Mutual predictiveness of sound correspondences for reconstruction and language subgrouping: The case of Gyalrongic preinitials

Yunfan Lai

Nanyang Technological University
Trinity College Dublin

Abstract

This paper uses Gyalrongic languages, a conservative branch of Sino-Tibetan, to illustrate a new method to evaluate proto-language reconstructions in general historical linguistics and to conduct exploratory analyses in language phylogeny. It first reconstructs a proto-system of Gyalrongic preinitials and computes and compares the implicative entropies between reconstructed and modern systems. In a second step, Mutual Implicative Entropy (MIE) is used to measure genetic distance between related languages and to generate Neighbornet networks to visualize the subgrouping of Gyalrongic languages. The resulting networks are in agreement with qualitative historical linguistic analyses and allow adjustments to previous subgroupings obtained by Bayesian phylogenetic inference. Thus, this method can be used to detect nuances in lower sub-branches, which are sometimes neglected by lexicon-based methods. Using MIE in historical linguistics is therefore a quick and efficient means of checking the effectiveness of reconstructions and establishing the accurate preliminary shape of language subgrouping.

Keywords: Sino-Tibetan, Gyalrongic languages, Tangut, Khroskyabs, (Mutual) Implicative Entropy, Neighbornet, historical linguistics

1 Introduction

Historical linguistics, starting from the 19th century Neogrammarian school, was the first subfield of linguistics that used a falsifiable scientific method to tackle problems of language change. The main principles of the Neogrammarians' Comparative Method have not changed since its birth – *Ausnahmslosigkeit* of sound change, analogy (Osthoff & Brugmann 1878) and shared innovations for sub-branch identification (Leskien 1876) – and have proven successful with many language families. The twofold task of the Comparative Method is to first reconstruct the proto-language, and second to then explore relationships among related languages and determine their subgroupings. While reconstruction is still more or less reserved for specialists, language subgrouping has drawn the attention of various other disciplines, especially archaeology and biology, which seek to infer language subgrouping with computational methods. Since the development of Morris Swadesh's

lexicostatistical method (Swadesh 1952), scholars have sought to use lexical data to infer subgrouping. The lexicon is the main locus for achieving this goal and has produced convincing results through Bayesian inference in phylogeny (Gray & Atkinson 2003). As essential material for subgrouping, the lexicon has many advantages (Greenhill et al. 2020). It is easy to access and in many cases demonstrates similarities among languages directly. However, although lexical data need the Comparative Method for the identification of cognates and borrowings, these data are not, technically, considered by historical linguists as the best resource for language subgrouping (especially when compared to phonology and morphology). Phylogenetic inference nowadays does not explicitly rely on shared innovations (although in some way the process does imply this principle).

Quantitative language subgrouping adhering to actual phonological and morphological innovations started in the beginning of the 20th century, and continues today. Czekanowski (1927) used a mixture of phonological and morphological features to establish a distance matrix of several Indo-European languages. This method did gain some further applications in the 1930s, such as in the work done by Kroeber and Chrétien (1937), but gradually lost its influence with the rise of lexicostatistics. More recently, Baxter (2006), focusing on phonological mergers in Mandarin dialects, proposed an effective statistical model to solve Mandarin phylogeny. This method achieved only limited popularity. The major disadvantages of methods based on the Comparative Method is that they require in-depth knowledge on the family or branch in question (although all methods should require at least basic knowledge of the target family/branch), and biases in data selection. Macklin-Cordes et al. (2021) found that comparing phonotactics among sister languages allows for handy curation of data from low-resourced languages and a high degree of automation. Their results also show accurate phylogenetic signals with Pama-Nyungan languages.

This paper demonstrates the two-fold task of the historical linguist, and in particular proposes an information-theoretical method to approach these tasks. It requires a moderate, but not in-depth, familiarity with target families or branches, and is relevant to the principles of Neogrammarian historical linguistics. It shows that sound correspondences are implicative, and uses entropy to test proto-language reconstruction. It also uses mutual predictiveness of sound correspondences between two varieties to infer language subgrouping. This method is illustrated using the correspondences of preinitial consonants in Gyalrongic languages, one of the most conservative sub-branches of the Sino-Tibetan family. Gyalrongic languages are crucial to understanding the history of Sino-Tibetan in that they preserve a great deal of phonological and morphological features that can be reconstructed to Proto-Sino-Tibetan (Jacques 2012; Hill 2019; Zhang et al. 2019b; Arcodia & Basciano 2020).

The present paper offers two key contributions. In the first place, it offers a reconstruction of the preinitial system in Proto-Gyalrongic with evidence that has never been seriously treated before; an important step forward for the historical linguistics of Gyalrongic and Sino-Tibetan. Second, it demonstrates a new and phylogenetically accurate distance-based method of information theory to deal with problems in language subgrouping. This work first uses the information-theoretical method to test the accuracy of the reconstruction of Gyalrongic preinitials, and then explores the method's capacity for inferring language subgrouping. The structure of the paper is as follows. Section 2 provides a general background to the Gyalrongic languages. In §3, I propose a reconstruction of the proto-preinitial system based on Proto-Khroskyabs and Tangut. The second part of the paper, §§4, 5, and 6, examines quantitatively the predictiveness of

preinitials in Gyalrongic languages using implicative entropy, a method previously restricted to the measurement of morphological complexity.

