is as yet so much to be learned … the present lack of data may enforce, at

> a minimum … one or two generations of waiting until two or more richly described contiguous points in time are available for comparison.

5.3 Obstacles to progress

A few obstacles can be identified that prevent greater use of tonal evidence in historical linguistics. One such barrier to progress is our limited understanding of sound change processes in the tonal domain generally. While our understanding of sound change in tone systems is certainly not in its infancy, it has a long way yet to come. But another important barrier to progress has been the difficulty in identifying the appropriate *object of comparison* equivalent to the *segment* in the traditional Comparative Method. To apply the Comparative Method on segmental data, we need only locate cognate sets and we can immediately begin to align segments and build up sets of regular sound correspondences. Since we cannot do this with surface tones, how do we include tonal evidence, say, in proposing a subgrouping?

This relates directly to another key barrier to progress: the inability to evaluate specific tone changes for their historical informativity. What we know about segmental sound change, and the method by which we work backward from extant languages to infer facts about the phonology of extinct languages, is built on the logic of such things as phonetic plausibility, articulatory ease, and typological likelihood. We might choose not to use a particular segmental sound change as subgrouping evidence if it occurs independently with high frequency, and thus would be a likely candidate for convergent evolution. There is a body of received wisdom for segmental sound changes that a historical linguist learns or intuits over the course of their academic career. There has been no equivalent body of received wisdom for sound change in tone systems, no principled way to state with confidence that a particular tone split or merger would be good evidence for subgrouping.

We must overcome these obstacles in order to make progress: minimally, we must have a suitable *object of comparison* to do reliable comparative work, and we must be able to *evaluate specific tone changes* in order to infer facts about their descent. The Tonal Comparative Method answers both of these needs.

5.4 Tone diachrony and the Sinospheric Tonbund

Before presenting the general framework for the Tonal Comparative Method, it is helpful to contextualize the discussion by examining parallels between the tone systems of multiple families in East and Southeast Asia.

East and Southeast Asia constitute what Matisoff (2001: 315) dubbed a Sinospheric Tonbund.[1] Indeed, the parallels are striking, with directly comparable tonal categories proposed in Sinitic, Tai, Hmong-Mien, Vietnamese, and areas of Tibeto-Burman (see Table 5.1).[2]

Tai	A (no mark)	B (*mai ek*)	C (*mai tho*)	D	Reference
Chinese	A (*píng*)	C (*qù*)	B (*shǎng*)	D (*rù*)	Haudricourt (1954)
Vietnamese	A	C	B	D	Haudricourt (1954)
Hmong-Mien	A	C	B	D	Ratliff (2010)
Karenic	A	B′	B	C	Kauffman (1993)

Table 5.1: Correspondences between reconstructed tone categories.

Matisoff (2001) posits that Sino-Tibetan is the source of tonal diffusion throughout the area, though the timing of the diaspora of various groups presents challenges to a conclusive answer. Scholars in each of these traditions, and sometimes bridging them, noticed the correspondences. Maspero (1911) was perhaps the first Western linguist to identify the relationship between onset consonants and historical tone classes in Chinese, and Karlgren

[1] But see earlier usage as "Asian 'tonbund'" in Matisoff (1985: 26). Matisoff also coined the terms *tonogenesis* (Matisoff 1970) and Sinosphere (Matisoff 1990: 113) and its contrasting term *Indosphere*, for the two key areas of cultural and linguistic influence in the region.

[2] Note the order of B and C, reversed due to the conventional Thai tone tone ordering. Quite unfortunately, the labels Li chose are exactly the opposite of the corresponding conventional labels in other families. Thus, tone B in Vietnamese is historically parallel with Thai tone C, and vice versa.

(1915) was the first to explain in depth how onsets conditioned tone splits. The first to do so for Hmong-Mien languages was Chang (1947, 1953), and for Tai languages was Li (1943).[3] Haudricourt showed the same for Karenic languages (1946) and Vietnamese (1954), and presented a unified theory of tonogenesis affecting all of these tonal languages.

(Matisoff 2001: 317) posits that, like Vietnamese, the Tai stock became monosyllabic and tonal under the influence of Chinese, and that this change is what drove the family's split from the Austronesian stock, as was posited by Benedict (1942, 1975).

In light of these areal similarities, and possible historical macro-alignment, the utility of tone for historical analysis has been a point of understandable debate. Despite his strong initial stance against tonal evidence (Matisoff 1973), recently Matisoff Matisoff (2001: 293) has suggested a middle course, adhering to the ideal of regularity of sound correspondences while keeping a practical stance on the inevitability of linguistic variation.

In their survey of tone systems in Mainland Southeast Asia (MSEA), Brunelle & Kirby (2015: 18) Brunelle and Kirby found that "the phylogenetic signal for tone is extremely strong." They used a database of 197 languages and dialects spoken in MSEA to build a statistical model of the predictability of tone systems based on genetic relatedness, maximal native wordtype (monosyllabic, sesquisyllabic, or polysyllabic), geographic proximity, and comparative population size. They conclude that family and wordtype are significant predictors of the number of tonal categories and the number of contrastive pitches, but that neither geography nor population size has any clear effect on making a language looking tonally more like its neighbor languages (2015: 18-20). They are careful to point out that their model should not be interpreted as evidence that there is never contact-induced change, but rather that the burden of proof lies on the side of proving cross-family effects, for which much additional modeling is needed (2015: 21).

[3] Observations about the 'inherent tones' of certain Thai consonant graphemes date to at least Low (1828), and traditional Thai pedagogy has been teaching three consonant classes since the 17th century (Pittayaporn 2016). These classes govern the surface tones and derive from the consonant natural classes that conditioned tone splits, though it is unclear how directly that connection would have been understood in that period.

Thus tone may be better suited for comparative work than Matisoff imagined, even though our theory of tonal change and understanding of tonal contact phenomena remain as yet poorly understood. As Chamberlain (1979) points out in reaction to Matisoff (1973), tone split patterns in Tai have been shown much more stable than in Tibeto-Burman (1979: 122-123). Among the most lasting demonstrations of the stability of tone in historical analysis is the system developed for Tai by Gedney.

5.5 The theoretical basis of the Tonal Comparative Method

An axiom of tone diachrony, now well understood, is that lexical tone arises as compensatory for the loss of segmental contrasts (Krauss 1973: 963). We can find this observation quite early, when de Lacouperie wrote of the Tai languages:

> The [Taï-Shan] language has developed tones originally as a compensation by natural equilibrium to the phonetic losses undergone in the everlasting process of intermingling.
>
> (de Lacouperie 1887: 69)

This observation, dating back well over a century, is at the core of the Tonal Comparative Method: lexical tone has its origins in former segmental contrasts. We can use the tools of comparison and reconstruction to uncover that past, and once we do, then the rest of the Comparative Method opens to us.

5.5.1 Desegmental phonology: the genealogy of tone and register

Concurrent with his work on tonogenesis, Haudricourt also studied registrogenesis, the diachronic origins of *register* (also known as *pitch register*, not to be confused with *pitch accent*). The term 'register' has an unfortunately large number of uses in linguistics, spanning multiple subfields. In this context it means the lexically contrastive use of phonetic cues originating in the larynx—phonation, glottalization, and even pitch.

Tone and register are both classified as suprasegmental, a category defined as much by what its members are not—segments—than by genuine group coherence. Suprasegmental features may scope over a wide range of phonetic and phonological units: vowel, rime, syllable, word, phrase, utterance. Fox et al. (2000: 1-11) succinctly summarizes the history of thought on suprasegmentals, noting the tendency to assume that "prosodic features are merely secondary modifications" on segments. Bloomfield (1935: 109) described them as "secondary phonemes," while Ladefoged & Johnson (2011: 24) state that they are "characterized by the fact that they must be described in relation to other items in the same utterance." These definitions are primarily intended to describe non-contrastive prosodic features, and do not fit well with lexical tone. It is not new to note that lexical tone is exceptional in this way. Bloomfield (1935: 90) stated that, unlike English, pitch variations in Chinese should be considered primary phonemes.

Register has often been studied separately from tone, and has clearly received vastly less attention in the literature, in part because some families where register is most prominent, such as Austroasiatic, are also considered predominantly atonal. In reality tone and register have a very close relationship, and as documentation of tone and register systems has increased, it has become clear that 'tonal' vs. 'atonal' is a false dichotomy, and the two concepts should not or cannot always be distinguished (Brunelle & Kirby 2015).

The diachronic unity of tone and register can be stated simply: both represent a transfer of phonemic complexity from segments onto suprasegments—a rebalancing of functional load. This close kinship highlights the need to recognize them as a meaningful subset of the suprasegmental domain. I propose the term *desegmental*, with this accompanying definition:

> **Desegmental phoneme**: a lexically contrastive suprasegmental feature that historically derives from a segmental contrast.

Tone and register can be different outcomes of the same areal sound change process. The most dramatic case of this is the East Asian Voicing Shift, a term I introduce here to refer to the massively cross-linguistic loss in onset voicing contrasts that swept across East and Southeast Asia in the early second millennium CE. The result was a doubling of the number of tones in tonal languages of many families, and new register contrasts in the atonal languages of e.g. Austroasiatic. Thurgood (2007) argued that non-modal phonation is an important trigger for tonogenesis, and possibly even a necessary intermediate stage. This idea is supported by phonetic work such as Abramson & Luangthongkum (2009) and DiCanio (2012).

Moving towards a fieldwide norm of treating these two phenomena as a meaningful *subtype* of suprasegments, and as two sides of the same diachronic coin, will assist in developing a more complete theory of desegmental sound change. This in turn will enable us to unravel those changes, since identifying the segmental origins of desegmental phonemes is a crucial step of the Tonal Comparative Method.

5.5.2 Comparing tonemes, not tones

The Comparative Method is used with wide success due to common and generally reproducible principles for building sets of sound correspondences and distinguishing innovations from retentions and chance resemblance. This method has undergone refinement over the course of two centuries. Thus, it is understandable that principles for tone system reconstruction have not emerged, due to the relatively recent understanding of tonogenesis, and relatively less access to diverse tonal data.

Having recognized that lexical tone has segmental lineage, we can identify why tone has been dismissed in the past, and explain both why we can actually expect it to work and how to make it work. To put it simply: the tones themselves have led linguists astray. In most synchronic descriptions, tones are conceived of as a set of possible melodies on syllables or words, and a language is frequently described as having *n* possible tones.

When we come along with the CM and try to compare surface tones, we will only be able to make connections in a very narrow genealogical or geographical domain. For example, of the Tai dialects spoken around Bangkok, we might have some success in identifying the varieties most closely related to Bangkok Thai simply based on the number of shared tone melodies. It does not get us far, however, and is fraught with danger, as another longstanding observation is that while tone distribution is not easily borrowed, actual tone melodies are more promiscuous.

The notion of desegmental phonology points us in the proper direction: to reconstruct tone diachrony, we are identifying correspondences not *between the surface tones*, but rather *between the historical conditioning environments* which led to those tones. Any evidence from tone melody must be secondary. There are scenarios where phonetic evidence from some particular pitch or melody may be appropriate as supporting evidence, but even so it must be dealt with extremely cautiously.

Consider a simple sound change rule like this one:

*p > b / V_V

What does a rule like this mean? Any trained linguist could restate the rule in prose: when a sound /p/ in some proto-language occurred between two vowels it became /b/ in a daughter language. But what does a rule like this actually pick out from a language? The answer is that it identifies a subset of the native lexicon, a set of etyma that pattern together historically because of a shared conditioning environment. Framing it in this way should make the connection to tone change clearer. In the segmental realm, when we propose a characteristic sound change for a subgroup, what we are observing is that for that set of languages, a reliably stable subset of the lexicon patterns together in some way. The same goes for lexical tone. When we compare tones across languages, we are comparing not their melodies, but the underlying tonal categories. In both the segmental and desegmental domain, we are comparing how the lexicon is partitioned into subsets that pattern together, and reconstructing what conditioned a particular change shared by more than one language.

Thus, two languages having a particular tone melody in common is no better evidence for a common ancestor than two languages having a particular consonant. It only becomes significant when we can demonstrate regularity and a shared conditioning environment. And unlike the reputation tone has for changing haphazardly—which their surface realization can certainly do—treating tones in this way yields extremely robust phylogenetic information, as demonstrated statistically in chapter 4.

5.6 Applying the method: early stage

This is the first of two sections in which I discuss how to apply the Tonal Comparative Method. Each section addresses one of the obstacles to progress discussed in . First is the early stage, in which I illustrate how to begin from first principles to identify tonal correspondence sets, without prior knowledge of the history of tone in a set of related languages. This addresses the obstacle of the lack of an appropriate object of comparison.

After that, in 5.7 I discuss the late stage application of the method, on distinguishing shared innovation from parallel innovation and shared retention. That stage comes after the segmental origins are well understood, and the native lexicon has been partitioned into historically coherent tone categories. This addresses the obstacle of the lack of way to evaluate the historical informativity of particular tone changes.

It may seem that the explanation of this early stage is not really needed, given the fact that tone has been reconstructed in multiple families for decades. I explain it in detail because moving toward a more complete theory of tone change dictates being better able to illustrate *why* and *how* the method works from first principles, if we hope to make progress in other families and clades where there has been less success. Indeed, the general progression of the method as I describe it below is is not far off from how Gedney worked prior to his arrival at the well-known tone box, as described in Gedney (1964).

5.6.1 Identifying correspondence sets

As discussed above, the phonetic tones are not the immediate object of comparison in the Tonal Comparative Method. Rather, it is the historical tone categories and their conditioning environments that must be identified through uncovering their segmental origins. Thus, before there can be tonal correspondence sets, there must be segmental correspondence sets. The beginning stages of the Tonal Comparative Method is simply the usual Comparative Method. Basic lexical items from multiple languages are organized into tentative cognate sets, and a general picture of the segmental inventory is compiled.

When the time comes to build tonal correspondence sets, the tone categories act as evidence for neutralized segmental contrasts. There is no fixed time at which to begin to compare tonal evidence. It entirely depends on how much is known of the family and the varieties under study. Here I explicate the method as if there is no prior knowledge of the history of tones in that set of languages, which may seldom actually be the case. One thing that is certain, though, is that the process must be a feedback loop, as in the Comparative Method generally. Reconstruction of any past stage of linguistics is necessarily an iterative process of refining. Prior to developing his tone box, Gedney described the system of slips of paper and boxes that he used to develop tonal correspondence sets for several Tai languages (Gedney 1964). Sans physical paper and boxes, the TCM follows a similar path.

Consider the toy dataset in Table 5.2, which uses data adapted from Gedney (1964). The actual data originally comes from Thai, White Tai, Black Tai, and Red Tai, from left to right, but as this illustration is meant to build from first principles and does not assume the standard knowledge of Tai tone diachrony, I have simply assigned them the arbitrary labels W, X, Y, and Z.

		W	X	Y	Z
1.	‘to come’	ma:1	ma:4	ma:4	ma:4
2.	‘hundred’	rɔ:y4	hɔy6	hɔy6	hɔy5
3.	‘to burn’	phaw5	phaw1	faw1	faw1
4.	‘to hate’	chaŋ1	caŋ4	caŋ4	caŋ4
5.	‘two’	sɔ:ŋ5	sɔŋ1	sɔŋ1	sɔŋ1
6.	‘finger, toe’	niw4	niw6	niw6	niw5
7.	‘deer’	kwaaŋ1	kwaaŋ1	kwaaŋ1	kwaaŋ1
8.	‘salt’	klɯa1	kə1	kɯa1	kɯa1
9.	‘bean’	thua2	tho2	thua2	thua2
10.	‘fever’	khay3	chay3	say3	say3
11.	‘to go up’	khɯn3	xɯn3	khɯn3	khɯn3
12.	‘to leak’	rua3	ho5	hua5	hua3
13.	‘goose’	haan2	haan2	haan2	haan2
14.	‘dry field’	ray3	hay5	hay5	hay3

Table 5.2: Stage 1: Laying out cognate sets

At this point, we might also temporarily discard the segmental form for convenience in grouping, retaining only the symbolic tone category numbers, as seen in Table 5.3.

		W	X	Y	Z
1.	‘to come’	1	4	4	4
2.	‘hundred’	4	6	6	5
3.	‘to burn’	5	1	1	1
4.	‘to hate’	1	4	4	4
5.	‘two’	5	1	1	1
6.	‘finger, toe’	4	6	6	5
7.	‘deer’	1	1	1	1
8.	‘salt’	1	1	1	1
9.	‘bean’	2	2	2	2
10.	‘fever’	3	3	3	3
11.	‘to go up’	3	3	3	3
12.	‘to leak’	3	5	5	3
13.	‘goose’	2	2	2	2
14.	‘dry field’	3	5	5	3

Table 5.3: Stage 2: Isolating tone correspondences

Next we identify the number of unique correspondence set types by collapsing each row of digits into a string, and each unique string forms a type. Here we have a total of seven unique correspondence sets: 1:1:1:1, 1:4:4:4, 2:2:2:2, 3:3:3:3, 3:5:5:3, 4:6:6:5, and 5:1:1:1. We can sort the table accordingly, as in Table 5.4, with alternating correspondence sets shaded for visual clarity.

		W	X	Y	Z
1.	‘deer’	1	1	1	1
2.	‘salt’	1	1	1	1
3.	‘to come’	1	4	4	4
4.	‘to hate’	1	4	4	4
5.	‘bean’	2	2	2	2
6.	‘goose’	2	2	2	2
7.	‘fever’	3	3	3	3
8.	‘to go up’	3	3	3	3
9.	‘to leak’	3	5	5	3
10.	‘dry field’	3	5	5	3
11.	‘hundred’	4	6	6	5
12.	‘finger, toe’	4	6	6	5
13.	‘to burn’	5	1	1	1
14.	‘two’	5	1	1	1

Table 5.4: Stage 3: Grouping correspondence sets

This toy dataset has exactly two cognate roots for each tonal correspondence set, but in a real use case we would do this for as much data as was available, the ideal minimum being the 400 lexical items needed to reliably observe every phoneme in a language (Dockum & Bowern 2019). In this way, we would know which correspondence sets are the most robust, because they would have the largest number of regular correspondences. There is no sense in pinning it to a precise number, but the idea is that tone correspondence types that are infrequent are possibly spurious, contaminated by loans, or subject to other problems. This is part of the reason more data than, say, a minimal Swadesh 100 list is important.

Assuming that our underlying dataset was correctly organized into cognate groups, at this point we can make observations about tone changes in these languages. Note that independently of other evidence, we would not yet be able to ascertain directionality of change. Observations to be made from Table 5.4 might include:[4]

- Tone W5 either split from W1, or there was some merger in XYZ1.
- W1 may have undergone a merger (as evidence from the tone contrast of XYZ1 and XYZ4).
- XY5 either split from XY3, or a corresponding category merged in W3 and Z3.
- There is total correspondence for 2:2:2:2 and 3:3:3:3, perhaps indicating that these tone categories are retained from a common ancestor.