2 Gyalrongic languages

Gyalrongic languages are a group of Burmo-Qiangic languages (Sino-Tibetan) with divergently complex phonology, especially in the onset domain. In particular, Khroskyabs attests more than 700 onsets in the Wobzi dialect (Lai 2017). The rich onset inventory means that elaborate consonant clusters are commonplace. In Wobzi Khroskyabs, an initial consonant cluster can be composed of a sequence of up to six consonants: *ʁ-j<n-l-z>dɜ̂* (PASS<AUT-CAUS-CAUS>buy.I) 'be asked to let someone buy something for one's own profit'. Most Gyalrongic languages attest consonant clusters ranging between two to four successive consonants (Jacques 2021; Gong 2018; Honkasalo 2019; Zhang 2020; Gates 2021). Nevertheless, we observe one exception. Tangut, recently recognized as a West Gyalrongic language (Lai et al. 2020), has an eroded phonological system with very few consonant clusters. Scholars have however shown that some phonation types in Tangut rhymes, such as vowel tensing, lengthening, labial medializing and rhoticization, are transphonologized from the loss of preinitials, comparing them with modern Burmo-Qiangic languages (Gong 1999; Miyake 2012; Jacques 2014).

There are two main sub-branches among Gyalrongic languages, the east branch and the west branch, and their exact subgrouping is presented in §6.2. The varieties used in this paper are cited below.

- East Gyalrongic
    - Japhug (Jacques 2021)
    - Situ
        - Bragbar (Zhang 2020)
        - Somang (Huang & Sun 2002; Jacques 2004)
        - Zbu (Gong 2018; Jacques 2004)
        - Tshobdun (Sun & Bstan'dzin 2019)
    - West Gyalrongic
        - Khroskyabs
            - Siyuewu (author's field notes)
            - Wobzi (Lai 2017)
        - Horpa
            - Geshiza (Honkasalo 2019, author's field notes)
            - Bawang (Yang 2021)
        - Tangut (Li 1997)

In the following section, I first present the syllable structures of Gyalrongic languages (§2.1) before giving a brief introduction to Tangut transphonologization (§2.2).

2.1 Syllable structure of Gyalrongic languages

Gyalrongic languages share similar syllable structures. A syllable is generally comprised of an onset and a rhyme. There are very few instances of onsetless syllables, most of which are grammatical words. While a Gyalrongic rhyme can be intuitively divided into a nucleus (usually a vowel) and a coda (usually a consonant), the structure of the onset is more complex. An onset consists of three parts: one or several preinitial consonant(s) (C_p), an initial consonant (C_i) and a medial (C_m), as shown in Figure 1.

$$\overbrace{C_{p1}C_{p2}[\ldots].C_i.C_m}^{onset}\ \overbrace{VC_f}^{rhyme}$$

Figure 1. Syllable structure of Gyalrongic languages

In a Gyalrongic onset, only the initial consonant is obligatory; medials and preinitials are both optional. The initial consonant can be any consonant from the inventory, and the medial consonant is usually an approximant, such as *-l-*, *-r-* or *-w-*. Preinitial consonants, in most cases, are continuants such as sibilants, liquids and nasals. Some Gyalrongic languages allow plosives as well. Although preinitial clusters can be very long, most lexical, non-derived clusters do not exceed three consonants.

In at least some varieties, phonological tests can be applied to identify the role of each consonant in a cluster; see Jacques and Chen (2004), Jacques (2004: 12-15), Lai (2013), and Lai (2017: 21-24).

2.2 Tangut syllable types and transphonologization

Tangut syllable types with special phonation types are believed to be caused through transphonologization by the loss of preinitials. There are four types of transphonologization in Tangut: tensing, rhoticization, labial medializing (a labial preinitial becoming a medial *-w-*) and vowel lengthening. Jacques (2014) proposes a straightforward reconstruction of Tangut tensing, rhoticization and labial medializing in Pre-Tangut. Tense vowels are reconstructed as *S.C- (> *-Ṿ*),[1] rhoticized vowels as *r.C-* (> *-Vr*), and labial medializing as *P.C- (> *-wV*).[2] Although there are general tendencies in the correspondences of Tangut transphonologization – for instance, tense vowels are likely related to a sibilant preinitial and rhoticized vowels to a rhotic preinitial – actual correspondences are not easily reconstructed. To cite an example, the Tangut forms 𗍫 *za^{ɤ}r^{1}* 'be spicy', 𗠈 *tsə̣^{ɤ1}* 'lung' and 𗥃 *mē^{2}* 'name' exhibit three distinct types of transphonologization, namely rhoticization, tensing and lengthening, but they all correspond to the preinitial consonant *r.C-* in Siyuewu Khroskyabs: *rdzáv* 'be spicy', *rts^{h}óz*

[1] Tangut tense vowels are noted with a dot under the vowel.

[2] Reconstructed forms in Khroskyabs and those cited from other works are preceded by an asterisk '*'. Other hypothetical forms proposed in this paper are preceded by a number sign '#', as DeLancey (2014) utilizes for person markers in Tibeto-Burman. Following the convention of Baxter and Sagart (2014), the uncertainty of a reconstructed sound is noted by round brackets, as in *ʁʊ(ɣ) 'head'. Square brackets are used to indicate a reconstructed sound that may have alternative reconstructions, as in *ɣm[o]r 'last night'. A hyphen indicates a morpheme boundary, and a period is used when the morphological relation between two suspect morphemes is unclear.

'lung' and *rmê* 'name'.[3] No work has yet been done to sort out these non-trivial correspondences. Table 1 demonstrates the three types of transphonologization with modern Gyalrongic cognates and Pre-Tangut reconstructions proposed by Jacques (2014).

Table 1. Tangut transphonologization

	Tangut	Japhug	Siyuewu	Pre-Tangut
star	0109 *gẹ¹*	*ʑŋgri*	*zgrɜ̂*	*S.C-
cook	0439 *ʁʕi̯ᵚ¹*	*sqa*	*sʁî*	*S.C-
forget	2325 *mə²*	*jmɯt*	*lmə̂d*	*S.C-
throat	0458 *qoᵚr¹*	*tɯ-rqo*	*rqê*	*r.C-
leopard	5480 *ziᵚwr²*	*kɯ-rtsɤɣ*	*ʁrdzôɣ* 'grizzly bear'	*r.C-
wife	1894 *jaᵚr¹*	*tɤ-rʑaβ*	*rjáv*	*r.C-
hail	4032 *muᵚr²*		*lmúɣ*	*r.C-
lead	1516 *swi²*	*mtsʰi*	*ʃsêr*	*P.C-
win	3596 *ʁʕwiᵚ¹* 'power'	*βʁa*	*vʁî* 'be arrogant'	*P.C-
light	5120 *swiᵚw¹*	*ʃsoʁ*	*ʃsôɣ*	*P.C-

Jacques (2014) does not propose a Pre-Tangut reconstruction to account for vowel lengthening. Based on his observation of Chinese loanwords in Tangut, Gong Xun (2021) questions the nature of "long vowels" in Tangut, originally proposed by Gong (2003), and believes that the syllables in which long vowels were previously identified actually contain nasal preinitials and neutral vowels (for example, 2396 *dzūᵚ²* is reconstructed as *ndzuᵚ²* by Gong Xun). I share Gong Xun's opinion concerning the actual realization of the "long vowels", but will continue to use this term for the sake of convenience.

Gong Xun's (2021) proposal accounts for a significant portion of Tangut forms originally reconstructed with long vowels, evidenced by comparisons with modern Gyalrongic languages. See Table 2 where modern Gyalrongic cognates have nasal preinitials and Tangut has long vowels.

[3] Tangut forms are given with an original Tangut character above its number in the dictionary. Reconstruction is based on the Gong Hwang-Cherng system (Gong 2003), refined by Gong Xun (2017, 2020). For instance, the notation 0001 *swə¹* 'sprout' shows a Tangut character numbered 0001 in Li Fanwen's dictionary. Its reconstructed phonological from is given as *swə¹*, with a superscript number indicating that it bears tone 1. Its English gloss is given as 'sprout'.

Table 2. Tangut long vowels :: Gyalrongic nasal preinitials

	Tangut	Japhug	Bragbar	Siyuewu	Geshiza	Other
kill (animal)	0716 *ɕī¹*	*ntɕʰa*	*ntɕʰiɛ̂*	*nɕî*	*ntɕʰə*	
think	2621 *sē²*	*sɯso*	*səsó*	*ntsʰə̂tsʰe*	*səsji*	
sit	2396 *dzū˞²*	*amdzɯ*	*mdzû* (Cogtse)		*ndzo*	*amdzóʏʔ* (Zbu)
wait	5322 *lē²*	*nɤjo*	*nɐjɛ̂*	*njé*	*lji*	*ᵑji* (Bawang)
invite	4040 *kʰū¹*	*qru*	*kró* (Somang)	*ɴqʰrú*		*ᵑqʰlə* (Bawang)
nine	3113 *gə¹*	*kɯngɯt*	*kə-ngû*	*ŋgə́d*	*ŋgæ*	
eat/gnaw	0750 *gī¹*	*nɤ-ŋka*			*ŋgə*	

However, there are cases that cannot be explained by nasal preinitials from a comparative point of view. As shown in Table 3, Tangut long vowels also correspond to various preinitial consonants in modern Gyalrongic languages, which indicates that long vowels do not have a single origin.