We can begin to evaluate the direction of change, and the segmental origins of the tones, by returning to information from Table 5.2. A table organized by correspondences, listing attested onsets that occur with each one, is given in Table 5.5. Two notes: (1) For better illustration I have included onsets from additional cognates beyond the 14 included in Table 5.2; and (2) this table gives only onsets. In a more complete version we would also make separate tables for onsets, vowels, codas, or other factors we suspect might have conditioned tone change in those languages.

[4] Since I am using arbitrary sequential integers to identify the tone categories of each language (in fact borrowing those assigned by Gedney), I use the shorthand W1 for tone 1 of language W, X3 for tone 3 of language X, etc. Where more than one language patterns together and they share an integer label, I use multiple consonants, e.g. XYZ1 refers to tone category 1 in those three languages.

	Correspondence	W	X	Y	Z
1.	1:1:1:1	p, t, c, k	p, t, c, k	p, t, c, k	p, t, c, k
2.	1:4:4:4	f, r, th, kh	f, h, t, x	f, h, t, k	f, h, t, kh
3.	2:2:2:2	n, th, h, k	n, th, k, h	n, th, k, h	n, th, k, h
4.	3:3:3:3	m, k, kh, h	m, ch, k, x, h	m, s, k, kh, h	m, s, k, kh, h
5.	3:5:5:3	ph, th, kh, r	p, t, x, h	p, t, k, h	p, t, k, h
6.	4:6:6:5	m, r, l, s, j	m, h, s, ɲ	m, h, l, s, ɲ	m, h, l, s, ɲ
7.	5:1:1:1	ph, s, m, n	ph, s, m, n	f, s, m, n	f, s, m, n

Table 5.5: Stage 4: Identifying tone-onset correspondences

First, look to see if any tonal correspondence set matches with an obvious natural class of segments, or very nearly matches. In our dataset, correspondence 1:1:1:1 should immediately jump out, with the only onsets attested being voiceless unaspirated stops: /p, t, c, k/. Next, look for onset clusters that minimally differ from a natural class. In correspondence 4:6:6:5, we have the voiced sonorants, /m, r, l, j, ɲ/, and then two voiceless fricatives /s, h/. Note that all are continuants as well. This could tip us off to voicing as a conditioning environment, where a merger affected only voiced obstruents in that environment. In fact, this is precisely what happened, but the point is that we generated this hypothesis from a very small dataset. The tonal evidence enriches and clarifies the potential historical analysis.

It is here where we hit the limits of the usefulness of our toy dataset. But a further suggestion would be to look for any onset clusters that are a proper subset of one another, or where this is is the case for some languages but not others. Depending on the quantity and complexity of data, it may also be helpful to subdivide each tonal correspondence set into its component segmental correspondence sets. For instance, consider the additional data on tone-onset correspondence in Table 5.6.

	Tones	Onsets
1.	1:1:1:1	p:p:p:p
2.	1:1:1:1	t:t:t:t
3.	1:1:1:1	c:c:c:c
4.	1:1:1:1	k:k:k:k
5.	5:1:1:1	th:th:th:th
6.	5:1:1:1	kh:kh:kh:kh
7.	5:1:1:1	ŋ:h:h:h
8.	5:1:1:1	h:h:h:h

Table 5.6: Stage 5: Detailed tone-onset correspondences

When laid out like this, the picture immediately becomes clearer. For tone correspondence set 1:1:1:1 we have total correspondence among voiceless unaspirated stops, as mentioned above. For 5:1:1:1, we have total correspondence for voiceless aspirated stops, as well as h:h:h:h. We might infer from this a split based on aspiration. The exception in Table 5.6 above is ŋ:h:h:h. This hints at the solution: a voiceless nasal *ŋ̊, intermediate between /ŋ/ and /h/, which became /ŋ/ in language W, but merged with /h/ in X, Y, and Z. Once again, I suggest this because we know from other analyses that this is what actually happened. But once again the point is to demonstrate how we can come to these conclusions from the first principles of the Comparative Method, which is then clarified and extended by the tonal evidence.

5.7 Applying the method: late stage

In this section we now jump ahead to a scenario where historical tone categories have been established that tie tones back to segmental origins. As discussed in §5.4 above, this is the case for several language families and subgroups in East and Southeast Asia.

This is where Tai tone diachrony has stood for nearly half a century. The key obstacle to overcome at this stage is differentiating between tone changes that are valid for e.g. subgrouping, and those that are not.

5.7.1 Evaluating tone changes

At this point of the method, tonal correspondences are well developed, some set of proto-tone categories have been inferred, and a set of tone changes—splits and mergers—have been observed between related lects. While the previous section was intended for those who want to apply the method from scratch in a set of related languages, this section is very much intended for cases like the Tai languages, where knowledge of the historical tone categories is relatively advanced.

A key use of the conventional CM is for reconstructing proto-environments, in order to posit sound changes. These sound changes are then examined to determine which ones constitute shared innovations from a common ancestor, and are characteristic of some group of more closely related languages, thus forming a subgroup. In the segmental CM, this is where we bring all of the received wisdom to bear, on the phonetic plausibility, articulatory ease, or typological likelihood of some sound change over some other one.

This is not yet directly possible for tone changes, i.e. tone splits and mergers. There are a number of reasons why the same kind of tone change may appear in multiple lects. Shared innovations must be distinguished from shared retention, and parallel innovation. In past cases where tone has been used to argue for a particular subgrouping in Tai (e.g. Chamberlain 1972b, 1975), this issue has not received any consideration at all, because there was no way to evaluate tone split or merger.

In §4.5, I demonstrated that tonal splits and mergers contain phylogenetic signal. I also identified a set of tonal traits that allows us to evaluate whether particular tone changes are likely to be innovations, and thus phylogenetically informative, or retentions, and thus phylogenetically uninformative. Consider the tree of the Southwestern Tai subgroup proposed by Chamberlain (1975) in Figure 5.1.

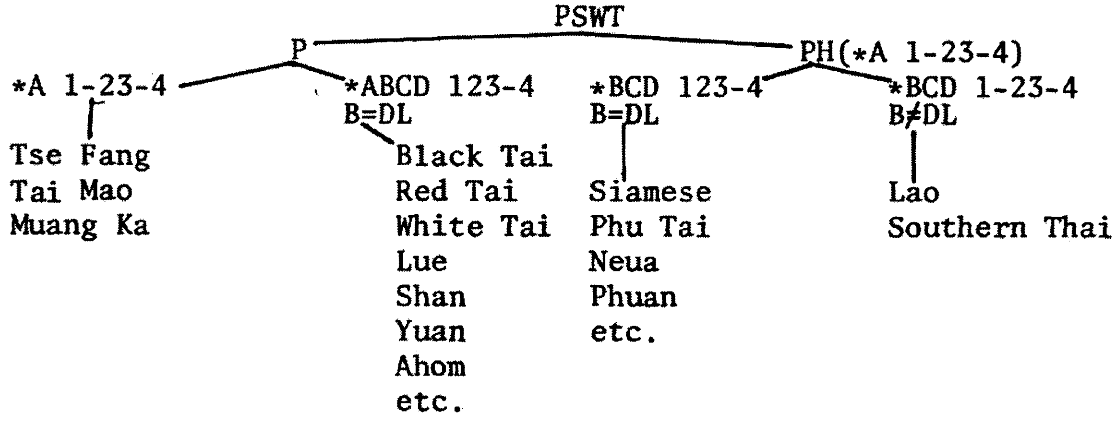


Figure 5.1: Subgrouping of the Southwestern Tai branch proposed in Chamberlain (1975).

The primary division of the tree is into the so-called P and PH groups, based on whether the voiced stop series in Proto-Southwestern Tai became voiceless aspirated or voiceless unaspirated. But within each of those branches, the remaining characteristic sound changes are entirely tonal. At this stage, we want to set aside the question of the D tones, but this leaves us with a few tone changes to consider. Consider one such tone split: *ABCD123-4, claimed as characteristic for the subgroup within the P group that includes Black Tai, Red Tai, White Tai, Lue, etc. The way to read Chamberlain's notation is that if the first three rows of each tone all share a surface tone, and only the fourth row differs, then languages that have this tone box configuration form a coherent subgroup.

But now consider what I described in §4.5 as the most recent common ancestor (MRCA) tone box, or the tone system that all Tai languages descended from at some point in time, given in 5.2.

The tonal split that Chamberlain has proposed as characterizing this subgroup is in fact exactly what we see in the MRCA. In other words, he has proposed subgrouping based on a tonal trait that can now be stated with a high degree of certainty to be a shared retention from a common ancestor proto-language. Exactly the same is true for the *BCD123-4 within the PH group (the A column differs in this case). Since a 123-4 split is a feature of the MRCA tone box of all languages in the family, this is not a coherent subgrouping

	A	B	C
$*p^h$	A1	B1	C2
*p	A2	B2	C2
*ʔb	A3	B3	C3
*b	A4	B4	C4

Figure 5.2: The 6-tone MRCA of all Tai language tone systems with Gedney cell labels.

criterion.

In §4.5 I also presented a table of several traits that are the most historically informative, and in which direction (split or merger). This is reproduced in Table 5.7.

Trait	Type
C1! = C23	split
A1! = A23	split
A23 = A4	merger
C123 = B4	merger
C23 = C4	merger
B123 = B4	merger
A1 = B123	merger
B123 = C1	merger

Table 5.7: Maximally informative tonal traits, and the type of tone change event that each represents.

These are some of the traits that we should be looking for. And indeed, in Chamberlain we see something similar to the trait A1! = A23 in his subgroup of Tse Fang, Tai Mao, and Muang Ka: the trait *A1-23-4 in the P group. We actually know that the split between 3 and 4 is not informative, but if we charitably interpret it as *A1-23 instead, then this is in fact one of the traits on our table: A1! = 23. So in theory this may be a useful subgrouping criterion.

Ultimately, we will need more than just the top several traits to assist us in making subgrouping decisions and deciding what categories to reconstruct for internal nodes of the tree. But this kind of reasoned argumentation, making reference to the statistical signal of a given trait overall, and comparing against an MRCA tone system, help to move us down the path toward complete tone system reconstruction.

5.8 Limitations of the Tonal Comparative Method

One potential criticism of the TCM is that the situation with the regularity of tone in Tai (and the larger Kra-Dai) is unique, and so the method is not generalizable beyond the family. Just as with the traditional segmental approach, the Tonal Comparative Method is subject to limitations. The extent of those must be left to future study. The present aim is make the case for why we should expect the method to generalize, even if it does not do so universally.

Even if we stipulate that the method is generally valid, it may still not be the case that the method is useful for every tonal language clade. Weiss (2014: 137-139) discusses some of the limitations to the traditional Comparative Method: complete loss, time depth, and convergence. The TCM certainly has similar weaknesses. For instance, the complete loss of evidence for some phoneme, such as through total merger, is not directly recoverable by the CM. The way around this is by uncovering indirect evidence for a former contrast by looking in related languages. Weiss uses the example of an *a*   *a* correspondence in Gothic and Sanskrit. It is only by expanding the comparison set to include Greek that we would know that two phonemes once existed in those data: *a and *o, and that they merged independently of one another.

Tone presents similar challenges. The gradual increase in complexity over time of the tone boxes used to study Tai tone diachrony (see §3.3) is an example of this same limitation at work. The Gedney box (Gedney 1972), with its 20 cells, stood for decades as the state of the art, and for the most part remains so. Each of the 20 cells represents a subset of the native lexicon, sharing a conditioning environment wherein some (usually) neutralized segmental contrast yielded to a new tonal contrast. And yet Gedney was under no illusion that he had learned all there was to know. He correctly predicted the discovery of additional conditioning environments (Gedney 1967), recently documented by Hanbo (2016).

Chapter 6

Illustrating the Tonal Comparative Method: Tai Khamti

6.1 Introduction

This chapter presents a case study of the application of the Tonal Comparative Method using data from multiple dialects of Tai Khamti. In this chapter I: (1) demonstrate the reconstruction of tone categories of Proto-Tai Khamti, the common ancestor of tonally divergent Tai Khamti dialects, using the Tonal Comparative Method; and (2) survey previous work on Southwestern Tai subgrouping, combining new field data with data from the literature to reexamine the place of Tai Khamti within Southwestern Tai, and demonstrate how tonal evidence corroborates segmental evidence and assist in resolving competing subgrouping hypotheses proposed with traditional methods.

Tai Khamti (hereafter TK), also known as Khamti or Khamti Shan, is a member of the Southwestern Tai (SWT) subgroup within the Tai branch of the Kra-Dai family, one of the five major language families of Southeast Asia. The term Khamti and its variations have been widely used to refer to language varieties in both India and Myanmar, and scholarship has largely treated them as the same, despite little comparative work between the two, and indeed, almost no data on TK from Myanmar in the literature at all. While TK of India is still far from well described, it has received the bulk of the attention to date (e.g. Robinson 1849; Needham 1894; Grierson 1904; Harris 1976; Weidert 1977). This basic need for documentation of TK in Myanmar, as an underdocumented and endangered language, provided the motivation for a pilot field research trip to Khamti Township, Myanmar, in summer 2014, followed by annual fieldwork trips from 2015 to 2018.

Chindwin TK shows just four lexical tonemes, as opposed to the five previously described for TK spoken in Arunachal Pradesh by Harris (1976), and analysis following the Gedney (1972) tone box method (see 3.3) shows a very different history of tone splits and mergers. Another doculect from India, this time a historical one, was documented in a language sketch by Robinson (1849), one of the earliest examples of Tai lexical documentation that explicitly notes the tone for each word in a wordlist. From that lexicon Morey (2005b) reconstructed the tone system of this variety, showing a tone system in some ways more similar to modern Chindwin TK than modern Indian TK, despite itself being a description of the language as spoken in India in the 19th century. Previous subgroupings that take TK into account have been based solely on Indian data.

Southwestern Tai is a branch that consists of some 32 languages, and the internal structure of the branch remains unsettled. Typically this is attributed to the difficulty of low-level subgrouping for languages that are so geographically close (leading to sustained language contact), have such a high cognacy rate (making lexicostatistics difficult), and are relatively phonologically homogeneous (limiting the traction of the conventional Comparative Method). Expert subgroupings for the larger Tai family have sometimes avoided proposing any internal structure within SWT at all, and subgroupings that do exist for SWT usually only consist of one or two major divisions (see §6.3).

The position of Tai Khamti [kht] within SWT has been a point of frequent disagreement. In the literature to date, TK has largely been treated as a monolithic language. This is problematic because the majority of work on TK has been from data gathered in India (e.g. Harris 1976; Weidert 1977), despite the scholarly consensus that TK speakers in India migrated centuries earlier from northern Myanmar (Inglis 2004: 26), which is consistent with local tradition. As a result, most classification work that takes TK into account has been made based on descriptions of Indian TK, while little data from Myanmar has been available in the literature. I conducted language documentation work over the course of five years, 2014-2019, in Khamti District, Myanmar, and refer to this doculect hereafter as Chindwin Tai Khamti (abbreviated Chindwin TK) to distinguish it from the language of the more populous Tai Khamti speaker community spoken in and around Putao in Kachin State, the area traditionally held to be the migratory source of both other dialects.

The Tonal Comparative Method (see Chapter 5) allows us to reconstruct a Proto-TK tonal system and identify its closest relatives in SWT with greater confidence. This case study demonstrates the utility and reliability of the Tonal Comparative Method for use in language classification, while also serving as a critique of the lack of rigor that has sometimes befallen the use of tonal evidence in past work in Tai historical linguistics.

In the remainder of this chapter, §6.2 discusses data sources; in §6.3, I review past proposals on subgrouping within the SWT branch. §6.4 presents the tone systems of Chindwin TK and modern and historical Indian TK, and I resolve discrepancies between them to reconstruct the tone system of their nearest common ancestor, Proto-Tai Khamti. In §6.5, I discuss the nature of sound change in tone systems and why we see only tone mergers posited between the Proto-TK stage and its daughters; §6.6 compares tonal evidence for subgrouping against segmental evidence, showing that the two are in concord; §6.7 discusses the implications this paper has for subgrouping in SWT, finding some support for the SWT "northern tier" of Edmondson and Solnit (1997) and Edmondson (2008). §6.8 concludes the chapter.

6.2 Data

6.2.1 Elicited data

For wordlist elicitation in my fieldwork, the Southeast Asia 436 Word List prepared by SIL (2002), was used as the primary wordlist for elicitation. This list is subdivided by semantic domain, covering such areas as nature, plants, food, animals, body parts, kinship terms, home, numbers, dimensions, question words, basic verbs, and basic adjectives. The list includes glosses in English, Central Thai (i.e. Siamese), Northern Thai, and Burmese. Despite a few errors discovered on the wordlist itself, access to a wordlist with Burmese glosses proved extremely helpful, as all TK speakers who acted as consultants also spoke Burmese natively, and were literate in Burmese. Over the course of elicitation and through the process of textual analysis with consultants, many hundreds more words were recorded beyond those contained in the SIL wordlist. Text elicitation took several forms, including narrations of the Pear Story video, a number of "frog story" wordless picture books by Mercer Mayer, traditional work chants, songs, rhymes, and proverbs, and numerous traditional folk stories. Audio recordings of these materials, many with transcriptions, glosses, and translations, are permanently archived in PARADISEC Dockum (2014-2018).

6.2.2 Compiled data

The main sources of data used in this chapter are: (1) Hudak (2008), a comparative set of 1159 cognate sets from 19 Tai languages, of which 8 are SWT, compiled from Gedney's original field data; (2) Jonsson (1991), a reconstruction of Proto-SWT, which compares several hundred words from 10 SWT languages, including TK (data from Harris 1976); (3) Inglis (2004), a lexicostatistical comparison of 120 words from 6 varieties of "Khamti Shan"; (4) Yaowen & Meizhen (2001), a Chinese government survey consisting of 1757 words in 9 SWT languages spoken in southern Yunnan Province, China; (5) Cuirong

(2009), another Chinese minority language survey containing 1034 items for 2 SWT languages, and (6) Morey (2005a), with data drawn from grammars and texts for several SWT languages of Northeast India. Across all of these datasets, around 30 distinct SWT doculects are represented.

6.3 Southwestern Tai as a subgroup

It was on primarily lexical evidence that Li (1959) first proposed Southwestern Tai (SWT) as one of three branches of Tai, along with Northern Tai (NT) and Central Tai (CT). These labels are roughly geographic, corresponding to lower mainland Southeast Asia (SWT), upper mainland Southeast Asia (CT), and southern China (NT), but not without significant geographical overlap. This tripartite division was reinforced by Li's (1960) subsequent phonological analysis and his ultimate large-scale reconstruction work (Li 1977). The SWT-CT-NT division remained the conventional formulation for half a century, and is often still cited. See Figure 6.1 for Li's original classification.