Table 3. Tangut long vowels :: Gyalrongic non-nasal preinitials

	Tangut	Japhug	Situ	Siyuewu	Geshiza	Other
tail	5677 *mē¹*	*tɤ-jme*	*ta-jmî*	*lmî*		*tɐ-lméʔ* (Zbu)
name	2639 *mē²*	*tɤ-rmi*	*tə-rmiê*	*rmê*	*lmə*	
nose	5700 *nī²*	*tɯ-ɕna*	*tə-ʃna* (Cogtse)	*snæ̂* (Wobzi)	*sni*	
daytime	2440 *nə²*	*sŋi*	*sni* (Cogtse)	*snə̂*	*sɲele*	
wound	5702 *mā¹*	*tɯ-ɣmaz*	*tə-rmâs*	*ɣmí*	*wmæ*	
two	4027 *nə¹*	*ʁnɯz*	*kənə̂s*	*ɣnæ̂ɣ*	*wne*	
fire	4408 *mə˞¹*	*smi*		*ʁmə̂*	*wmə*	*ʁmə* (Bawang)
accumulate	5149 *dū˞¹*	*ajtɯ*	*ka-sə-jtə̂n*	*rdó*		*altûʏ* (Zbu)
weigh	5682 *qā˞r¹*	*skɤr*	*skár*	*skár*		*skôr* (Zbu)

I agree that Tangut vowel lengthening is due to preinitial loss, just as all the other types of transphonologization. However, vowel lengthening is sure to have heterogeneous origins at least before Tangut branched off, and is thus the result of the merger of various preinitials.[4]

In a nutshell, Tangut transphonologization merges different types of preinitials into four categories. The only problem is that we do not know why exactly some preinitials fall into a particular category, while others fall into another. In the following section, I will explore the factors conditioning these mergers.

[4]I proposed that Tangut "long vowels" be reconstructed as *ɦ.C-* in my invited talks on 28 October 2022 (Lai 2022a) at the Hong Kong Polytechnic University and 23 November 2022 at Trinity College Dublin (Lai 2022b).

3 Reconstruction of the preinitial system

In this section, I propose a three-way contrast in the preinitial system of Proto-Gyalrongic based on Tangut and Proto-Khroskyabs, as reconstructed in Lai (2023a). The preinitial system proposed here seeks to account for both Tangut transphonologization and the various preinitial systems of modern Gyalrongic languages. It also serves to test the information-theoretical method proposed in Section 4. The reconstructed system works best for West Gyalrongic, as it is based on two West Gyalrongic languages, however, it is predictive of East Gyalrongic preinitials as well, and therefore satisfactorily represents Proto-Gyalrongic. I do not attempt to reconstruct every preinitial in full detail, which would require comparisons of several Gyalrongic varieties and eventually make this paper difficult to follow. I will instead focus on the essential problems of reconstruction of Gyalrongic preinitials.

In §3.1, I present the preinitial system of Proto-Khroskyabs. In §3.2, I analyze correspondences between preinitials in Proto-Khroskyabs and Tangut. In §3.3, I reconstruct the proto-preinitial system by proposing a three-way syllabicity contrast to account for transphonologization in Tangut.

3.1 The preinitial system of Proto-Khroskyabs

Lai (2023a) provides a thorough reconstruction of the preinitial system of Proto-Khroskyabs. Proto-Khroskyabs distinguishes a two-way syllabicity contrast in the domain of preinitials: a syllabic preinitial with a short vowel, noted *Cə̆.C-, and a non-syllabic preinitial with a single consonant, noted *C.C-. Syllabic preinitials caused lenition of the following voiceless unaspirated plosive or affricate before shortening to a single consonant, while non-syllabic preinitials did not modify the following consonant, a process shown in (1).[5]

(1) a. *C.T- > *C.T-*
Example: *ɕ.koɣː > *skû* 'onion'
b. *Cə̆.T- > *C.W-*
Example: *sə̆-qæ > *sʁî* 'sound'

See Table 4 for the reconstructed system of preinitials in Proto-Khroskyabs, with comaprative examples from modern Siyuewu Khroskyabs. Note that the preinitials assimilate to the voicing of the initial consonants in modern Siyuewu, except *s-* before *v.C-*.[6]

[5] I use a capital *T* to represent a non-lenited consonant, and *W* to represent a lenited consonant.
[6] For the reconstruction of rhymes in Proto-Khroskyabs, see Lai (2021; 2022c; forthcoming in 2025).