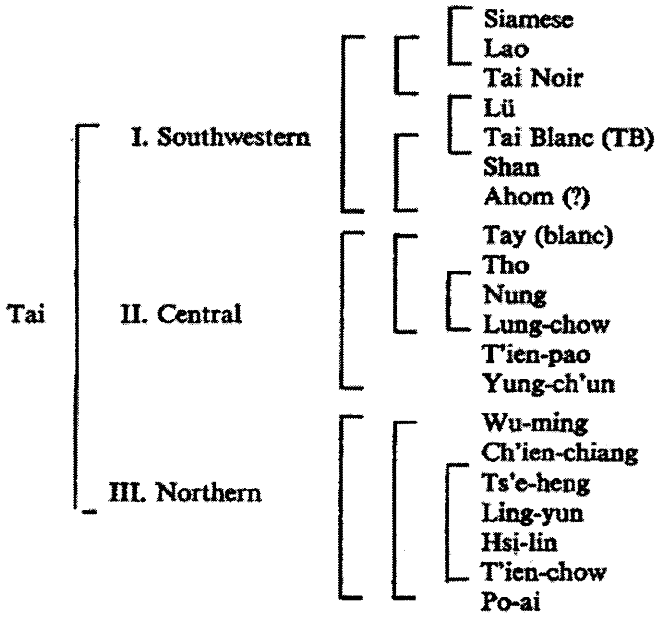


Figure 6.1: Tai subgrouping from Li (1960).

The most common alternate view (Chamberlain 1975) proposes the same three groups, but favors a binary division of Proto-Tai into Proto-NT on the one side and the common ancestor of both Proto-CT and Proto-SWT, sometimes referred to as Proto-SCT, on the other. This view has never been firmly established, however.

The most significant challenge to Li's conventional classification, and the current state of the art in Tai subgrouping, is Pittayaporn (2009), who conducted extensive additional fieldwork in Vietnam and China, resulting in a new phonological reconstruction of Proto-Tai based on the segmental Comparative Method. The result is very different phoneme inventories of Proto-Tai, and a drastically different tree. Pittayaporn found that SWT was the only valid subgroup of three proposed by Li (Pittayaporn 2009: 7), though he places it much lower tree. SWT corresponds to Pittayaporn's Group Q, which is a subgroup within his Group A. See Figure 6.2 for Pittayaporn's tree.

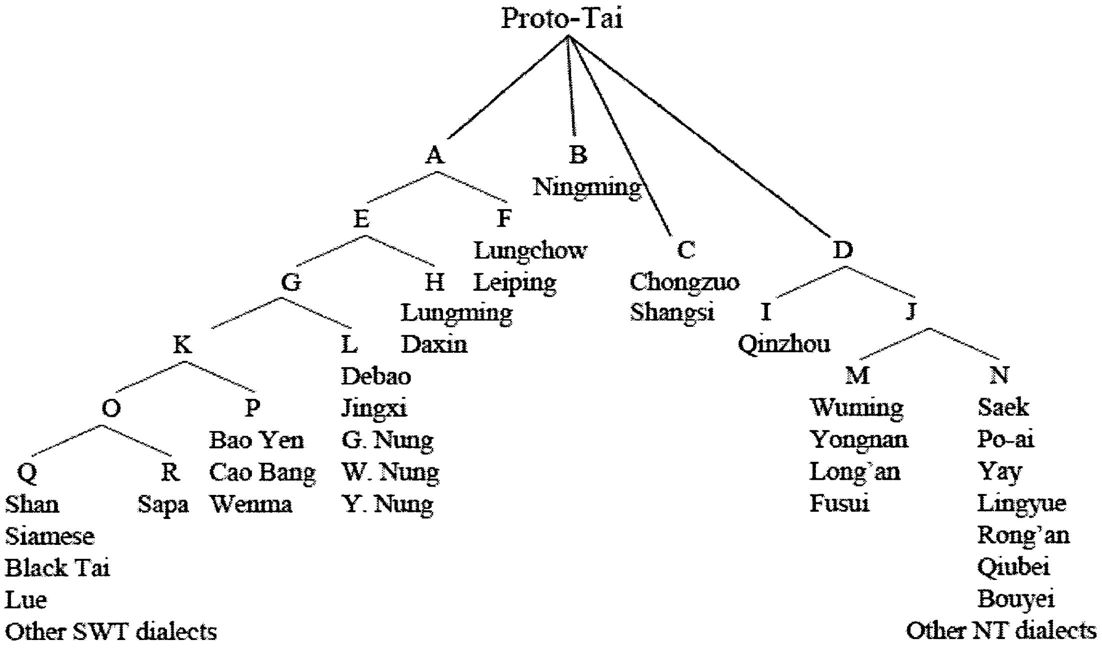


Figure 6.2: Tai subgrouping from Pittayaporn (2009).

6.3.1 Internal structure of Southwestern Tai

Previous studies of the internal structure of the SWT subgroup, and of Tai dialectology more generally, have relied on three types of evidence: (i) lexical evidence, looking at the distribution of words within members of the subgroup to identify characteristic lexemes and define groups therefrom, (ii) tonal evidence, using patterns of shared splits and mergers to group languages together, or (iii) segmental evidence, using the traditional Comparative Method to observe patterns of shared sound changes between languages. This section examines past proposals regarding the internal structure of SWT in terms of these three types of evidence. Only those proposals that have specific implications for TK are considered, as some proposals (e.g. Haas 1957; Brown 1965) have not dealt with TK at all, presumably due to a lack of available data.

The problem of doculect names

Inconsistency in naming can make it difficult to assess and verify past claims. On the larger level, terms like "Shan" and "Tai" are used generically, which can obscure significant variation. Further down the tree, the use of lect names is at best confusing, and the endonym for any given Tai language is frequently just "Tai." The names in the literature often give a false impression of specificity, making it impossible to positively identify two varieties as the same based on the name cited in the literature alone. Color-based names like Black Tai, White Tai, and Red Tai are derived from the color of traditional native dress, which we should not expect to be a good predictor of genetic language affiliation. In my fieldwork in Khamti Township, Myanmar, a neighboring group referred to both their own language and the local TK variety as 'Red Tai,' despite important classificatory differences between the two, and no connection with doculects in Vietnam described with the same name by Gedney. Another point of confusion is Tai Nuea, also frequently spelled Tai Nüa or Tai Nɯa, which simply means "Northern Tai." It is usually an exonym, and its application thus dependent on relative geography. What others call Dehong or Chinese Shan is also often identified as Tai Nuea, further confusing the picture. There is no simple solution for this problem, and often lacking the original data used by authors, we can only interpret their proposals as best as we can with whatever evidence is provided.

Chamberlain (1972, 1975)

Chamberlain 1975 is a revision of his earlier proposal, Chamberlain 1972a. In both papers he proposes two top-level branches of SWT, dubbed the P group and the PH group. These names derive from a characteristic sound change from the Proto-SWT voiced initials (*b, *d, *g), i.e. whether they simply devoiced (P group), or devoiced and became aspirated (PH group).

Though not mentioned explicitly, TK fits into Chamberlain's P group. Within the P group, Chamberlain originally proposed that all P languages exhibited a 123-4 split in all four of its proto-tones, and that this was characteristic of the P group. In 1975, Chamberlain revised his assessment of the P group, proposing instead a two-way split based on the A tone: A1-23-4 vs. A123-4. This revision was based on additional data from doculects he called Nuea, Tai Mao, and Tse Fang. Chamberlain identifies Tse Fang as "probably what has been referred to as Chinese Shan" (1975: 50). Chamberlain's 1975 tree is reproduced in Figure 6.3.

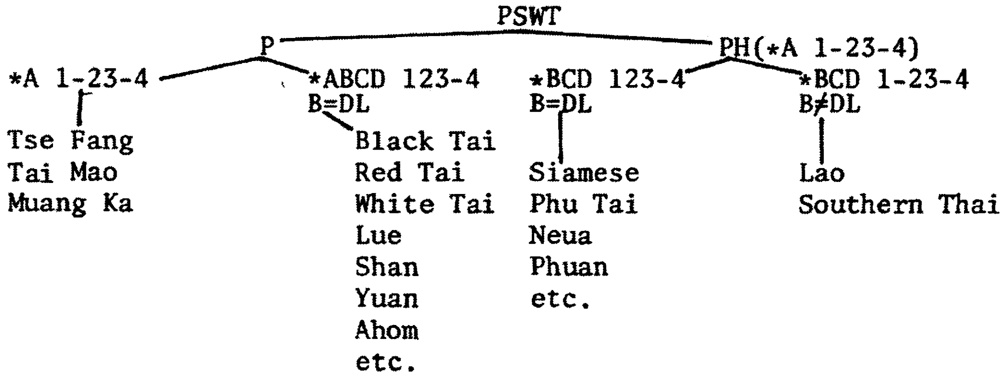


Figure 6.3: Southwestern Tai subgrouping from Chamberlain (1975).

Although he labels the newly defined split A1-23-4, Chamberlain is grouping A1-23-4 and A1-234 together, which other scholars have written as A1-23(4), as it is the A1 split that is considered diagnostic. Since then many additional doculects also confirm this tone split as characteristic of one subgroup within SWT. Others of Chamberlain's proposed tonal diagnostics do not hold up to scrutiny, as

Hartmann (1980)

The alignment proposed by Hartmann (1980) is a three-way subdivision based primarily on shared tonal splits: Upper SWT, Middle SWT, and Lower SWT. Upper SWT covers a geographic swath from the uppermost part of Myanmar east into southwestern China; Middle SWT corresponds roughly with Shan State, Myanmar and northern Thailand; and Lower SWT encompasses Laos and the rest of Thailand. Though Hartmann believes the evidence is less clear for the unity of Lower SWT, with respect to Upper and Middle SWT he states: "There seems to be little doubt about the unity of these two subgroups if examined from the standpoint of tonal splits." Notably, Hartmann's proposal does not account for Tai varieties of Northeast India, presumably as he did not have data for them. The varieties Hartmann identifies as representative of each subdivision are presented in Table 6.1.

Upper SWT	Middle SWT	Lower SWT
Shan		
Tai Nuea		
Lue (Chiang Rung)	Shan (Kengtung)	Luang Prabang
Lue (Chiang Tung)	Khuen (Kengtung)	Sam Neua
White Tai	Lue (Muong Yong)	Vientiane
Red Tai	Chiang Rai	Savannakhet
Black Tai	Chiang Mai	Loei
Western Nung	Nan	Roi-Et
Nung	Phrae	Ubon
Lung Chow	Phayao	Khorat
Ning Ming	Tak	Bangkok
Wuming	Uttaradit	Chumphon
Puyi South		Nakhon Si Thammarat
Chuang		

Table 6.1: Divisions within SWT languages according to Hartmann (1980).

Jonsson (1991)

Jonsson (1991) explicitly rejects lexically based subgrouping as unsound, and proposes a two-way SWT division based on shared phonological innovations:

First: Thai, Lao, Red Tai, Burmese Shan, Khamti, Lue, Tai Nuea, Ahom, Southern Thai

Second: Black Tai, White Tai, Chinese Shan (Tai Mao)

Robinson (1994)

Robinson III (1994) claims five shared innovations—three tonal and two segmental—to argue for Tai Nuea and TK in a separate branch apart from the rest of SWT:

Branch 1: Tai Nuea, Khamti

Branch 2: Burman Shan, other Southwest Tai

A similar claim is made in Luo (2001), discussed in 6.3.1, but with different evidence.

Edmondson and Solnit (1997)

Edmondson and Solnit (1997) review past work on tones in Shan and present evidence from tone shapes and tone splits for a three-way division in the Shan languages, which they take to include the languages of upper Myanmar and adjacent southern China, but excluding TK and the Tai languages of northeast India. In subsequent work, Edmondson (2008) presents a survey of segmental data and some tone data from dozens of additional locations throughout China and Myanmar in support of this three-way division, but terming them TK, Northern Shan, and Southern Shan.

Kullavanijaya and L-Thongkum (1998)

Kullavanijaya & L-Thongkum (1998) gathered SWT data from 42 locations in Thailand, Laos Vietnam and China, supplemented with Phake and Khamti data from India. They use shared tonal splits and mergers to posit a six-way division in a northern tier of Southwestern Tai (apparently synonymous with Chamberlain's P group):

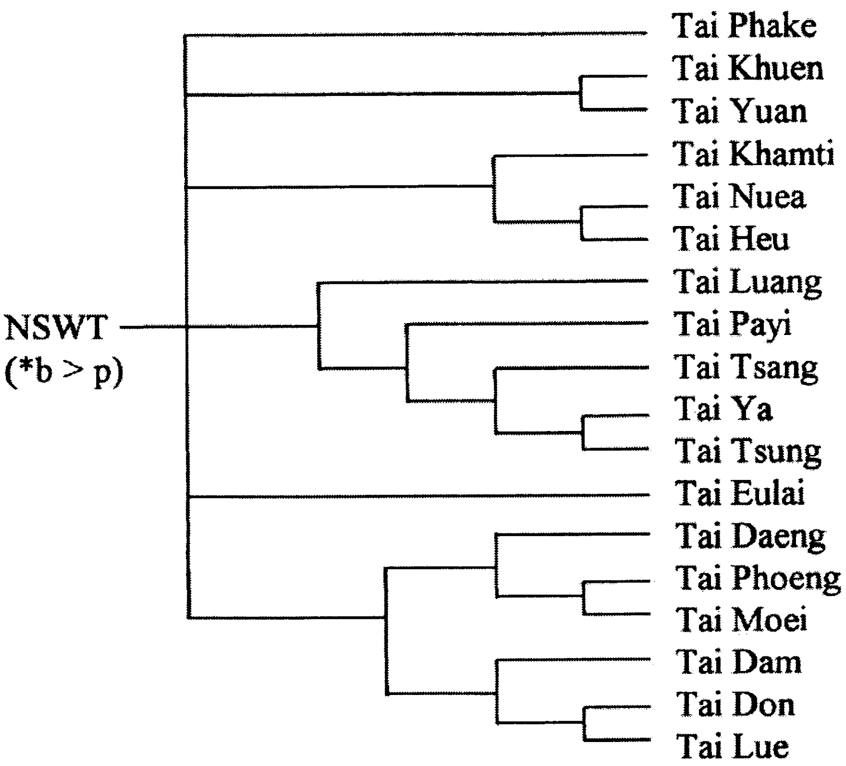


Figure 6.4: Proposed Northern SWT tier from Kullavanijaya & L-Thongkum (1998).

Their method does not capture intermediate levels relatedness for the top six divisions, for instance putting Tai Phake as its own divergent branch of NSWT based on having A1-234, B123-4, and C123-4 splits, while the branch that contains TK and Tai Nuea has A1-23-4, B123-4, and C123-4, differing only in the behavior of the A4 category and an additional A23-B4 merger seen in the TK cluster. This may be a naïve interpretation of surface tonal evidence, not distinguishing likely retentions from parallel but independent tonal innovations in different members of the group.

Below the six main branches that they use a mix of tonal and segmental differences to make further divisions. However, because they take tone as the primary evidence for subgrouping, segmental sound correspondences are irregular and perhaps unlikely. For instance, in Tai Payi, recorded in Yunnan, forms cited by Kullavanijaya & L-Thongkum indicate that *ʔd > d has occurred, and yet on the tonal evidence it is grouped as most closely related to Tai Tsang, Tai Ya, and Tai Tsung, which all exhibit a *ʔd > l sound change (1998: 283).

Luo (2001)

Rather than a northern division with SWT, Luo (2001) presents phonological and lexical evidence for a fourth top-tier division of Proto-Tai, which he terms Northwestern Tai (NWT), consisting of languages generally held to be part of the SWT stock. The primary piece of phonological evidence Luo presents is an /x/   /s/ alternation in words that have been reconstructed to Proto-Tai *xr-. Luo believes this is a remnant of a historical sound change that differentiates Dehong from Southwestern Tai languages. On lexical evidence, Luo identifies Dehong cognates of several words from the other Tai branches that Li (1977) claimed did not exist in SWT languages. These Dehong reflexes such as /tau/ 'shuttle of a loom', or /sam/ 'dirty'. Luo thus believes that Dehong supplies a lexical "missing link" between the non-SWT branches, leading to his proposal of a new branch.

The only language Luo firmly claims in his NWT proposal is Dehong, but he states that his proposed new branch "may include Northern Shan varieties such as Khamti, and perhaps Southern Shan varieties in Myanmar as well" (2001: 186). This is similar to the proposal of Robinson III (1994), who posited a relationship between TK and Tai Nuea separate from the rest of SWT based on five shared innovations, one shared tone split, two shared tone mergers, and two segmental mergers.

6.4 Tai Khamti tones

Early documentation work on TK, Needham (1894) and Grierson (1904), did not mark lexical tone at all, though they did include descriptions of the tones. This may have been influenced by the fact that tone was not marked in the Shan script in that time period. Regardless, we can gain no information for this category from those sources.

However, an earlier record of TK does record tone. This account was published by

Robinson (1849) using data from an American missionary, Nathan Brown. Comprising approximately 200 words gathered in Assam, Robinson is the first of three major sources on TK tone that this paper employs. The second data source on TK tone comes from Harris (1976). It is a wordlist of approximately 800 words from a village in Lohit District, of what is now Arunachal Pradesh Province in Northeast India. Finally, the third key source of TK tone data is from my own fieldwork in Khamti Township, Myanmar.

6.4.1 Tones of 19th century Indian TK

Robinson (1849: 312) describes three tones in TK as follows:

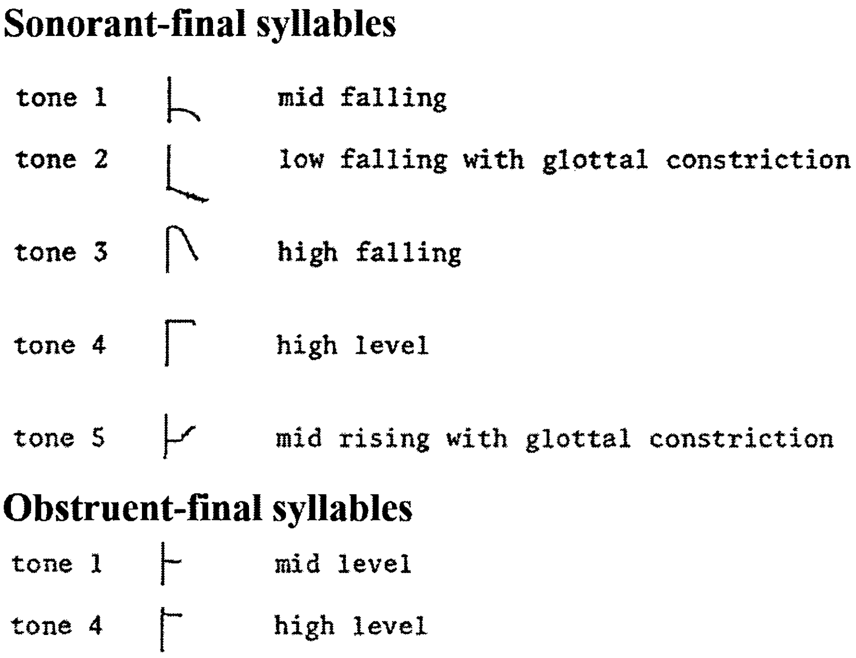
> "Thus má, for instance (with the rising tone) signifies a dog; má (the Italic m denoting the falling tone) signifies to come; while the same syllable, with an abrupt termination, or a sudden cessation of the voice at the end of it, mạ́, denotes a horse."