Table 4. Proto-Khroskyabs preinitial system with examples

Proto-Khroskyabs	Siyuewu	Example
*pə̆.C-	*v-*	*pə̆.qæ > *vʁî* 'be arrogant'
*p.C-	*v-*	*f.tɕʰə > *ftɕʰə̂* 'melt (vt)'
*mə̆.C-	*v-*	*mə̆.to > *vdê* 'see'
*m.C-	*v-*	*m.to > *ftê* 'slope'
*sə̆.C-	*s-*	*sə̆.put > *svə̂d* 'pus'
*s.C-	*s-*	*s.qæ > *sqí* 'female sibling'
*ɕə̆.C-	*s-*	*ɕə.pæːk > *svîɣ* 'be thirsty'
*ɕ.C-	*s-*	*q.ɕ.pæ > *χspî* 'toad'
*rə̆.C-	*r-*	*rə̆.pæ > *rví* 'axe'
*r.C-	*r-*	*r.p[i] > *rpî* 'honey'
*kə̆.C-	*ɣ-*	*kə̆.tu > *ɣdə́* 'buy'
*k.C-	*ɣ-*	*k.tsu > *xtsə̂* 'male genitalia'
*q.C-	*ʁ-*	*q.ⁿboʸk > *ʁbôɣ* 'be numerous'

The preinitial system of Proto-Khroskyabs offers the oldest reconstruction of any Gyalrongic language, making it the most straightforward and suitable material for investigating Tangut transphonologization.

3.2 Correspondences between Proto-Khroskyabs and Tangut

In this section, I examine preinitial correspondences between Proto-Khroskyabs and Tangut in terms of Tangut transphonologization. Forms in Modern Siyuewu will be provided for comparison.

In my data, there are 469 Tangut forms that have cognates in Gyalrongic languages, among which 148 involve transphonologization. I have excluded nasal preinitial correspondences as these have been superbly reconstructed by Gong Xun (2021). I show here only cases not explained by Gong Xun (2021). The entire dataset is provided in the Supplementary Material.

Table 5 shows correspondences between Tangut forms with long vowels and their Proto-Khroskyabs cognates. Proto-Khroskyabs preinitials *r.C-, *l.C-, *s.C- and *p.C- correspond to Tangut long vowels.

Table 5. Tangut long vowels vs. Khroskyabs

	Tangut	Siyuewu	Proto-Khroskyabs
dream	0330 *mē¹*	*rmô*	*r.moʏ
accumulate	5149 *dūᵚ¹*	*rdo*	*r.doʏ
knead	5321 *nēᵚ²*	*lnî*	*l.n[i]
tail	5677 *mē¹*	*lmî*	*l.m[i]
daytime	2440 *nə²*	*snô*	*s.nə
nose	5700 *nī²*	*snæ-*	*s.næ
bile	3582 *kər²*	*skʰrɔ́*	*ɕ.kʰrə
marrow	2801 *ɬū²*	*pʰjû*	*p.[ɬj]oʏː

Table 6 shows correspondences between Tangut forms with rhoticized vowels and their Proto-Khroskyabs cognates. A good number of cognate sets involve Proto-Khroskyabs forms with rhotic preinitials, *rə.C- and *r.C-, which shows that Tangut vowel rhoticization is in many cases related to rhotic preinitials. There are other correspondences as well, such as Proto-Khroskyabs *lə.C-, *l.C-, *s.C-, *kə.C- and *k.C-.[7]

Table 6. Tangut rhoticized vowels vs. Khroskyabs

	Tangut	Siyuewu	Proto-Khroskyabs
winter	1490 *tsuᵚr¹*	*rtsô*	*r.tsoʏ
face	3158 *ɴwəᵚr²*	*rŋá*	*r.ŋæʏ
be spicy	0045 *zaᵚr¹*	*rdzáv*	*rə.tsæʏp
ask	0268 *jər²*	*rɣǽd*	*rə.kæt
hail	4032 *muᵚr²*	*lmúɣ*	*l.moʏː k
testicles	3102 *gur¹* ‘kidney’	*lʁɔ́d*	*lə.qut
right side	2547 *tɕʃiᵚr¹*	*scʰɔ́ɣ*	*s.cʰ[əː]k
millstone	0973 *wer²*	*ɣví*	*kə.pæ
drum	5328 *baᵚr¹*	*ɣbê*	*k.ⁿbo

Table 7 shows correspondences between Tangut forms with tense vowels and their Proto-Khroskyabs cognates. Tangut tensing corresponds to a surprisingly wide range of

[7]There is an alternative reconstruction for ‘be spicy’, namely *mV-rts(ʰ)æʏp > *rndzæʏp, according to Jacques and d’Alpoim Guedes (2023: 19). This reconstruction is partly based on comparison with other Gyalrongic languages preserving the prefix **mV-*, such as Japhug *mɤ-rtsaβ* ‘be spicy’ and Bragbar Situ *mə-rtsiɛp*. There is however no Khroskyabs-internal evidence for this prefix. Therefore, I keep *rə.tsæʏp in the present paper.

preinitials in Proto-Khroskyabs, covering preinitial syllables and consonants with rhotic, liquid, sibilant and plosive articulations.

Table 7. Tangut tense vowels vs. Khroskyabs

	Tangut	Siyuewu	Proto-Khroskyabs
cough	4615 *tsụᵚ²*	*rtsʰə̂*	*r.tsʰə
axe	5203 *wị¹*	*rvî*	*rə̆.pæ
oath	4600 *ɴwụ¹*	*lŋû*	*l.ŋ[oᵛː]
shoulder	5170 *wạᵚ¹*	*lváɣ*	*lə̆.pæᵛk
cloud	2738 *dəj²*	*zdə̂m*	*s.ⁿdəm
rust	4966 *wịᵚ¹*	*nzɣí*	*nə̆-ⁿsə̆.[k]æ
snot	5731 *nạᵚ¹*	*snív*	*ɕ.næːp
be thirsty	4532 *pạᵚ²*	*svíɣ*	*ɕə̆.pæːk
three	5865 *sọᵚ¹*	*xsə̂m*	*k.soᵛm
shoe	1321 *zị¹*	*ɣzî*	*kə̆.tsæ
grass	0585 *ɕị²*	*χɕí*	*q.ɕ[i]

Table 8 shows correspondences between Tangut forms with labial medializing and their Proto-Khroskyabs cognates. In most cases, labial medializing in Tangut corresponds to labial preinitials in Proto-Khroskyabs, such as *p(ə̆).C- and *m(ə̆).C-. It can also be related to Proto-Khroskyabs *k.C- and *q.C-.

Table 8. Tangut vowels with *-w-* vs. Khroskyabs

	Tangut	Siyuewu	Proto-Khroskyabs
neighbor	4889 *dʑwə¹*	*vdʑə́* 'mate'	*p.ⁿdʑə
be bright	5120 *swiᵚw¹*	*fsôɣ*	*p.soᵛk
nephew	2134 *zwị¹*	*(s-ɣə-)vzí* 'uncle and nephew'	*pə̆.zæ
win	3596 *ʁʃwiᵚ¹* 'power'	*vʁî* 'be arrogant'	*pə̆.qæ
lead	1516 *swi²*	*fsêr*	*m.seːr
sharpen	1670 *swe¹*	*fsə̂*	*[m].sə
bird	2262 *dʑwoᵚw¹*	*ɣbjə̂m*	*k.ⁿbjəm
be sour	2739 *tɕʰwər²*	*χtɕʰə̂r*	*q.tɕʰ[u]r

At first glance, rhoticization and labial medializing in Tangut allow us to make the intuitive generalizations that rhoticization is related to rhotic preinitials and labial

medializing to labial preinitials. On the other hand, the correspondences between Tangut lengthening/tensing and Proto-Khroskyabs preinitials cannot be straightforwardly explained. In the next section, I will look closer at the distributions of these correspondences and propose a reconstruction.

3.3 A three-way syllabicity contrast among Proto-Gyalrongic preinitials

This section argues for a three-way syllabicity contrast among Proto-Gyalrongic preinitials based on Tangut transphonologization. Before coming to the reconstruction, we first look at the statistics of each Proto-Khroskyabs preinitial and their Tangut correspondences.[8]

Overview of preinitial correspondences The counts of each Proto-Khroskyabs preinitial corresponding to Tangut transphonologization types are illustrated in Tables 9 (focusing on individual preinitials), 10 (focusing on the syllabicity of preinitials) and 11 (focusing on the articulation types). Cells with zeros are shaded gray.[9]

Table 9. Tangut transphonologization vs. Proto-Khroskyabs preinitials

Tangut vowel	*r.C-	*rə̆.C-	*l.C-	*lə̆.C-	*s.C-	*sə̆.C-	*ɕ.C-	*ɕə̆.C-	*k.C-	*kə̆.C-	*q.C-	*p.C-	*pə̆.C-	*m.C-
Long	6	0	2	0	4	0	2	0	0	0	0	2	0	0
Rhotic	16	2	1	1	1	0	0	0	1	1	0	0	0	0
Tense	6	1	5	5	14	9	8	1	5	2	2	0	0	0
w-medial	0	0	0	0	0	0	0	0	2	1	0	10	3	2

Table 10. Tangut transphonologization vs. Proto-Khroskyabs preinitials (syllabicity)

Tangut vowel	R	RV	L	LV	S	SV	P	PV	K	KV	Q
Long	6	0	2	0	6	0	2	0	0	0	0
Rhotic	16	2	1	1	1	0	0	0	1	1	0
Tense	6	1	5	5	22	10	0	0	5	2	2
Rounded	0	0	0	0	0	0	10	3	2	1	0

Table 11. Tangut transphonologization vs. Proto-Khroskyabs preinitials (articulation type)

Tangut vowel	R	L	S	P	K	Q
Long	6	2	6	2	0	0
Rhotic	18	2	1	0	2	0
Tense	7	10	32	0	7	2
Rounded	0	0	0	13	3	0

Several generalizations can be made from these statistics.