Morey (2005b) also notes a fourth tone used in the wordlist but not described by Robinson. Morey's reconstruction of the 1849 TK tonal system is presented in Figure 6.5.

	A	B	C	DS	DL
1	1 Rising	4 Level	3 Glottal	1	
2	2 Falling				
3					
4			?? falling, glottalized	2	

Figure 6.5: Historical tone categories of 1849 Indian TK (Morey 2005b: 191).

Some uncertainties remain in this reconstruction, as some of the 20 slots in the Gedney box had no corresponding vocabulary, and 13 of the 20 boxes had fewer than 10 tokens in the wordlist. Paucity of the data notwithstanding, Morey has gone to considerable lengths to identify a useful historical signal amidst the noise. He notes considerable typographic



Sonorant-final syllables

tone 1	mid falling
tone 2	low falling with glottal constriction
tone 3	high falling
tone 4	high level
tone 5	mid rising with glottal constriction

Obstruent-final syllables

tone 1	mid level
tone 4	high level

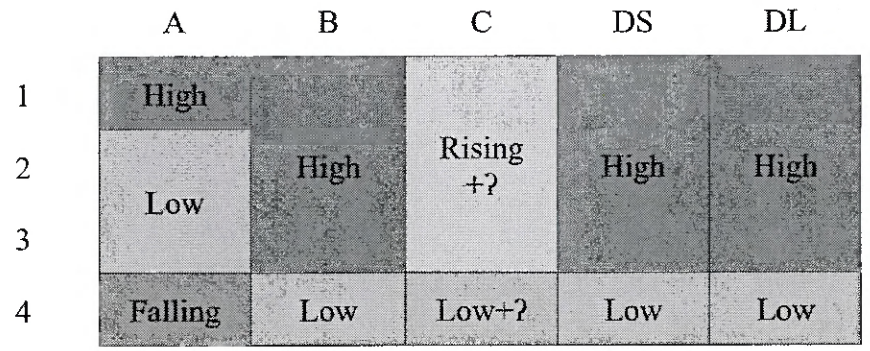
Figure 6.6: Chart of Tai Khamti tones from Harris (1976).

inconsistency due to the error-prone combination of italicization and subscript dots (Morey 2005b: 189). In this reconstruction, the A column is the most reliably reconstructed, as it was best represented in the wordlist. This is par for the family, where the number of native A tone words tends to be slightly more than the number of B and C tone words combined.

6.4.2 Tones of modern Indian TK

Harris (1976) presents the tonal system of TK in Arunachal Pradesh as five lexical tones on sonorant-final syllables, and two allotones on obstruent-final syllables. This is reproduced in Figure 6.6:

When analyzed according to the (Gedney 1972) tone box, we can construct a chart of the historical tone categories, and thus the historical splits and mergers, for Indian TK, as presented in Table 6.7.



	A	B	C	DS	DL
1	High	High	Rising +?	High	High
2	Low				
3					
4	Falling	Low	Low+?	Low	Low

Figure 6.7: Historical tone categories of Indian TK.

Splits and mergers that have been taken as characteristic of TK in previous work include (i) the A1-23-4 tone split, (ii) the A1 = B123 tonal merger, (iii) an A23 = B4 merger, and (iv) B = DL.

6.4.3 Tones of Chindwin TK

Unlike Indian TK, Chindwin TK has only four tonal contrasts on open and sonorant-final syllables, and two allotones on obstruent-final syllables. The tones of consultant SAM, a 38-year-old male, and LSAT, a 75-year-old male, both natives of Khamti Township, are presented in Figure 6.8.

Both speakers exhibited four distinct tonemes, with speaker intuitions of tone categories confirmed instrumentally in Praat. The most significant difference in their tone systems is in the realization of the rising tone. SAM, the younger speaker, exhibited a clear rise, equivalent to a 35 on the Chao (1930) tone scale, while the rising tone of LSAT was closer to 45 or even very nearly a 55 high level. Despite their different surface forms, both tones have the exact same lexical coverage, and are thus allotones of the same tonal category.

The tones presented in Figure 6.8 are averages of a minimum of 30 tokens of the citation form of each tone, pronounced on an open syllable. Recordings were segmented using Praat (Boersma 2019), followed by demarcating the rimes of each syllable through analysis of the waveform and spectrogram. A Praat script was used to measure F0 values

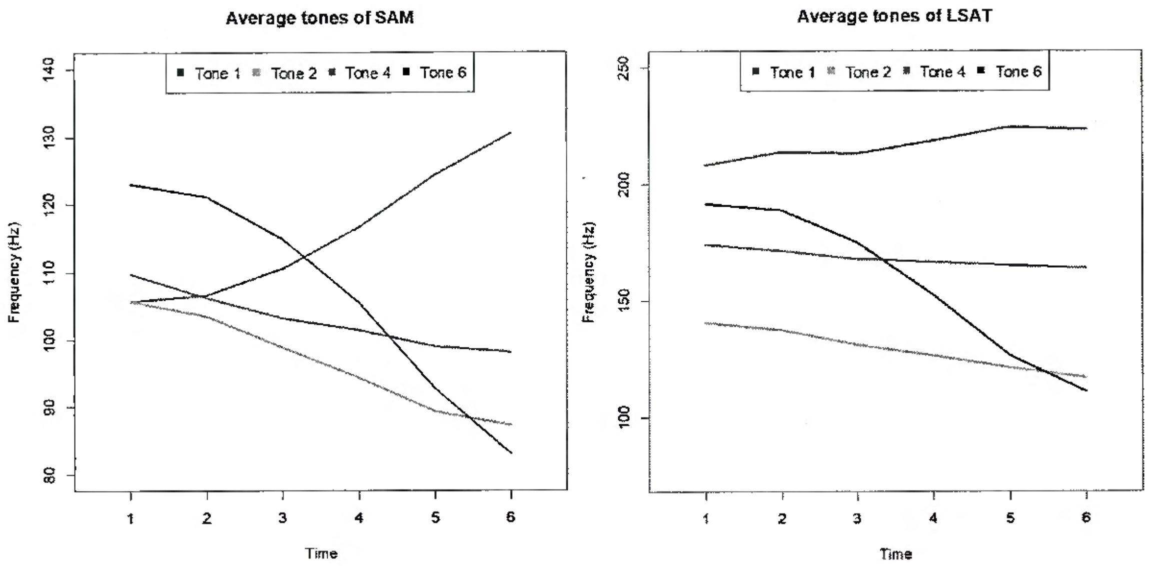


Figure 6.8: Average tones of consultants SAM, age 38, and LSAT, age 75.

at six intervals throughout each syllable DiCanio (2007), and visualized with R. The labels Tone 1, Tone 2, Tone 4, Tone 6 seen in Figure 6.8 are a simple numbering system used for convenience. They are not sequential because they are named based on the tone marks used in Chindwin TK orthography. Unlike other varieties of TK, the speaker community of Chindwin TK adopted a modified version of Shan script, and since it has fewer tones they use only a subset of the tone marks, hence the non-sequential tone numberling.

Examining the lexical coverage of each tone according to the Gedney system, Chindwin TK historical tone categories are as seen in Table 6.9

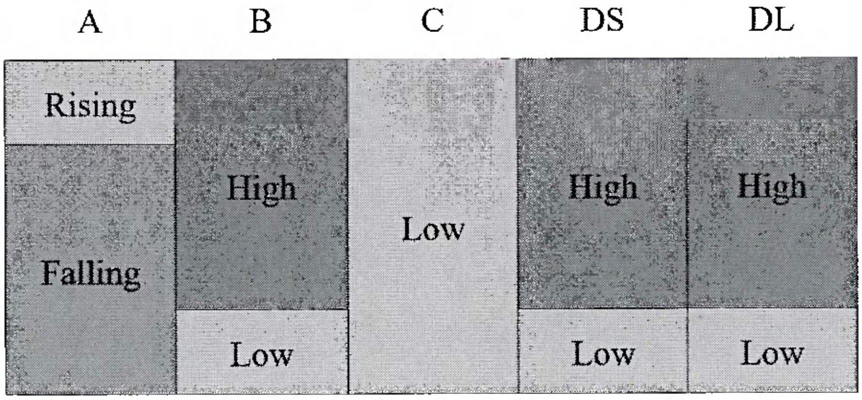


Figure 6.9: Historical tone categories of Chindwin TK.

6.4.4 Comparing tonal systems of Indian TK and Chindwin TK

In this section I examine the differences in tonal development between the modern tone systems of modern Chindwin TK and Indian TK, and 19th century Indian TK. Since tone splits have been used as diagnostic in determining genetic subgrouping, and past analyses involving TK have been based solely on Indian tonal data, it is important to establish which splits and mergers are retained from their common ancestor, and which are more recent innovations.

As discussed in depth in Chapter 5, it is important to keep in mind that what we are comparing here is not the phonetic tones, but rather the pattern of splits and mergers, i.e. how the historical tone categories conditioned by loss of segmental onset contrasts became the tonal systems as documented. The phonetic plausibility of a particular split may serve as additional corroborating evidence, but are not stable or reliable enough to form the core of the analysis. The splits are regular and systematic within the conditioning environment, but the phonetic realization of the same historical category can vary widely across even closely related languages.

Similarity and difference

While the phonetic realization is not identical, the tonal categories corresponding to the B and D proto-tones are identical between Indian TK and Chindwin TK. The B123 and the D123 categories are all a high or mid-high level tone, while the B4 and D4 categories are a mid-to-low falling tone.

The real value in comparing Indian TK and Chindwin TK comes in resolving the discrepancies in their tonal systems and considering what conditions were necessary in order for those discrepancies to arise. In this manner we can reconstruct the tonal system of their common ancestor, which we can term Proto-TK.

A tones

The first significant discrepancy is in the A series. In modern Indian TK, there is a three-way split between A1, A23, and A4, while both 1849 Indian TK and Chindwin TK have two tones, representing A1 and A234, as seen in Table 6.10.

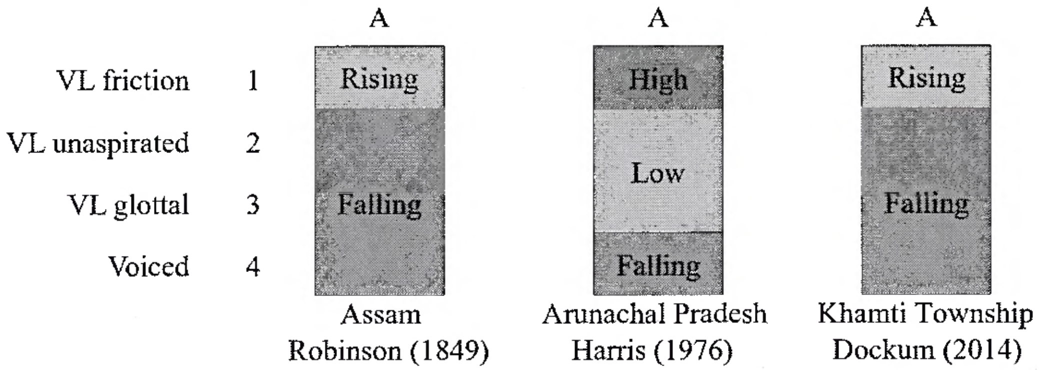


Figure 6.10: Splits in A-series tones for Indian TK and Chindwin TK.

The three-way split in the A series in the Harris (and Weidert) data is one of the core features that previous subgrouping proposals have been based upon. The division seen in the 1849 Indian data and the modern Chindwin TK is unusual because, as discussed in §4.5, the split of A123 from A4 is characteristic of the Great Tone Split, and yet neither TK dialect has it. That split happened hundreds of years before TK varieties would have begun to diverge from each other, and in fact the conditioning environment for it no longer exists, as the voiced stops from the proto-language merged with the voiceless unaspirated consonants across the board in TK. In the absence of a conditioning environment that could select for the correct lexical subset, there would be no way for A4 to have split off from Proto-TK. We can thus conclude that A23 must have merged with A4 in both 1849 Indian TK and Chindwin TK subsequent to their divergence from Indian TK, though we cannot yet say whether it was likely a single tone change or two parallel ones. We can also reconstruct a three-way split in the A category in their nearest common ancestor.

The fact that the older Indian TK looks more like modern Chindwin TK in this respect also deserves some comment. We must not forget that this is a reconstruction by Morey based upon a short wordlist of typographically problematic evidence. As such, there are a few possible explanations: (1) The A234 merger is a retention, and the two varieties are more closely related to each other than either is to modern Indian TK; (2) Morey's reconstruction is not correct, either due to an error of analysis or typographical error in the original document caused by the error-prone system of tone marking; (3) since A23 and A4 would have both had falling contours, only from slightly different starting pitch heights, the person recording the wordlist could not distinguish the two tones; or (4) the mergers happened independently and coincidentally.

Of these possibilities, it is difficult to know which to favor. But since the A234 merger certainly cannot be reconstructed back to Proto-TK, the correct reconstruction is an A1-23-4 tripartite division, as seen in modern Indian TK. Furthermore, the proto-tone for the A1 category can likely be reconstructed as a rising tone, as discussed in the next section.

B tones

The next discrepancies that must be resolved are those that arise in the historical B tone category. These are twofold: first, the equivalence between the A1 tone and the B123 tone in modern Indian TK, which is not reflected in either of the other two varieties; second, the merger in the B category seen exclusively in Morey's reconstruction of Robinson. These are as shown in Table 6.11.

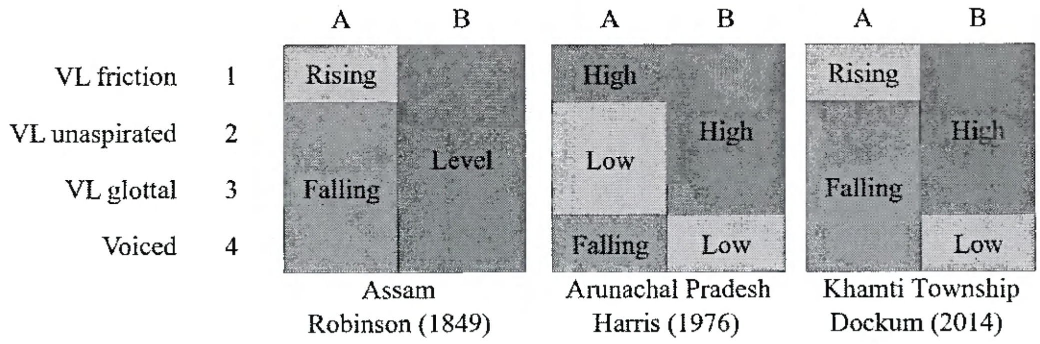


Figure 6.11: A1 = B1 merger seen only in modern Indian TK.

Using the logic of the Tonal Comparative Method, similar to that used in the previous section, the only possible analysis here is that there has been a merger in Indian TK, as the A-B category distinctions are original to Proto-Tai. There is no way such a split could have occurred in the older Indian or modern Burmese data, as the environment that gave rise to the original tonogenesis in Proto-Tai would have disappeared many centuries earlier. Ruling out the high level tone as being retained from Proto-TK also tells us that the likely A1 proto-tone was a rising tone. And a merger between a high level tone and a high rising tone is also phonetically very plausible, as they would share a pitch target.

As for the second problem in the B series, the B1234 merger in the older Indian TK, Morey notes that this merger is found nowhere else in Northeast India, and only attested in some Lao dialects (2005b: 192). While it is certainly not impossible that this merger had taken place in Assam at that time, it does not affect the larger analysis of what the common ancestor of these three varieties must have looked like. In this case it seems at least as likely that the data is simply too noisy and scarce on this point than that such the merger can be reliably claimed to have occurred.

C tones

The final area of discrepancy to consider is in the C series of the tone box. Both varieties of Indian TK exhibit a split between C123 and C4, and both are glottalized, as opposed to Chindwin TK, which has only a single low falling tone for the entire C category, as seen in Table 6.12.

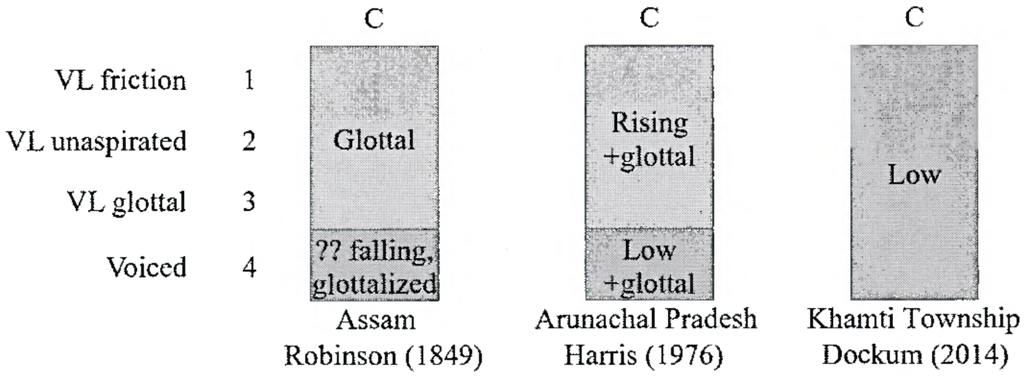


Figure 6.12: Splits in C-series tones for Indian TK and Chindwin TK.

This merged C tone in Chindwin TK does indeed exhibit frequent glottalization, especially on open vowels, but it is has merged with unglottalized mid-low falling tone that is the modern Indian TK reflex of the historical B4 and D4 tones. Glottalization is expected, as it is glottal phonation that has been proposed by scholars such as Sagart (2004) and Pittayaporn (2009) as the original phonation contrast that caused the genesis of the C tone series, prior to subsequent splits.

Since the C category has merged in Chindwin TK, so that now B4 = C1234 = D4, it might appear devoid of useful information for purposes of reconstructing. I would propose that a difference in the salience of phonetic cues for the C category developed at some point after Chindwin TK and Indian TK were separated by migration. The glottalization likely became the most salient contrastive cue for this category in Chindwin TK, leading to a collapse of the pitch contrast into a single glottalized low falling tone. This tone then further merged with the unglottalized low falling tone, bringing the total tonal inventory down to just four lexical tones.