1. Tangut long vowels only correspond to non-syllabic preinitials in Proto-Khroskyabs.

[8] Note that not all Tangut forms have a corresponding Proto-Khroskyabs form.

[9] R stands for rhoticity, L stands for laterals, S for sibilants, P for labials, K for velars and Q for uvulars.

2. Tangut long vowels do not correspond to Proto-Khroskyabs preinitials with posterior articulations (*k(ə̆).C- and *q.C-).

3. Tangut rhoticization is overwhelmingly associated with rhotic preinitials in Proto-Khroskyabs.

4. Tangut labial medialization is overwhelmingly associated with labial preinitials in Proto-Khroskyabs.

5. Tangut vowel tensing corresponds most regularly to Proto-Khroskyabs sibilant preinitials, although rhotic, liquid and plosive initials do occur.

Reconstruction of syllabicity Since Tangut long vowels only correspond to non-syllabic preinitials in Proto-Khroskyabs, we can immediately conclude that Tangut long vowels are transphonologized from non-syllabic preinitials in the proto-language, written as *C.C- here. See (2).

(2) a. Khroskyabs
*C.C- > *C.C-*
Example: *s.nə > *snə̂* 'daytime'

b. Tangut
*C.C- > *CV̄*
Example: *s.nə > *nə̄* 'daytime'

Because *C.C- derives long vowels in Tangut, it cannot be reused to also explain Tangut rhoticization and tensing, features which correspond to both syllabic and non-syllabic preinitials in Proto-Khroskyabs. Here we must propose an additional syllable type to account for the correspondences concerning rhoticization and tensing. I reconstruct *Cŭ̆ɹ.C- for Proto-Khroskyabs non-syllabic preinitials corresponding to Tangut rhoticization and tensing, and *Cə̆.C- for Proto-Khroskyabs syllabic preinitials corresponding to Tangut rhoticization and tensing.[10] In Khroskyabs, *Cŭ̆ɹ.C- merged with *C.C- before the presyllable in *Cə̆.C- triggered lenition of the initial. See (3), (4) and (5).

(3) Siyuewu

a. *sŭ̆ɹ.T- > *s.T-
Example: *sŭ̆ɹ.t[oʸː] > *s.t[oʸː] > *stû* 'be straight'

b. *sə̆.T- > *sə̆.W- > *s.W-*
Example: *sə̆.k[əːt] *s.k[əːt] > *sɣə̂d* 'ten'

c. *rŭ̆ɹ.T- > *r.T-
Example: *rŭ̆ɹ.qo > *r.qo > *rqê* 'throat'

d. *rə̆.T- > *rə̆.W- > *r.W-*
Example: *rə̆.tsæʸp > *rzɑv > *rdzáv* 'be spicy'

e. *pŭ̆ɹ.T- > *p.T- > *v.T-*
Example: *pŭ̆ɹ.tɕʰə > *p.tɕʰə *ftɕʰə̂* 'melt (vt)'

[10]The reason for using the back closed vowel [ɯ] here is that it has a higher degree of closure than the schwa, and can thus presumably simplify to a consonantal preinitial more easily than a schwa. This reconstruction may be revised with more evidence. It is noteworthy that presyllables of this type exist in other Gyalrongic or Burmo-Qiangic languages. For example, Xumi *lɛ³³ma⁵⁵* 'forget' corresponds to Pre-Proto-Khroskyabs *lŭ̆ɹ.məːt 'forget' and Xide Yi *le⁴⁴ba³³* 'shoulder' (Huang & Dai 1992) to Pre-Proto-Khroskyabs *lə̆.pæʸk 'shoulder'.

f. *pə̆.T- > *pə̆.W-* > *v.W-*
Example: *pə̆.qæ > *pə̆.ʁæ > *vʁî* ‘be arrogant’

(4) Tangut tensing

a. *Cŭ.TV > *TV̩*
Example: *[s]ŭ.twu¹ > 祋 *twụ¹* ‘be straight’

b. *Cə̆.TV > (*Cə̆.W-) > *WV̩*
Example: *[s]ə̆.qaᵚC¹ > 祋 *ʁạᵚ²* ‘ten’

(5) Tangut rhoticization

a. *rŭ.TV > *TVr*
Example: *rŭ.qoᵚ¹ > 蘸 *qoᵚr¹* ‘throat’

b. *rə̆.TV > (*rə̆.WV) > *WVr*
Example: *rə̆.tsaᵚC¹ > 蔆 *zaᵚr¹* ‘be spicy’

(6) Tangut labial medializing

a. *Pŭ.TV > *TwV*
Example: *Pŭ.tɕʰi¹ > 祋 *tɕʰwi¹* ‘melt (vt)’

b. *Pə̆.TV > (*Pə̆.WV) > *WwV*
Example: *Pə̆.qʕaᵚ¹ > 祋 *ʁʕwiᵚ¹* ‘power (cognate to Siyuewu *vʁî* ‘be arrogant’)’