Meanwhile, Indian TK retains what is likely the original tonal categories, and quite possibly the same phonetic realization as well. Thus I would posit that in the C category, the Proto-TK tones were C123, realized as a rising tone with glottal constriction, and C4, a mid-low falling glottalized tone.

6.4.5 Uniting the analysis

From the five-tone system of modern Indian TK, and the four-tone system of Chindwin TK, after considering each discrepancy between them we arrive at a completed reconstruction of the tonal system of their nearest common ancestor, which I have been referring to as Proto-Tai Khamti. Based upon the principles of the Tonal Comparative Method, it must have had a total of six lexical tones, as shown in Table 6.13.

<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>DS</th><th>DL</th></tr>
<tr><td>1</td><td>Rising</td><td rowspan="3">High Level</td><td rowspan="3">Rising +glottal</td><td rowspan="3">High Level</td><td rowspan="3">High Level</td></tr>
<tr><td>2</td><td rowspan="2">Mid/low Falling</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td><td>High Falling</td><td>Mid/low Falling</td><td>Low +glottal</td><td>Mid/low Falling</td><td>Mid/low Falling</td></tr>
</table>

Figure 6.13: Reconstructed tonal system of Proto-TK.

This reconstruction of the Proto-TK tones has all of the same tonal categories as Morey's reconstruction of what he terms Proto-Assam/Dehong/Northern Myanmar Tai, and which he dates to the 13th century (2005b:196), prior to their geographic dispersal and linguistic divergence. This reconstruction does differ from Morey in that he does not reconstruct the B and D categories as being allotonic. He reconstructs the B123 tone as 'level', and B4 as 'low level', while D123 as 'high' and D4 as 'low'. The B=D merger could thus be a subsequent innovation that may be useful for subgrouping as well, but given how prevalent the B=D equivalency is throughout the Tai languages, it is a good candidate for being a relatively old retention and not an innovation. So the more likely conclusion is that it Morey's reconstruction is incorrect on this point, unless evidence arises that a B=D merger is a frequent sound change.

6.5 Why only mergers since Proto-TK?

In going from Proto-TK to both modern TK doculects treated in this paper, I have reconstructed exclusively tone mergers. There have been no subsequent splits. This raises questions about the phonological plausibility of this trajectory of change, and its potential motivations, given the tendency in phonological change for the total number of segmental contrasts to remain roughly constant. Just as with unconditioned segmental mergers, tone mergers decrease the number of phonemic contrasts in the language, and thus increase homophony in the language.

However, we can account for the tonal mergers we see across the modern Tai languages if we view them not as creating new homophony, but rather as the culmination of the process of rephonologization (Hyman 1976) that started with initial tonogenesis.

For Proto-Tai (PT), Li (1977) reconstructed 36 simple onsets and 31 clusters vs. Pittayaporn's 36 simple onsets and 27 clusters (2009a:70,139). On non-obstruent final syllables, PT had the A, B, and C tones. Moving forward in time to Proto-Southwestern Tai (PSWT), Jonsson reconstructed 37 simple onsets and 12 clusters in Proto-SWT (1991:52-53), versus Pittayaporn's 39 simple and 15 clusters (Pittayaporn 2009b:121). While there remains disagreement on the exact number of contrasts, the trajectory of the phonology is clear: there was a drastic decrease in the number of complex onsets between PT and PSWT. This coincides with the Great Tone Split, which would have doubled the number of tonal contrasts in the language. This collapse in onset distinctions represents a large scale rephonologization, a rebalancing of the system following the Great Tone Split and the loss of former cluster contrasts, but resulting in the total phonemic complexity found in the system remaining roughly constant.

With the number of theoretically possible tonal categories that would have existed after subsequent splits in the voiceless proto-initials, corresponding to rows 1, 2 and 3 of the Gedney box, then we would expect to see a drastically reduced number of onsets in the modern daughters of PSWT. And in fact this is exactly what we do see, both in TK and elsewhere: Chindwin TK and Indian TK both have 16 simple onsets.

Viewing the phonology of a tonal language as a system where the number of tonal contrasts and segmental contrasts are in balance one another, such variation may represent pressure for or against tonal mergers even among closely related languages. Thus the modern varieties may be viewed as continuing the process of rephonologization that began with tonogenesis. And under this view, more recent tonal mergers would not cause undue homophony at all, if they are maintaining the balance of the number of phonemic and tonemic contrasts.

6.6 Tonal and segmental evidence: conflict or concord?

Subgroupings based primarily or exclusively on tonal evidence, e.g. Chamberlain (1972a, 1975) and Kullavanijaya & L-Thongkum (1998), have not received wide adoption, some combination of the skepticism against tonal evidence that exists in historical linguistics generally, and specific problems with their proposals. Tones change as a system, and thus the tonal categories are reconstructible, even if the phonetics of proto-tone systems remain obscure. Indeed, that may be the rule as we move forward with more tone box reconstructions, not the exception. Nonetheless, in this section I consider one of the characteristic tonal splits found in TK and nearby languages, the A1-234, and compare its distribution within SWT to some other characteristic sound changes in SWT languages.

6.6.1 *ʔd > n, *ʔb > m, *f > ph

From the perspective of the segmental evidence, among the shared sound changes that we can use to compare varieties of TK, and determine their closest relatives, are *ʔd > n, *ʔb > m, and *f > ph. The distribution of modern reflexes of some of the languages examine for this study are given in Table 6.14.

	Chindwin TK (Dockum 2014)	Indian TK (Harris 1976)	Indian TK (Weidert 1979)	Namti Khamti (Inglis 2004)	Homalin Khamti (Inglis 2004)	Khamti Khamti (Inglis 2004)	Putao Khamti (Inglis 2004)	Phake (Morey 2015)	Khamyang (Morey 2015)	Aiton (Morey 2015)	Ahom (Morey 2015)	Black Tai (Hudak 2008)	Shan (Hudak 2008)	Shan (Jonsson 1991)	Black Tai (Jonsson 1991)	Tai Nuea (Jonsson 1991)	Tai Nuea (Zhou 2001)	Tai Hongjin (Zhou 2001)	Tai Ya (Zhou 2001)	Tai Nuea (Yu 1980)	Siamese (Jonsson 1991)	Lao (Jonsson 1991)	White Tai (Hudak 2008)	Black Tai (Hudak 2008)	Nong Khai (Hudak 2008)	Chieng Hung (Hudak 2008)	Lue Muong Yong (Hudak 2008)	Chiang Mai (Hudak 2008)
*ʔd	n	n	n	n	n	n	n	n	n	n	n	l/n	l	f	l	f	l	f	f	f	d	d	d	d/l	d	d	d	d
*ʔb	m	m	m	m	m	m	m	m	m	m	m	b	w	m	b	v,m	v	v	v	m	b	b	b	b	b	b	b	b
*f	ph	ph	ph	ph	f[8]	ph	ph	ph	ph	ph	ph	f	f	f	l	f	f	f	f	f	f	f	f	l	f	f	f	f

Figure 6.14: Modern reflexes of Proto-SWT *ʔd, *ʔb, and *f.

As the table shows, all doculects of TK pattern together, including Chindwin TK as I have described it, both major modern sources for Indian TK, and all four locations in northern Myanmar recorded in Inglis' (2004) comparative wordlist. Furthermore, TK patterns most closely with the Tai languages of Northeast India, including Phake, Aiton, Khamyang, and Ahom. This would be an unremarkable finding if we still had data from only Indian TK to go on. However, given the disparate migrations that separated the TK-speaking groups from each no later than the 18th century (Gogoi 1971:21), the fact that they still pattern so close together segmentally only serves to highlights the need for a better understanding of TK tones before we base any subgrouping alignment on them. It also strengthens the case for the reconstruction of the Proto-TK tone system proposed above.

6.6.2 A tone splits

Divisions in the A category demonstrate the viability of tonal evidence for historical subgrouping arguments. Since TK itself shows variation between A1-23-4 and A1-234, we must treat these two together, since it was explained in section above that A234 must be a recent merger, anyway, as the A123-4 split was one of the results of the Great Tone Split.

Thus any SWT subgrouping argument that uses A1-23-4 vs. A1-234 as evidence for anything other than a recent divergence must be incorrect. Using the same set of languages as in the previous table, the distribution of an A1-A23(4) split is given in Table 6.15.

Chindwin TK (Dockum 2014)	Indian TK (Harris 1976)	Indian TK (Weidert 1979)	Namti Khamti (Inglis 2004)	Homalin Khamti (Inglis 2004)	Khamti Khamti (Inglis 2004)	Putao Khamti (Inglis 2004)	Phake (Morey 2015)	Khamyang (Morey 2015)	Aiton (Morey 2015)	Ahom (Morey 2015)	Shan (Jonsson 1991)	Tai Nuea (Jonsson 1991)	Tai Nuea (Yu 1980)	Tai Nuea (Zhou 2001)	Black Tai (Hudak 2008)	Shan (Hudak 2008)	Black Tai (Jonsson 1991)	Tai Hongjin (Zhou 2001)	Tai Ya (Zhou 2001)	White Tai (Hudak 2008)	Black Tai (Hudak 2008)	Chieng Hung (Hudak 2008)	Lue Muong Yong (Hudak 2008)	Chiang Mai (Hudak 2008)	Siamese (Jonsson 1991)	Lao (Jonsson 1991)	Nong Khai (Hudak 2008)
1	1	1	1	1	1	1	1	1	1	?[9]	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1

Figure 6.15: Distribution of A1-23(4) tone split in SWT languages (1 = present, 0 = absent).

Some language names have been reordered to groups like values together, but the original coloring that marked segmental clustering from the previous table is retained to enable easier comparison. Table 6.15 shows that there is indeed a very clear cluster of A1-23(4) splits that coincide with the segmental distribution seen above. This split does partially bleed over into southern China, however, as Tai Nuea varieties described by both Zhou (2001) and Yu (1980) also display it. Finally, there is what can only be an unrelated cluster of A1-23(4) splits in Thailand and Laos.

Thus in our test case, the tonal evidence seems to broadly corroborate the segmental evidence, even if it identifies a slightly larger set. We could thus conclude that the A1-23(4) split is a strong piece of evidence for genetic subgrouping, even if it should not be taken as proof uncritically. Following this method, each tone merger or tone split may be considered in turn, just as with any segmental sound correspondence.

6.7 Implications for subgrouping

The debate over the position of TK may not be fully settled, but tonal evidence and segmental evidence are in agreement that TK patterns most closely with other Tai languages of Northeast India. This evidence casts doubt on claims by Robinson III (1994) and Luo (2001) that Tai Nuea and TK form a separate branch.

Further down the tree, another implication for subgrouping is that some tonal diagnostics used to classify TK are not necessarily reliable without consideration of the larger phonology. For example, the A23 = B4 merger of Indian TK does not exist in Chindwin TK, but it does exist in Khamyang (Morey 2005b) and Tai Nuea (Zhou 2001, Yu 1980). I have reconstructed it in the Proto-TK tone system, but lacking that, if we simply took the modern tone categories at face value, we could only conclude that Chindwin TK was a more distant relative of Indian TK than Tai Nuea, when the segmental (and anthropological) evidence contradicts such a conclusion.

In addition, the A1 = B123 merger that is used to identify Indian TK must be a recent innovation, within the last 200 years. The fact that Indian TK has this merger in common with its geographical neighbor Aiton, as documented by Morey (2005b), but not Chindwin TK, may mean that Indian TK and Aiton may have undergone this merger at the same time, or perhaps one under the influence of the other. This likelihood is especially strong given that Diller also records a tonal system in a second variety of Áiton which lacks this merger (1992:18). Furthermore, high-level divisions of SWT based on tonal splits, as posited by Chamberlain (1975), Hartmann (1980), and Kullavanijaya and L-Thongkum (1998), are insufficient without corroborating segmental evidence. At the same time, a purely segmental subgrouping would be sure to miss important pieces of the puzzle as well. It is only when the two are analyzed as parts of the same whole, applying the logic of the Comparative Method to both domains, that we can resolve many of the issues that have caused problems for determining the internal structure of SWT. Since they relied almost exclusively on tonal evidence, these proposals have already been challenged by subsequent comparative work that focuses on the segmental evidence, but this reconstruction further confirms that Proto-TK cannot have existed in Chamberlain's A123-4 division of the P group, where he had grouped both Shan and Ahom.

The combination of tonal and segmental evidence presented here also supports the "northern tier" of SWT languages proposed in Edmondson and Solnit (1997) and Edmondson (2008).

6.8 Conclusion

This chapter makes use of new data from Chindwin TK to demonstrate the Tonal Comparative Method, applying the same type of reasoning to tonal splits and mergers as has long been done for segmental evidence as part of the Comparative Method. While the TK situation is one of the very smallest pieces of the unresolved subgrouping puzzle in SWT, it is exactly this level of careful, bottom-up, granular reconstruction that will help us to eventually resolve the internal structure of SWT and the Tai branch in its entirety.

Thus the Tai Khamti case study contributes towards the development of an improved understanding of the mechanisms of tone change and how it fits into the larger system of historical phonology. Equipped with the Tonal Comparative Method, and the large amounts of data on SWT languages that has been documented, we will be able to make subgrouping proposals that are more transparent and testable, better motivated on empirical grounds, and ultimately stronger.

Chapter 7

Conclusion

In many ways the Comparative Method has remained unchanged since the early 19th century, and seems like a monolithic method. The reality is of course, that while the core of the method—systematic identification of regular sound correspondences—is ingenious in its simplicity, the conventional wisdom of how best to apply the method, and how to reconstruct, has accrued gradually and continually over the course of the history of linguistics. Certainly it remains true to this day, to paraphrase Alexander Pope, that a little knowledge of the Comparative Method is a dangerous thing. As historical linguists, we hope to distinguish ourselves from hobbyists who occasionally fill our inboxes or bend our ear at some event, convinced they have evidence of some implausible macro family that linguists have somehow failed to notice. And yet the history of historical linguistics is also filled with the discarded phylogenies and reconstructions of many great linguists. That is, of course, the nature of scientific discovery. The Comparative Method has enjoyed two centuries of building up of a body of conventional wisdom. This conventional wisdom deals with such things as the reasons why some sound change may have occurred, or the likelihood whether some innovation is shared or parallel. Everything taken as conventional wisdom today was once not so, and some things current now will be discarded before long.

The study of tone diachrony is yet in its youth. We have come a long way—from the earliest observations by de Lacouperie (1887) on its compensatory nature, to the explicit connection between tonal contrasts and their segmental forebears by Li (1943), to the cogent explication of the basic principles of tonogenesis by Haudricourt (1954). We are finally beginning to be equipped with the scale of data and the sophistication of methods to build up a conventional wisdom for tone change that is broadly applicable at least throughout the Sinospheric Tonbund, and perhaps to languages worldwide.

At the International Conference on Sino-Tibetan Languages and Linguistics in 1983, held in Seattle, Gedney gave an after-dinner conference address titled "Confronting the Unknown" (Gedney 1985). At that point he was already retired, four decades removed from commencing his doctoral studies on Thai linguistics with Franklin Edgerton, Bernard Bloch, and Isidore Dyen at Yale. On that occasion, Gedney was frank about how little we still knew about how lexical tone had so thoroughly conquered the region. He spoke of "the great wave of tonal splits that swept across Southeast Asia and the Far East," and detailed various gaps in our knowledge, including its geographic extent, its familial extent, and the dates involved.

Gedney described the need for cooperative investigation, and his belief that with the combined knowledge in the room they could sort it out, if they only could just all sit down to do it together. He discussed a plan to write to each person in attendance and piece together the answer to the major questions for the next year's conference. In the next breath, however, he said that he was no longer up to the demands of the task, and welcomed anyone else who wished to do so.

We have certainly made progress in our understanding of tone diachrony in the decades since that meeting. And yet no one quite took up the torch that Gedney offered. The Tonal Comparative Method is in some respects my way of doing that, but it is also a reiteration of Gedney's plea for a collaborative effort. The fortunate circumstances of tone in the Tai family—its regularity, its relative youth, and its long written record—provide us with the model whereby the method can be laid out. But it is only in years to come that we will fully refine it as we probe the true extent of its reach.

Appendix A

Appendix 1

A.1 Tai language documentation theses compiled

The following is a list of the theses on the documentation of various aspects of Tai languages, compiled from extensive library research in Thailand. The majority of these theses are in written in Thai, with a minority written in English. See the References section for complete bibliographic details.

Theses focusing on tone systems of Tai doculects

Koowatthanasiri 1981; Debavalya 1983; Ratanadilok Na Phuket 1983; Sritararat 1983; Kopprayun 1986; Taengko 1987; Malaichalern 1988; Tanlaput 1988; Chinchest 1989; Prapaipet 1989; Aruneeung 1990; Tingsabadh 1990; Panroj 1991; Sawangwan 1991; Hanpanich 1992; Kobsirikarn 1992; Nualjansaeng 1992; Banditkul 1993; Pornsib 1994; Krisnapan 1995; Komontha 1996; Sittiprapaporn 1997; Akharawatthanakun 1998; Anusurain 1998; Khumdee 2000; Worawong 2000; Namwang 2001; Pratankiet 2001; Khamrueangsi 2002; Khemkhaeng 2002; Khotchanthuek 2002; Nasanee 2002; Akharawatthanakun 2003; Kewkasem 2003; Tanprasert 2003; Pintasaard 2004; Kitivongprateep 2005; Lertthana 2005; Kongthong 2006; Saeng-ngam 2006; Sitthi 2006; Thawarorit 2006; Bunmee 2007; Kantong 2007; Chaimano 2009; Soiyana 2009; Awirutthiyothin 2010; Canilao 2010; Yooyen

2013

Theses focusing on more general Tai language documentation

Jantanakom 1983; Pungpawpun 1984; Ampornpan 1986; Chativong 1986; Siriwisitkun 1986; Chaokhamin 1988; Sukpiti 1989; Mapawongse 1979; Sungkep 1983; Boonsner 1984; Charoenphol 1985; Poo-Israkij 1985; Tisapong 1985; Chotecheun 1986; Eam-eium 1986; Maneewong 1987; Thongrat 1988; Subcharoen 1989; Thongphiew 1989; Rakpaet 1998; Thianthaworn 1998; Teeranuwat 2002; Lengtai 2009; Suppasin 2011; Thavorn 2013; Plungsuwan 1981; Kitprasert 1985; Pimpa 1986; Lamchiagdase 1984; Ploykaew 1985; Weesakul 1983; Chulkeeree 1991; Udomphan 2000; Suwanratt 1991; Sila 1975; Massupong 1982; Paiboonwangcharoen 1984; Sungvanthrup 1991; Bencha 2000; Nakorn 2000; Chanavong 1980; Yensamut 1981; Narkphong 1982; Tanyong 1983; Ratanapraseart 1985; Peamphermphoon 1986; Nakpuntawong 1987; Kongsuwan 1988; Beadnok 1989; **?**; Ninjinda 1989; Panarat 1990; Sombatmaungkan 1990; Charoenvalaya 1991; Vaitayavanich 1991; Yoojaroensuk 1991; Chawsuan 1994; Praphin 1996; Seangsrichan 1998; Matchikanang 1999; Jitbanjong 2002; Pumma 2003; Wuttheerapon 2004; Laksanawong 2008; Tippol 1988; Kummun 1992; Unakornsawat 1993; Hasonnary 2000; Soongsumaln 2002; Chummalee 2010; Junlawan 2011; Khwanritti 1987; Wetchasit 1987; Thumsaro 1993; Rittiwong 1997; Angsuwiriya 2003; Suwanmusik 2004; Rakmoh 2007; Plodkaew 2008; Jidlang 2012; Tebpawan 2012; Chai-arun 1998; Kamwachirapitak 2005; Sornjitti 2007; Poonpholwattanaporn 2010; Sutadarat 1978; Shen 2003; Osatananda 1997; Sumransook 1995; Withayasakpan 1979b; Sukpreedee 1988; Arpakul 1995; Sakdanuwatwong 1995; Mahaphunthong 1996; Phantachat 1983; Thongmark 1983; Khamboonchoo 1985; Saeneetontikul 1985; Ache 1986; Maryprasith 1992; Choophan 2004; Boonabha 1969; Chanaingoon 1970; Dumruks 1970; Rinprom 1977; Panka 1980; Somnuk 1982; Senisrisant 1983; Jurjanad 1987; Petsuk 1978; Worachin 2009; Boonsawasd 2012; Punthong 1979; Buranasing 1988; Manoosawet 1993; Patpong 1997; Rakpaet 2010

Bibliography

Abouheif, E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research* 1(8). 895–909.

Abramson, Arthur S & Theraphan Luangthongkum. 2009. A fuzzy boundary between tone languages and voice-register languages. *Frontiers in phonetics and speech science* 149–155.

Ache, Duangjai. 1986. *Lexical geography of Southern Thai spoken in Surat Thani and Nakhon Si Thammarat.* Chulalongkorn University M.A. thesis. [In Thai].

Akharawatthanakun, Phinnarat. 1998. *Tones in Lao, Nyo and Phutai in That Phanom district, Nakhon Phanom province.* Chulalongkorn University M.A. thesis. [In Thai].

Akharawatthanakun, Phinnarat. 2003. *Tone change: a case study of Lao languages.* Chulalongkorn University M.A. thesis. [In Thai].

Ampornpan, Nipa. 1986. *Description of Tai Lue dialect of Pakha subdistrict, Thawangpha district, Nan province.* Silpakorn University M.A. thesis. [In Thai].

Angsuwiriya, Chanokphorn. 2003. *Study of vocabulary in Southern Thai as spoken by residents in urban, suburban, and rural areas of Hat Yai district, Songkhla province.* Thaksin University M.A. thesis. [In Thai].

Anusurain, Ekkapol. 1998. *Synthesis of tones and vowels in Thai open syllables using microphonemes.* Chulalongkorn University M.A. thesis. [In Thai].

Arpakul, Pruchya. 1995. *Language map of Pattani and Narathiwat provinces*. Chulalongkorn University M.A. thesis. [In Thai].

Aruneeung, Arunee. 1990. *Variation in the falling tone in Bangkok Thai according to speaker age*. Chulalongkorn University M.A. thesis. [In Thai].

Awirutthiyothin, Tamjai. 2010. *Acoustic characteristics of consonant, vowel and tone in Standard Thai with southern accent in comparison with Standard Thai and Southern Thai*. Chulalongkorn University M.A. thesis. [In Thai].

Banditkul, Panchanit. 1993. *Tones in monosyllabic and disyllabic words in the central Thai dialect of Prachuap Khiri Khan*. Chulalongkorn University M.A. thesis. [In Thai].

Beadnok, Chuucheep. 1989. *Lexical study of Khorat dialect*. Silpakorn University M.A. thesis. [In Thai].

Bencha, Nutthapong. 2000. *Description of Tai Lue dialect in Thung Mok village, Bang Mang district, Chiang Muan district, Phayao province*. Silpakorn University M.A. thesis. [In Thai].

Benedict, Paul K. 1942. Thai, kadai, and indonesian: a new alignment in southeastern asia. *American Anthropologist* 44(4). 576–601.

Benedict, Paul K. 1973. Tibeto-burman tones, with a note on teleo-reconstruction. *Acta Orientalia* 35. 127–138.

Benedict, Paul K. 1975. *Austro-thai language and culture, with a glossary of roots*. Human Relations Area Files.

Bengtson, John D. 1994. Edward sapir and the "sino-dene" hypothesis. *Anthropological Science* 102(3). 207–230.

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, the Data Citation and Attribution in Linguistics Group & the Linguistics Data Interest Group. 2018. The Austin principles of data citation in linguistics. `http://site.uit.no/linguisticsdatacitation/austinprinciples`. Version 1.0.

Birchall, Joshua, Michael Dunn & Simon J. Greenhill. 2016. A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics* 82(3). 255–284.

Blomberg, S. P., T. Garland & A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57(4). 717–745.

Bloomfield, Leonard. 1935. *Language*. Allen & Unwin.

Boersma, David, Paul Weenink. 2019. Praat: doing phonetics by computer [computer program]. `http://www.praat.org/`.

Boonabha, Oraphim. 1969. *Phonemes in Rayong Thai*. Chulalongkorn University M.A. thesis. [In Thai].

Boonkao, Phonthiph. 1989. *Lexical geography of Thai dialects in Mahasarakham province*. Silpakorn University M.A thesis. [In Thai].

Boonsawasd, Attasith. 2012. *A grammar of Bouyei*: Mahidol University Ph.D. dissertation.

Boonsner, Thepbangon. 1984. *Phonology of Yooy*. Mahidol University M.A. thesis.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard & Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science* 337(6097). 957–960.

Bouckaert, Remco R, Claire Bowern & Quentin D Atkinson. 2018. The origin and expansion of pama–nyungan languages across australia. *Nature ecology & evolution* 2(4). 741.

Bowern, Claire. 2012. The riddle of tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences* 279(1747). 4590–4595.

Bowern, Claire. 2018. Computational phylogenetics. *Annual Review of Linguistics* 4. 281–296.

Bowern, Claire & Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88. 817–845.

Bradley, Cornelius Beach. 1911. Graphic analysis of the tone-accents of the siamese language. *Journal of the American oriental society* 31(3). 282–289.

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupillai. 2008. Automated classification of the world′ s languages: a description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung* 61(4). 285–308.

Brown, J. Marvin. 1965. *From Ancient Thai to modern dialects*. Social Science Association Press of Thailand.

Brown, J. Marvin. 1975. The great tone split: did it work in two opposite ways? .

Brunelle, Marc & James Kirby. 2015. Re-assessing tonal diversity and geographical convergence in Mainland Southeast Asia. In *Languages of Mainland Southeast Asia: The State of the Art*, 82–110. Mouton de Gruyter.

Bryant, D. & V. Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* 21(2). 255–265.

Bunmee, Aimkamon. 2007. *Tone geography of Khammueang Lampang.* Mahidol University M.A. thesis.

Buranasing, Anchulee. 1988. *An analysis of lexical change among three generations in Thai Song dialect.* Mahidol University M.A. thesis. [In Thai].

Campbell, Lyle. 2003. *How to show languages are related: methods for distant genetic relationship* 264–284. Blackwell.

Canilao, Kritsana. 2010. *Tonal geography of the provinces of Central Thailand.* Mahidol University M.A. thesis.

Chacon, Thiago Costa & Johann-Mattis List. 2015. Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationship* (3). 177–203.

Chai-arun, Praphai. 1998. *Lexical geography of Sa Kaeo province.* Burapha University M.A. thesis. [In Thai].

Chaimano, Kanita. 2009. *Tone variation of Tai Lue spoken in Thailand.* Mahidol University M.A. thesis.

Chaisri, Jiraporn. 1984. *Verb modifiers in Lamphun dialect.* Chulalongkorn University M.A thesis. [In Thai].

Chamberlain, James R. 1972a. The Origin of the Southwestern Tai. *Bulletin de Amis du Laos* 7(8). 233–244.

Chamberlain, James R. 1972b. Tone borrowing in five northeastern dialects. In Jimmy G. Harris & Richard B. Noss (eds.), *A conference on tai phonetics and phonology*, 43–46. Mahidol University.

Chamberlain, James R. 1975. A new look at the history and classification of the Tai languages. *Studies in Tai linguistics in honor of William J. Gedney* 49–66.

Chamberlain, James R. 1979. Tone in tai: A new perspective .

Chanaingoon, Wiroonrat. 1970. *Phonemes of the Petchabun dialect (used in Tambon Lom Sak, Amphoe Lom Sak) / Wiroonrat Chanaingoon.* Chulalongkorn University M.A. thesis. [In Thai].

Chanavong, Pramechit. 1980. *Words and syntax in Southern Thai dialects: Chumphon and Nakhon Si Thammarat.* Silpakorn University M.A. thesis. [In Thai].

Chang, Charles B & Yao Yao. 2007. Tone production in whispered mandarin. *UC Berkeley PhonLab Annual Report* 3(3).

Chang, Kun. 1947. Miáoyáoyŭ shēngdiào wèntí [On the tone system of the Miao-Yao languages]. *Bulletin of the Institute of History and Philology* 16. 93–110. [In Chinese].

Chang, Kun. 1953. On the tone system of the miao-yao languages. *Language* 29(3). 374–378.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. doi:10.1353/lan.2015.0005.

Chao, Yuen-Ren. 1930. A system of tone-letters. *Le Maître Phonétique* 45. 24–27.

Chaokhamin, Sanae. 1988. *Description of Tai Yuan dialect at Tha Chang subdistrict, Sao Hai district, Saraburi province.* Silpakorn University M.A. thesis. [In Thai].

Charoenphol, Nipa. 1985. *Phonological description of the Thai dialect at Chanthaburi.* Mahidol University M.A. thesis.

Charoenvalaya, Kanokphorn. 1991. *Lexical distribution of some lexical items in Phetchaburi Province.* Silpakorn University M.A. thesis. [In Thai].

Chativong, Jinda. 1986. *Description of Lao Khrang dialect of Huay Duan subdistrict, Dontum district, Nakhon Pathom province.* Silpakorn University M.A. thesis. [In Thai].

Chawsuan, Em-orn. 1994. *Comparative lexicon of Lao dialects of Nakhon Pathom province, Thailand, and Borkaew province, Laos.* Silpakorn University M.A. thesis. [In Thai].

Chinchest, Pornsri. 1989. *Lao Ngaew tones in citation forms and in connected speech.* Chulalongkorn University M.A. thesis. [In Thai].

Ching, Wai-Ki & Michael K. Ng. 2006. *Markov chains: Models, algorithms and applications.* Springer.

Choophan, Sirirat. 2004. *Lexical variation in the Thai dialect of Koh Samui island by speakers' area of residence and age.* Chulalongkorn University M.A. thesis. [In Thai].

Chotecheun, Siwaporn. 1986. *Phonology of Nan Thai with comparisons to Phrae.* Mahidol University M.A. thesis.

Chulkeeree, Uthumphorn. 1991. *Lexical geography of Thai dialects in Phichit province.* Mahidol University M.A. thesis. [In Thai].

Chummalee, Penprapa. 2010. *Phonology of the Phu Thai dialect, Thamcharoen subdistrict, So Phisai district, Bueng Kan province.* Silpakorn University M.A. thesis. [In Thai].

Collins, Jeremy. 2016. The role of language contact in creating correlations between humidity and tone. *Journal of Language Evolution* 1(1). 46–52. doi:10.1093/jole/lzv012.

Court, Christopher. 1998. The "gedney boxes" for southwestern tai: the need for a dʔ column, .

Cuirong, Yu. 2009. *Dǎi yǔ jiǎn zhì [Description of Dai]* (Zhōngguó shào shù mínzú yǔyán jiǎn zhì cóngshū [Brief Descriptions of Chinese Minority Languages Series; revised edition] 3). [In Chinese].

Cushing, Josiah Nelson. 1871. *Grammar of the shan language.* American Baptist Mission Press.

Cushing, Josiah Nelson. 1887. *Grammar of the shan language, second edition.* American Baptist Mission Press.

Debavalya, Kesmanee. 1983. *Isogloss (tonal) between Central Thai and Southern Thai.* Chulalongkorn University M.A. thesis. [In Thai].

DiCanio, Christian. 2007. Pitch script for praat. `https://www.acsu.buffalo.edu/~cdicanio/scripts/Get_pitch.praat`.

DiCanio, Christian T. 2012. Cross-linguistic perception of itunyoso trique tone. *Journal of Phonetics* 40(5). 672–688.

Diller, Anthony V.N. 1996. Thai orthography and the history of marking tone. *Oriens Extremus* 39(2). 228–254.

Dockum, Rikker. 2018. Undocumented labor: how old fieldwork sheds new light on tai tone system diversification. doi:10.5281/zenodo.1136317.

Dockum, Rikker & Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16. 35–54.

Dockum, Rikker (collector). 2014-2018. Tai Khamti of the Upper Chindwin River Valley (RD2), digital collection managed by PARADISEC [open access]. doi:10.26278/5b520716aa520.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *Wals online*. Max Planck Institute for Evolutionary Anthropology. `https://wals.info/`.

Dumruks, Vilaiwan. 1970. *Phonemes of the Nakhon Si Thammarat dialect*. Chulalongkorn University M.A. thesis. [In Thai].

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson & Neele Becker. 2011a. Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronica* 28(3). 291–323.

Dunn, Michael, Simon J Greenhill, Stephen C Levinson & Russell D. Gray. 2011b. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345). 79.

Dunn, Michael & Angela Terrill. 2012. Assessing the lexical evidence for a Central Solomons Papuan family using the Oswalt Monte Carlo Test. *Diachronica* 29(1). 1–27.

Eam-eium, Chalong. 1986. *Phonology of Phuan at Hatsiaw subdistrict, Si Satchanalai district, Sukhothai province*. Mahidol University M.A. thesis.

Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2019. Ethnologue: Languages of the world (twenty-second edition). SIL International. `http://www.ethnologue.com`.

Edmondson, Jerold A. 2008. Shan and other northern tier southeast tai languages of myanmar and china: themes and variations. In *The tai-kadai languages*, 200–222. Routledge.

Egerod, Søren. 1961. Studies in thai dialectology. *Acta Orientalia* 26(1-2). 43–91.

Elsevier. 2012. Elsevier withdraws support for the research works act. `https://www.`

elsevier.com/about/policies/message-on-the-research-works-act. Accessed: 2019-06-01.

Enfield, Nicholas J. 2008. *A grammar of lao*, vol. 38. Walter de Gruyter.

Everett, Caleb, Damián E. Blasi & Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots 112(5). 1322–1327. doi:10.1073/pnas.1417413112.

Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1). 1–15.

Fox, Anthony et al. 2000. *Prosodic features and prosodic structure: The phonology of suprasegmentals*. Oxford University Press.

Fritz, S. A. & A. Purvis. 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic strength in binary traits. *Conservation Biology* 24(4). 1042–1051. doi:10.1111/j.1523-1739.2010.01455.x.

Gasser, Emily & Claire Bowern. 2014. Revisiting phonological generalizations in Australian languages. *Proceedings of the Annual Meetings on Phonology* doi:10.3765/amp.v1i1.17.

Gedney, William J. 1964. A comparative sketch of white, black and red tai. *The Social Science Review* 14. 1–47.

Gedney, William J. 1966. Linguistic diversity among tai dialects in southern kwangsi. In *41st meeting of the linguistic society of america* 43, .

Gedney, William J. 1967. Future directions in comparative tai linguistics .

Gedney, William J. 1972. A checklist for determining tones in Tai dialects. *Studies in linguistics in honor of George L. Trager* 423–37.

Gedney, William J. 1985. Confronting the unknown: Tonal splits and the genealogy of tai-kadai. *Linguistics of the Sino-Tibetan area: The state of the art* 116–124.

Golla, Victor. 1984. The sapir-kroeber correspondence: Letters between edward sapir and al kroeber, 1905-1925.

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

Gray, Russell D, David Bryant & Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559). 3923–3933.

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.

Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of austronesian expansion, 2000. *Nature* 405. 1052–1055.

Greenhill, Simon J, Russell D Gray et al. 2009. Austronesian language phylogenies: Myths and misconceptions about bayesian computational methods. *Austronesian historical linguistics and culture history: a festschrift for Robert Blust. Canberra: Pacific Linguistics* 375–397.

Grierson, G. A. 1904. Linguistic Survey of India, Vol. II: Mon-Khmer and Siamese-Chinese Families (Including Khassi and Tai).

Haas, Mary R. 1957. The tones of four tai dialects. *Bulletin of the Institute of History and Philology* 29. 817–826.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019a. About languoids. Max Planck Institute for the Science of Human History. `https://glottolog.org/glottolog/glottologinformation`.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019b. Glottolog 4.0. Max Planck Institute for the Science of Human History. `https://glottolog.org/`.

Hanbo, Liao. 2016. *Tonal development of tai languages*. Payap University M.A. thesis.

Hanpanich, Sasitorn. 1992. *Statistical distribution of consonants, vowels and tones within Thai syllables existing as words and word constituents*. Chulalongkorn University M.A. thesis. [In Thai].

Harris, Jimmy G. 1976. Notes on Khamti Shan. In *Tai linguistics in honor of Fang-Kuei Li*, 113–141.

Hartmann, John F. 1980. A model for the alignment of dialects in Southwestern Tai. In *10th International Conference on Sino-Tibetan Languages and Linguistics, Washington DC published in JSS*, vol. 68, 72–86.

Hasonnary, Siwaporn. 2000. *Phonology of Lao Luang Prabang: a comparative study with Lao Khrang in Thachin river basin and Lao Dan Say*. Silpakorn University M.A. thesis. [In Thai].

Haudricourt, André-Georges. 1946. Restitution du karen commun. *Bulletin de la Société de Linguistique de Paris* 42(1). 103–11.

Haudricourt, André-Georges. 1954. De l'origine des tons en vietnamien [translated 2018]. *Journal Asiatique* 242. 69–82.

Hock, Hans Henrich & Brian D. Joseph. 2009. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*, vol. 218. Walter de Gruyter.

Hoenigswald, Henry M. 1993. On the history of the comparative method. *Anthropological Linguistics* 54–65.

Holden, Clare J & Russell D. Gray. 2006. Rapid radiation, borrowing and dialect continua in the bantu languages. *Phylogenetic methods and the prehistory of languages* 19. 31.

Holland, B. R., K. T. Huber, A. Dress & V. Moulton. 2002. Delta plots: A tool for analyzing phylogenetic distance data. *Molecular biology and evolution* 19(12). 2051–2059.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker & Pamela Brown. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52(6). 000–000.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42(3-4). 331–354.

Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen & Niklas Wahlberg. 2013. Cultural and climatic changes shape the evolutionary history of the uralic languages. *Journal of Evolutionary Biology* 26(6). 1244–1253.

Hruschka, Daniel J, Simon Branford, Eric D Smith, Jon Wilkins, Andrew Meade, Mark Pagel & Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1). 1–9.

Hudak, Thomas J. 2008. *William J. Gedney's comparative Tai source book* (Oceanic linguistics special publications no. 34). Honolulu: University of Hawai'i Press.

Hudak, Thomas John. 2004. William j. gedney's elicitation questionnaire. *Journal of the American Oriental Society* 124(3). 549–559.

Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23(2). 254–267.

Hyman, Larry M. 1976. Phonologization. In A. Juilland (ed.), *Linguistic studies offered to joseph greenberg, vol. 2*, 407–418.

Hyman, Larry M. 2013. Enlarging the scope of phonologization. *Origins of sound change: Approaches to phonologization* 3–28.

Hyman, Larry M. 2018. Presidential address at the 2018 annual meeting of the linguistic society of america.

Hyslop, Gwendolyn. 2007. Toward a typology of tonogenesis. In *7th biennial conference of the association for linguistic typology, paris, france*, .

Inglis, Doug. 2004. Preliminary wordlist and lexicostatistical analysis of Khamti Shan.

Janda, Richard D. & Brian D. Joseph. 2003. *On language, change, and language change; or, of history, linguistics, and historical linguistics* 3–180. Blackwell.

Jantanakom, Wanna. 1983. *Description of Tai Yai (Tai Aw) language in Mae Sai district, Chiang Rai province.* Silpakorn University M.A. thesis. [In Thai].

Jidlang, Athitthan. 2012. *Maintenance and variation in Southern Thai spoken by three generations of fishermen at Ban Yong Star community, Tha Kham subdistrict, Palian district, Trang province.* Thaksin University M.A. thesis. [In Thai].

Jitbanjong, Sarinya. 2002. *Lexical variation in Saek among three generations at Nawa district, Nakhon Phanom province.* Silpakorn University M.A. thesis. [In Thai].

Jones, John Taylor. 1842. *Brief grammatical notices of the siamese language.* Mission Press, Bangkok.

Jongman, Allard. 2013. Acoustic phonetics. *Oxford Bibliographies* doi:10.1093/OBO/9780199772810-0047.

Jonsson, Nanna L. 1991. *Proto Southwestern Tai*: SUNY Albany Ph.D thesis.

Junlawan, Taweeporn. 2011. *Phonology of Phithen dialect, Phithen subdistrict, Thung Yang Daeng district, Pattani province.* Silpakorn University M.A. thesis. [In Thai].

Jurjanad, Oratai. 1987. *Phonemes of Tai Yuan dialect in Sikhio district, Nakon Ratchasima province.* Chulalongkorn University M.A. thesis. [In Thai].

Kamwachirapitak, Papasara. 2005. *Lexical geography of Thai dialects in Chaiyaphum.* Thammasat University M.A. thesis. [In Thai].

Kantong, Ekapon. 2007. *Tonal variation in Chiang Mai Thai by age group.* Chulalongkorn University M.A. thesis. [In Thai].

Karlgren, Bernhard. 1915. 1926. *Études sur la phonologie chinoise* 1–4.

Kauffman, William G. 1993. *The great tone split and central karen.* University of North Dakota M.A. thesis.

Kewkasem, Praphaiphan. 2003. *Development of tones from the Jindamani textbook to modern Thai.* Srinakharinwirot University M.A. thesis. [In Thai].

Khamboonchoo, Chamlong. 1985. *Lexical study of Khammueang sub-dialects in Lampang.* Chulalongkorn University M.A. thesis. [In Thai].

Khamrueangsi, Pinkanok. 2002. *Classification of Nyo in Northeastern Thailand based on the tonal system.* Mahasarakham University M.A. thesis. [In Thai].

Khemkhaeng, Pornwali. 2002. *Tones of Nyo speakers of varying ages in Thakhonyang village, Kantharawichai district, Mahasarakham province.* Mahasarakham University M.A. thesis. [In Thai].

Khotchanthuek, Chintana. 2002. *Changes to tones in a multilingual community at Latbuakhao subdistrict, Sikhiw district, Nakhon Ratchasima province.* Mahasarakham University M.A. thesis. [In Thai].

Khumdee, Suttimas. 2000. *A tonal study to differentiate Banlad accents of Petchaburi province.* Chulalongkorn University M.A. thesis. [In Thai].

Khwanritti, Suphap. 1987. *Current Thai dialects in Songkhla.* Srinakharinwirot University, Songkhla M.A. thesis. [In Thai].

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa & Connie J. Mulligan. 2009. Bayesian phylogenetic analysis of semitic languages identifies an early bronze age origin of semitic in the near east. *Proceedings of the Royal Society B: Biological Sciences* 276(1668). 2703–2710.

Kitivongprateep, Sunisa. 2005. *Tones of koh samui thai dialect: variation by speaker age and residence.* Chulalongkorn University M.A. thesis. [In Thai].

Kitprasert, Chailert. 1985. *A tonal comparison of Tai dialects: Tak Bai group.* Mahidol University M.A. thesis.

Kobsirikarn, Rataya. 1992. *Tonal variation in the high-falling tone in Suphanburi Thai by some social variables.* Chulalongkorn University M.A. thesis. [In Thai].

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray & Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science* 5(3). 1–17.

Komontha, Manat. 1996. *Tone phonetics in Khorat Thai.* Thammasat University M.A. thesis. [In Thai].

Kongsuwan, Supavadee. 1988. *Lexical distribution of some lexical items in Loei province.* Silpakorn University M.A. thesis. [In Thai].

Kongthong, Ruangsuk. 2006. *Isoglosses (tonal) between Central Thai, Southern Thai, and Central-Southern Thai hybrid dialect: tonal variation by age group.* Chulalongkorn University M.A. thesis. [In Thai].

Koowatthanasiri, Kanjana. 1981. *Tones in Nyo.* Chulalongkorn University M.A. thesis. [In Thai].

Kopprayun, Suthida. 1986. *Tones in Tai Yoy.* Mahidol University M.A. thesis.

Krauss, Michael E. 1973. *Na-dene* 903–978. Mouton.

Krisnapan, Daranee. 1995. *Tonal study of connected speech: a case study of Phetchaburi Thai.* Chulalongkorn University M.A. thesis. [In Thai].

Kullavanijaya, Pranee & Theraphan L-Thongkum. 1998. Linguistic criteria for determining Tai ethnic groups: Case studies on central and south-western Tais. In *Tai Studies Proceedings*, 273–297.

Kummun, Wichit. 1992. *Phonology of Lao Khrang dialect at Banrai district, Uthai Thani province.* Silpakorn University M.A. thesis. [In Thai].

de Lacouperie, Terrien. 1887. *The languages of china before the chinese.* D. Nutt.

Ladefoged, Peter & Keith Johnson. 2011. *A course in phonetics, sixth edition.* Nelson Education.

Laksanawong, Mudjalin. 2008. *Comparative lexicon of Tai dialects in Sakon Nakhon province.* Silpakorn University M.A. thesis. [In Thai].

Lamchiagdase, Nanthariya. 1984. *Phonology of Tai Lue in Lampang province.* Mahidol University M.A. thesis.

Lee, Sean & 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.

Lengtai, Aggasena. 2009. *Shan phonology and morphology*. Mahidol University M.A. thesis.

Lertthana, Areewan. 2005. *Tone variation in falling tone by sex of speaker in Bangkok Thai and hearer's attitudes*. Thammasat University M.A. thesis. [In Thai].

Li, Fang-kuei. 1943. The hypothesis of a pre-glottalized series of consonants in primitive tai. *Bulletin of the Institute of Ethnology, Academia Sinica* (11). 177–188.

Li, Fang-Kuei. 1945. Some old chinese loan words in the tai languages. *Harvard Journal of Asiatic Studies* 333–342.

Li, Fang-kuei. 1954. Consonant clusters in Tai. *Language* 30(3). 368–379.

Li, Fang-kuei. 1959. Classification by vocabulary: Tai dialects. *Anthropological Linguistics* 1(2). 15–21.

Li, Fang-kuei. 1960. A tentative classification of Tai dialects. In *Culture in history: Essays in honor of Paul Radin*, .

Li, Fang-kuei. 1966. The relationship between tones and initials in tai. In Norman H. Zide (ed.), *Studies in comparative austroasiatic linguistics*, 82–88. London: Mouton Co.

Li, Fang-kuei. 1977. A handbook of comparative Tai. *Oceanic Linguistics Special Publications* (15). 1–389.

Longobardi, Giuseppe & Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* 119(11). 1679–1706. doi:10.1016/j.lingua.2008.09.012.

Longobardi, Giuseppe, Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini & Andrea Ceolin. 2013. Toward a syntactic phylogeny of modern Indo-European languages. *Journal of Historical Linguistics* 3(1). 122–152. doi:10.1075/jhl.3.1.07lon.

Low, James. 1828. *A grammar of the thai or siamese language.* Printed at the Baptist Mission Press. Sold by Messrs. Thacker and Company ....

Luo, Yongxian. 1997. *The subgroup structure of the Tai languages: a historical-comparative study.* Journal of Chinese Linguistics, Monograph Series No. 12.

Luo, Yongxian. 2001. The Hypothesis of a New Branch for the Tai Languages 11.

Macklin-Cordes, Jayden. 2015. *Phylogeny and phonotactics: quantifying historical signal in sequences of sound.* University of Queensland B.A. thesis.

Macklin-Cordes, Jayden L. & Erich R. Round. 2015. High-definition phonotactics reflect linguistic pasts doi:http://dx.doi.org/10.15496/publikation-8609.

Mahaphunthong, Challika. 1996. *Locating a dialect boundary between Eastern Southern Thai and Western Southern Thai using syllables with long vowels and final /k/ or /?/.* Chulalongkorn University M.A. thesis. [In Thai].

Malaichalern, Yajai. 1988. *Tones on the Thai dialects of Changwat Ang Thong and Phra Nakhon Si Ayutthaya.* Chulalongkorn University M.A. thesis. [In Thai].

Maneewong, Orapin. 1987. *Phonology of Lao Song in Phetchaburi and Nakhon Pathom provinces.* Mahidol University M.A. thesis.

Manoosawet, Chalermchai. 1993. *Lexical distribution in the area of the isogloss between Central Thai and Southern Thai: geographical and social variation.* Mahidol University M.A. thesis. [In Thai].

Mapawongse, Damrong. 1979. *Phonological description of the Nakhon Thai dialect.* Mahidol University M.A. thesis. [In Thai].

Maryprasith, Primrose. 1992. *Age-based variation of the linguistic transition area between Central Thai and Southern Thai: a lexical study.* Chulalongkorn University M.A. thesis. [In Thai].

Maspero, Henri. 1911. Contribution à l'étude du système phonétique des langues thaï [a contribution to the phonetic study of the thai languages]. *Bulletin de l'École Française d'Extrême-Orient* 19. 152–169.

Massachusetts Institute of Technology. 2013. Report to the president: MIT and the prosecution of Aaron Swartz. `http://swartz-report.mit.edu/`. Accessed: 2019-06-01.

Massupong, Chantra. 1982. *Grammar of the Tai dialect of Lung-Chow.* Silpakorn University M.A. thesis. [In Thai].

Matchikanang, Phra Sukhum. 1999. *Nyo lexicon and syntax at Thakhonyang village, Kantharawichai district, Mahasarakham province.* Silpakorn University M.A. thesis. [In Thai].

Matisoff, James A. 1970. Glottal dissimilation and the lahu high-rising tone: a tonogenetic case-study. *Journal of the American Oriental Society* 13–44.

Matisoff, James A. 1973. Tonogenesis in southeast asia. *Consonant types and tone* 1.

Matisoff, James A. 1985. New directions in east and southeast asian linguistics. In Graham Thurgood, James A. Matisoff & David Bradley (eds.), *Linguistics of the sino-tibetan area: The state of the art. papers presented to paul k. benedict for his 71st birthday*, 21–35. The Australian National University: Pacific Linguistics.

Matisoff, James A. 1990. On megalocomparison. *Language* 66(1). 106–120.

Matisoff, James A. 1991. Areal and universal dimensions of grammatization in lahu. *Approaches to grammaticalization* 2. 383–453.

Matisoff, James A. 2001. Genetic versus contact relationship: prosodic diffusibility in South-East Asian languages. *Areal diffusion and genetic inheritance: problems in comparative linguistics* 291–327.

Matisoff, James A. 2003. *Handbook of proto-tibeto-burman: system and philosophy of sino-tibetan reconstruction.* Univ of California Press.

Meillet, Antoine. 1914. Le problème de la parenté des langues. *Scientia* 15(35).

Mesoudi, Alex. 2011. Cultural evolution. *eLS* 1–8.

Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Sérgio Meira, Vivian Wauters & Zachary O'Hagan. 2015a. A bayesian phylogenetic classification of tupí-guaraní. *LIAMES: Linguas Indígenas Americanas* 15(2). 193–221. doi:10.20396/liames.v15i2.8642301.

Michael, Lev, Natalia Chousou-Polydouri, Zachary O'Hagan, Keith Bartolomei, Diamantis Sellis, Emily Clem & Erin Donnelly. 2015b. A bayesian phylogenetic internal classification of the tupí-guaraní family.

Ministry of Education of Thailand. 2016. 2016 educational statistics. `http://www.en.moe.go.th/enMoe2017/index.php/educational-statistics/educational-statistics-2016`. Accessed: 2019-06-01.

Moontuy, Sirirat. 2010. *Analysis of linguistic change in Yong word usage of three generations in Buak Khang sub-district, San Kamphaeng district, Chiang Mai province.* Chiang Mai University M.A thesis. [In Thai].

Morén, Bruce & Elizabeth Zsiga. 2006. The lexical and post-lexical phonology of thai tones. *Natural Language & Linguistic Theory* 24(1). 113–178.

Morey, Stephen. 2005a. *The Tai languages of Assam–a grammar and texts.* Pacific

Linguistics, Research School of Pacific and Asian Studies, the Australian National University.

Morey, Stephen. 2005b. Tonal change in the Tai languages of Northeast India. *Linguistics of the Tibeto-Burman Area* 28(2). 139–202.

Nakorn, Chen. 2000. *Morphology and syntax of Lao Khrang at Wanglao village, Nongkrot subdistrict, Muang district, Nakhon Sawan province.* Silpakorn University M.A. thesis. [In Thai].

Nakpuntawong, Panta. 1987. *Lexical geography of Uttaradit province.* Silpakorn University M.A. thesis. [In Thai].

Namwang, Pornsawan. 2001. *Tone in Northeastern Thai as spoken by Lao, Phu Thai and So of the communities at Naphiang sub-district, Kusuman District, Sakon Nakorn Province*. Thammasat University M.A. thesis. [In Thai].

Narkphong, Phongsiri. 1982. *Word classes in Khon Kaen dialect.* Silpakorn University M.A. thesis. [In Thai].

Nasanee, Kusuma. 2002. *Tones in Thai in tracheoesophageal speech: acoustic analysis and perception*. Chulalongkorn University M.A. thesis. [In Thai].

National Statistical Office of Thailand. 2019. `http://web.nso.go.th/`. Accessed: 2019-06-01.

Neamnark, Wisuttira. 1985. *Lamphun Yong phonology: a synchronic comparative study*. Chulalongkorn University M.A thesis. [In Thai].

Needham, J. F. 1894. *Outline Grammar of the Tai* (*Khamti*) *Language.*

Ngaorangsi, Kanchana, Phunphong Ngamkasem, Bancha Khucharoenphaibun & Arun Siprasert. 1982. Language map of Phitsanulok province. Tech. rep. Naresuan University. [In Thai].

Ninjinda, Nantaporn. 1989. *Lexical study of Nyo spoken in Sakon Nakhon, Nakhon Phanom and Prachin Buri.* Silpakorn University M.A. thesis. [In Thai].

Norman, Jerry. 1988. *Chinese.* Cambridge University Press.

Nualjansaeng, Janya. 1992. *Tones of the Thai dialect of Amphoe Muang Nakhon Pathom.* Chulalongkorn University M.A. thesis. [In Thai].

Orme, David, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac & Will Pearse. 2012. Caper: Comparative analyses of phylogenetics and evolution in r. *R package version 0.5* 2. 458.

Osatananda, Varisa. 1997. *Tone in Vientiane Lao*: University of Hawai'i Ph.D. thesis.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401(6756). 877.

Pagel, Mark. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10(6). 405–415.

Paiboonwangcharoen, Pimpan. 1984. *Description of Nakhonthai dialect.* Silpakorn University M.A. thesis. [In Thai].

Pallegoix, Jean-Baptiste. 1850. *Grammatica linguae thai.* Ex typographiâ collegii Assumptionis BMV.

Pallegoix, Johannes Baptista. 1854. *Dictionarium linguae thai sive siamensis, interpretatione latina, gallica et anglica illustratum.* Typograheum Imperatorium.

Panarat, Surat. 1990. *Lexical geography of Thai dialects in Lopburi province.* Silpakorn University M.A. thesis. [In Thai].

Panka, Kanchana. 1980. *Phonological characteristics of Lao dialects in Amphoe Mueang, Nakhon Pathom.* Chulalongkorn University M.A. thesis. [In Thai].

Pankhuenkhat, Ruengdet. 1978. *Yong phonology.*

Panrerk, Penchan. 2004. *The relationship between language and ethnicity in Lamphun province.* Chulalongkorn University M.A thesis. [In Thai].

Panroj, Piyachut. 1991. *Acoustic characteristics of tones in Bangkok Thai : Variation by age groups.* Chulalongkorn University M.A. thesis. [In Thai].

Patpong, Pattama. 1997. *The comparative study of lexical usage among three generations in Sukhothai dialect of Tambon Thungluang, Amphoe Khirimat, Sukhothai province.* Mahidol University M.A. thesis.

Peamphermphoon, Salub. 1986. *Lexical geography of Thai dialects in Buriram province.* Silpakorn University M.A. thesis. [In Thai].

Petsuk, Rasi. 1978. *General characteristics of the Khuen language.* Mahidol University M.A. thesis.

Phantachat, Wantanee. 1983. *Lexical study of regional varieties of Khammueang.* Chulalongkorn University M.A. thesis. [In Thai].

Pimpa, Wirat. 1986. *Tonal comparison of Northeastern Thai dialects in Khon Kaen province.* Mahidol University M.A. thesis.

Pintasaard, Rungwimol. 2004. *Development of tones *B and *C in Southwestern Tai languages.* Chulalongkorn University M.A. thesis. [In Thai].

Pittayaporn, Pittayawat. 2009. *The phonology of proto-Tai*: Cornell University Ph.D. thesis.

Pittayaporn, Pittayawat. 2016. Chindamani and reconstruction of thai tones in the 17th century. *Diachronica* 33(2). 187–219.

Pittayaporn, Pittayawat & James Kirby. 2017. Laryngeal contrasts in the Tai dialect of Cao Bằng. *Journal of the International Phonetic Association* 47(01). 65–85. doi: 10.1017/S0025100316000293.

Pittayaporn, Pittayaway. 2013. Tonal developments and Southwestern Tai subgrouping. *Journal of Letters* 42(2). 305–339.

Plodkaew, Achana. 2008. *Comparative lexicon of Southern Thai spoken by three generations at Lanska district, Nakhon Si Thammarat province.* Thaksin University M.A. thesis. [In Thai].

Ploykaew, Pornsawan. 1985. *Phonology of Tai Lue in Chiang Rai province.* Mahidol University M.A. thesis.

Plungsuwan, Wipawan. 1981. *Tonal comparison of Tai dialects in Ratchaburi.* Mahidol University M.A. thesis.

Poo-Israkij, Orawan. 1985. *Phonology of Tai Yai at Maelanoi district, Maehongson province.* Mahidol University M.A. thesis.

Poonpholwattanaporn, Mayuree. 2010. *Phonology of Northern Thai at Thasailuat subdistrict, Maesot district, Tak province.* Thammasat University M.A. thesis. [In Thai].

Pornsib, Apinya. 1994. *Tones in Phetchaburi Thai.* Chulalongkorn University M.A. thesis. [In Thai].

Prapaipet, Supot. 1989. *Comparative study of falling tone lexical items in Central Thai and other dialects.* Mahidol University M.A. thesis. [In Thai].

Praphin, Woranuch. 1996. *Comparative lexicon of Lao Song of Nakhon Pathom, Ratchaburi and Phetchaburi provinces.* Silpakorn University M.A. thesis. [In Thai].

Pratankiet, Panitha. 2001. *Tone in Northeastern Thai of Isaan Thai, Khmer and Kui people at Thakhoinang village, Sawai subdistrict, Prangku district, Sisaket province.* Thammasat University M.A. thesis. [In Thai].

Pumma, Samiththicha. 2003. *Comparative lexicon of Tai dialects in Ratchaburi province.* Silpakorn University M.A. thesis. [In Thai].

Pungpawpun, Nongnuch. 1984. *Description of Lao Ngaew language of Thong-En subdistrict, Inburi district, Singburi province.* Silpakorn University M.A. thesis. [In Thai].

Punthong, Gunyarat. 1979. *An analysis of lexical change among three generations in Kam Muang dialect.* Mahidol University M.A. thesis.

R Core Team. 2019. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

Rakmoh, Supa. 2007. *Differences between the sound system of Soang Thai dialect of Sai-ngam Hamlet and that of Standard Thai to construct exercises for solution of problems of oral reading of standard Thai.* Thaksin University M.A. thesis. [In Thai].

Rakpaet, Dueanpen. 1998. *Phonology of Sukhothai dialect with comparison to Nakhon Sithammarat dialect.* Mahidol University M.A. thesis. [In Thai].

Rakpaet, Dueanpen. 2010. *A study of the language situation and Lao Isan lexical shift in Roi Et province*: Mahidol University Ph.D. dissertation.

Rankin, Robert L. 2003. *The comparative method* 183–212. Blackwell.

Ratanadilok Na Phuket, Lorrat. 1983. *Tones in the Thai dialect of Ratchaburi province.* Chulalongkorn University M.A. thesis. [In Thai].

Ratanapraseart, Wanna. 1985. *Lexical study of Lao Wiang language in Chachoeng Sao province.* Silpakorn University M.A. thesis. [In Thai].

Ratliff, Martha. 2010. *Hmong-mien language history.* Pacific linguistics.

Ratliff, Martha. 2015. Tonoexodus, tonogenesis, and tone change. *The Oxford handbook of historical phonology* 245–261.

Revell, Liam J., Luke J. Harmon & David C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* 57(4). 591–601. doi:10.1080/10635150802302427. https://doi.org/10.1080/10635150802302427.

Rexová, Kateřina, Daniel Frynta & Jan Zrzavý. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19(2). 120–127. doi:10.1111/j.1096-0031.2003.tb00299.x.

Rinprom, Chalida. 1977. *Phonemes of the Khorat dialect.* Chulalongkorn University M.A. thesis. [In Thai].

Rittiwong, Nongyaow. 1997. *Lexical geography of Southern Thai in Narathiwat province.* Thaksin University M.A. thesis. [In Thai].

Rivera-Castillo, Yolanda & Lucy Pickering. 2004. Phonetic correlates of stress and tone in a mixed system. *Journal of Pidgin & Creole Languages* 19(2).

Robinson, William. 1849. Notes on the languages spoken by the various tribes inhabiting the valley of Asam and its mountain confines [Section: The Khamti]. *Journal of the Royal Asiatic Society of Bengal* 18(1). 183–237,310–349.

Robinson III, Edward Raymond. 1994. *Further classification of southwestern tai" p" group languages.* Chulalongkorn University M.A. thesis.

Rousselot, P. J. 1897. *Principes de phonétique expérimentale*, vol. 1. H. Welter.

Saeneetontikul, Prapapan. 1985. *Lexical study of Southern Thai in Surat Thani, Nakhon Si Thammarat, and Songkhla.* Chulalongkorn University M.A. thesis. [In Thai].

Saeng-ngam, Suntharat. 2006. *Lexical and tonal variation by age group and language attitude in Song (Black Tai) of Khao Yoi district, Phetchaburi province.* Chulalongkorn University M.A. thesis. [In Thai].

Sagart, Laurent. 2004. The higher phylogeny of austronesian and the position of tai-kadai. *Oceanic Linguistics* 43(2). 411–444.

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. `https://www.pnas.org/content/116/21/10317`.

Sakdanuwatwong, Jantana. 1995. *Language geography of Prachin Buri and Sa Kaeo provinces.* Chulalongkorn University M.A. thesis. [In Thai].

Sawangwan, Sawai. 1991. *Tone geography of Thai dialects in Chaiyaphum province.* Mahidol University M.A. thesis. [In Thai].

Schooler, Jonathan W. 2014. Metascience could rescue the 'replication crisis'. *Nature News* 515(7525). 9.

Seangsrichan, Chayanon. 1998. *Lexical usage of Tai Lue by three generations in Chiangkham district, Phayao province.* Silpakorn University M.A. thesis. [In Thai].

Senisrisant, Rachanee. 1983. *Comparative study of some phonetic and phonological aspects of Maapplakhao Lao Phuan of speakers of different age groups.* Chulalongkorn University M.A. thesis. [In Thai].

Shen, Yu-May. 2003. *Phonology of Sanchong Gelao.* University of Texas at Arlington M.A. thesis.

Sicoli, Mark A & Gary Holton. 2014. Linguistic phylogenies support back-migration from beringia to asia. *PLoS One* 9(3). e91722.

Sidwell, Paul. 2014. 3 austroasiatic classification. In *The handbook of austroasiatic languages (2 vols)*, 144–220. Brill.

Sila, Sarapee. 1975. *Comparative study of Thai and Khu Bua.* Silpakorn University M.A. thesis. [In Thai].

Siriwisitkun, Sriphin. 1986. *Description of Nyo (Yo) of Khlongnamsai subdistrict, Aranyaprathet district, Prachinburi province.* Silpakorn University M.A. thesis. [In Thai].

Sitthi, Rapeeporn. 2006. *Lexical and tonal variation in khorat thai by age group and ease of communication.* Chulalongkorn University M.A. thesis. [In Thai].

Sittiprapaporn, Wichian. 1997. *Tone geography of Thai dialects in Udonthani province.* Mahidol University M.A. thesis.

Smyth, David. 2001. Farangs and siamese: a brief history of learning thai. *Essays in Tai Linguistics* 277–285.

Sodsongkrit, Metcha. 2009. The correspondence /r/, /k/, /kh/ and other correspondences: evidence for chinese and thai as related languages. *Manutsat Paritat: Journal of Humanities* 31. [In Thai].

Sodsongkrit, Metcha. 2010. List of Isan dialect words in Northeast of Thailand which possibly connected with Tai-Chinese family language. *Journal of East Asian Studies, Thammasat University* 14(2). 124–162. [In Thai].

Sodsongkrit, Metcha. 2012. A historical linguistic study of disyllabled reduplicated words in isan dialect which are assumed to be sino-tai cognate words. *Journal of Humanities, Naresuan University* 9(1). [In Thai].

Soiyana, Disaraporn. 2009. *Acoustic analysis of tone in Yong: a comparison between citation and connected speech in two age-groups.* Chulalongkorn University M.A thesis. [In Thai].

Sombatmaungkan, Banyatporn. 1990. *Lexical geography of Sakon Nakhon province.* Silpakorn University M.A. thesis. [In Thai].

Somnuk, Jariya. 1982. *Phonology ofNakhon Si Thammarat sub-dialects.* Chulalongkorn University M.A. thesis. [In Thai].

Soongsumaln, Nonglak. 2002. *Phonology ofLao Derm dialect at Bo Kradan subdistrict, Paktho district, Ratchaburi province.* Silpakorn University M.A. thesis. [In Thai].

Sopheap, Eng. 2017. *Tone systems and the grouping of Lao languages in area of Cambodia-Laos border.* Mahasarakham University M.A. thesis. [In Thai].

Sornjitti, Nantaporn. 2007. *Comparative study of Southern Thai lexical usage among three generations in Chumpon province.* Thammasat University M.A. thesis. [In Thai].

Sritararat, Pojanee. 1983. *Tonal comparison ofPhuthai dialect in 3 provinces.* Mahidol University M.A. thesis. [In Thai].

Stark, Tammy Elizabeth. 2018. *Caribbean northern arawak person marking and alignment: a comparative and diachronic analysis*: University of California, Berkeley Ph.D. dissertation.

Subcharoen, Tippawan. 1989. *Phonology of Trat Thai with comparison to Rayong and Chanthaburi dialects.* Mahidol University M.A. thesis.

Sukpiti, Charuwan. 1989. *Description ofLao Phuan dialect ofHuawa subdistrict, Simahaphot district, Prachinburi province.* Silpakorn University M.A. thesis. [In Thai].

Sukpreedee, Janthiraporn. 1988. *Lexical geography of Thai dialects in Rayong, Chanthaburi and Trat provinces*. Chulalongkorn University M.A. thesis. [In Thai].

Sumransook, Kularb. 1995. *A study of tones in the Thai dialect of Chonburi.* Burapha University M.A. thesis. [In Thai].

Sungkep, Tanakorn. 1983. *Phonological study of Lao Ngaeo with comparison to five Tai dialects*. Mahidol University M.A. thesis.

Sungvanthrup, Chonlada. 1991. *Description of Lao Phuan dialect of Nongsaeng subdistrict, Pakphli district, Nakhon Nayok province.* Silpakorn University M.A. thesis. [In Thai].

Suntharawakun, Benchawan. 1962. *Phonemes of Chiang Mai Thai.* Chulalongkorn University M.A thesis. [In Thai].

Suppasin, Nattawit. 2011. *Phonology of displaced Thais in Prachuap Khirikhan, Chumphon, Ranong, and Phang-nga provinces.* Mahidol University M.A. thesis. [In Thai].

Sutadarat, Suntana Gungsadan. 1978. *A phonological description of Standard Thai*: University of Wisconsin-Madison Ph.D. dissertation.

Suwanmusik, Rangsita. 2004. *Lexical variation among three age-groups in the Koh Samui Southern Thai dialect, Surat Thani province.* Thaksin University M.A. thesis. [In Thai].

Suwanratt, Charoen. 1991. *Study of words and meanings of a Southern Thai dialect in Phithen subdistrict, Thung Yang Daeng district, Pattani province.* Prince of Songkla University M.A. thesis. [In Thai].

Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas

Wahlberg. 2013. Shedding more light on language classification using basic vocabularies and phylogenetic methods: a case study of uralic. *Diachronica* 30(3). 323–352.

Taengko, Jiraphon. 1987. *Tonal comparison of Tai dialects in Loei province.* Mahidol University M.A. thesis.

Tanlaput, Anirut. 1988. *Tonal variation of Lampang Kham Muang.* Mahidol University M.A. thesis.

Tanprasert, Pornpen. 2003. *A language classification of Phuan in Thailand: a study of the tone system.* Mahidol University M.A. thesis.

Tanyong, Utaiwan. 1983. *Lexical change among three generations in Phuan dialect.* Silpakorn University M.A. thesis. [In Thai].

Taylor, Ann, Tandy Warnow & Don Ringe. 1995. Character-based reconstruction of a linguistic cladogram. *Historical linguistics 1995* 1. 393–408.

Tebpawan, Jutada. 2012. *Lexical variation and maintenance among three generations in Khao Phra Narai community, Le subdistrict, Kapong district, Phang-nga province.* Thaksin University M.A. thesis. [In Thai].

Teeranuwat, Sutatip. 2002. *Phonology of Thai Berng-Thai Derng at Khoksalung subdistrict, Phatthananikhom district, Lopburi province.* Mahidol University M.A. thesis. [In Thai].

Thavorn, Sopita. 2013. *Phonological variation and change in Tai Dam.* Mahidol University M.A. thesis. [In Thai].

Thawarorit, Sriangkarn. 2006. Using a phonetic alphabet to record the sound of words without tone marks in Standard Thai. [In Thai].

Thianthaworn, Rungnapa. 1998. *Phonological comparison of four Tai Yuan dialects in central Thailand.* Mahidol University M.A. thesis. [In Thai].

Thongmark, Wannaporn. 1983. *Isogloss (lexical) between Central Thai and Southern Thai.* Chulalongkorn University M.A. thesis. [In Thai].

Thongphiew, Urairat. 1989. *Phonological comparison of Roi-et Thai and Vientiane Lao.* Mahidol University M.A. thesis.

Thongrat, Phutphong. 1988. *Phonology of Phuan at Suphanburi and Sukhothai provinces.* Mahidol University M.A. thesis.

Thumsaro, Chiraporn. 1993. *Lexical geography of Southern Thai in Pattani province.* Thaksin University M.A. thesis. [In Thai].

Thurgood, Graham. 2007. Tonogenesis revisited: Revising the model and the analysis. *Studies in Tai and Southeast Asian Linguistics* 263–291.

Tingsabadh, M.R. Kalaya. 1990. *Tones in Suphanburi Thai: comparative study of tones in words and in connected speech.*

Tippol, Achara. 1988. *The phonemic system of the dialect of Uthumphon Phisai district, Sisaket.* Silpakorn University M.A. thesis. [In Thai].

Tisapong, Areeluck. 1985. *Phonology of Kaloeng language at Ban Dong Ma Fai, Sakon Nakhon province.* Mahidol University M.A. thesis.

Udomphan, Cherdchai. 2000. *Genetic relationship between Sakom, Tak Bai, and Songkhla dialects.* Prince of Songkla University M.A. thesis. [In Thai].

Unakornsawat, Orapan. 1993. *Phonological comparison of Phu Thai and Lao Song.* Silpakorn University M.A. thesis. [In Thai].

University, Duke. 2019. Systematic reviews: the process: Grey literature. `https://guides.mclibrary.duke.edu/sysreview/greylit`. Accessed: 2019-06-01.

Vaitayavanich, Kuntalee. 1991. *Lexical study of Southern Thai spoken in Yala, Pattani and Narathiwat.* Silpakorn University M.A. thesis. [In Thai].

Walker, Robert S & Lincoln A. Ribeiro. 2011. Bayesian phylogeography of the arawak expansion in lowland south america. *Proceedings of the Royal Society B: Biological Sciences* 278(1718). 2562–2567.

Wangsai, Piyawat. 2007. *A comparative study of phonological Yong and Northern Thai language (Kammuang*). Kasetsart University M.A thesis. [In Thai].

Weesakul, Varee. 1983. *Lexical geography of Sukhothai province.* Chulalongkorn University M.A. thesis. [In Thai].

Weidert, Alfons. 1977. Tai-Khamti Phonology and Vocabulary.

Weiss, Michael. 2014. The comparative method. In Claire Bowern & Bethwyn Evans (eds.), *The routledge handbook of historical linguistics*, 127–145. Routledge.

Wetchasit, Sunikun. 1987. *Current Thai dialects in Narathiwat.* Srinakharinwirot University, Songkhla M.A. thesis. [In Thai].

Wheeler, Ward C & Peter M Whiteley. 2015. Historical linguistics as a sequence optimization problem: the evolution and biogeography of u to-a ztecan languages. *Cladistics* 31(2). 113–125.

Wichmann, Søren & Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24(2). 373–404.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2018. The ASJP Database (version 18) .

Withayasakpan, Sompong. 1979a. *A study of Rayong sub-dialects.* Chulalongkorn University M.A thesis. [In Thai].

Withayasakpan, Sompong. 1979b. *A study of Rayong sub-dialects.* Chulalongkorn University M.A. thesis. [In Thai].

Worachin, Sirikanya. 2009. *Study of linguistic status of the Phuthai Language in Kuchinarai district, Kalasin province.* Mahidol University M.A. thesis. [In Thai].

Worawong, Netnapa. 2000. *Tone in Kanchanaburi Thai.* Chulalongkorn University M.A. thesis. [In Thai].

Wuttheerapon, Yuwaret. 2004. *Comparative lexicon of the Northern Thai dialects in Failuang subdistrict, Laplae district, Uttaradit province; Bantuek subdistrict, Sisatchanalai district, Sukhothai province; and Tak-ok subdistrict, Bantak district, Tak province.* Silpakorn University M.A. thesis. [In Thai].

Yaowen, Zhou & Luo Meizhen. 2001. *Dǎiyǔ fāngyán yánjiū* [*Dai dialect study*]. Ethnic Publishing House. [In Chinese].

Yeh, Chia-Hsin. 2009. Acoustic correlates of tone 3 and tone 4 in mandarin. *The Journal of the Acoustical Society of America* 125(4). 2751–2751.

Yensamut, Panida. 1981. *Words and meaning in Lao Song: selected topics.* Silpakorn University M.A. thesis. [In Thai].

Yip, Moira. 2002. *Tone.* Cambridge University Press.

Yoojaroensuk, Yowvalux. 1991. *Lexical study of Khammueang dialects in Phrae province.* Silpakorn University M.A. thesis. [In Thai].

Yooyen, Penwipa. 2013. *Tone variation of Thai Song by age group in Ratchaburi province.* Mahidol University M.A. thesis. [In Thai].

Zhang, Junru. 1999. *Zhuàngyǔ fāngyán yánjiù* [*A study of Zhuang dialects*]. [In Chinese].

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.