Tangut lenition does not necessarily correspond to Khroskyabs lenition. However, out of the 39 cognate sets where lenition is relevant, only six confirmed instances cannot be predicted by our reconstruction. This minority of examples are however quite consistent with lenition in Horpa languages, such as Bawang, as shown in Table 12. This comparison also hints at the proximity between Tangut and Horpa varieties, which will be further discussed in §6.2.[11]

Table 12. Lenition discrepancies between Khroskyabs and Tangut

Gloss	Proto-Khroskyabs	Siyuewu	Bawang	Tangut
be thirsty	*ɕə̆-pæːk	*svîɣ*	*spa*	隙 *pạᵚ²*
know (how)	*p.s[oɣ]	*fsô*	*ⁿsʰwo*	蘞 *wị²*
pus	*sə̆.put	*svə̂d*	*spo*	鵝 *pəᵚ¹*
be thick (diametre)	*pom	*pə̂m*	*jwə*	蘞 *woᵚ¹*
shuttle	*lə̆.poːɣk	*lvûɣ*		祋 *puᵚ²*
be satiated	*pə̆.kæ	*vɣí*	*wkwə*	蘞 *kwị¹*

Reconstruction of articulation types It is impossible to reconstruct an exact proto-preinitial for every instance only based on Proto-Khroskyabs and Tangut, as there is a small portion of discrepancies like the ones in Table 12. An entire reconstruction would require a thorough comparison with modern Gyalrongic languages of both the Eastern and Western branches, which is beyond the scope of this paper. It is however possible to use an

[11] For a preliminary anaylsis of lenition in Tangut and Modern West Gyalrongic, see Lai (2023c).

early form of Proto-Khroskyabs with a three-way preinitial contrast reconstructed above as a proxy for Proto-Gyalrongic. I provisionally call this version of Proto-Khroskyabs 'Pre-Proto-Khroskyabs (PPK)'. The PPK preinitial system accounts for many relevant and previously unexplained phenomena across Gyalrongic, including Tangut transphonologization. See Table 13 for a comparison between PPK preinitials and preinitials in modern Gyalrongic languages (two West Gyalrongic languages, Siyuewu and Tangut, and two East Gyalrongic languages, Japhug and Somang, are selected).[12]

Table 13. PPK preinitial system compared to modern Gyalrongic languages

Gloss	PPK	Proto-Khroskyabs	Siyuewu	Tangut	Japhug	Somang
nose	*ɕ.næ	*ɕ.næ	*snæ-*	𗴢 *nī²*	*tɯ-ɕna*	*tə-ɕná*
frog	*q[ŭ].ɕŭ.pæ	*q.ɕ.pæ	*χspî*	𘘛 *pʕi̯ᵚ¹*	*qa-ɕpa*	*kʰa-ɕpâ*
be thirsty	*ɕə̆-pæːk	*ɕə̆-pæːk	*svîɣ*	𗗙 *paᵚ²*	*ɕpaʁ*	*ɕpâk*
wound	*k.mæ	*k.mæ	*ɣmí*	𗾞 *mā¹*	*tɯ-ɣmaz*	*tə-nmâs*
burn	*kŭ<n>ⁿtɕəᵛv	*k<n>ⁿtɕəᵛv	*ɣndʑôv*	𗈪 *dʑə¹*	*ɣndʑʁβ*	*ndʑôp*
shoe	*kə̆-tsæ	*kə̆-tsæ	*ɣzî*	𗥃 *zi¹*	*tɯ-xtsa*	*tə-ktsâ*
tail	*l.mi	*l.mi	*lmî*	𗋕 *mē¹*	*tʁ-jmɐ*	*ta-jmî*
forget	*lŭ.məːt	*l.məːt	*lmâd*	𗖵 *mə²*	*jmɯt*	*jmâs*
shoulder	*lə̆.pæᵛk	*lə̆.pæᵛk	*lváɣ*	𗍫 *wąᵚ¹*	*tʁ-jwaʁ*	*tʁ-jwɐ́k*
marrow	*p.[ɬ]oː	*p.[ɬ]oᵛː	*pʰjû*	𗘂 *ɬū²*	*tɯ-pju*	*tə-pjo*
be bright	*pŭ.soᵛk	*p.soᵛk	*fsôɣ*	𗾭 *swiᵚw¹*	*fsoʁ*	*pʰsók*
be hungry	*pə̆.kæ	*pə̆.kæ	*vɣí*	𗆄 *kwi̯¹*	*fka*	*pkâ*
lead	*mŭ.seːr	*m.seːr	*fsêr*	𗍘 *swi²*	*mtsʰi*	
grass	*qŭ.ɕi	*q.ɕi	*χɕí*	𗍢 *ɕi̯²*	*xɕaj*	
be heavy	*r.lᵛə	*r.lᵛə	*rdâ*	𗏁 *lə¹*	*rʑi*	*lî*
wife	*rŭ.jæᵛp	*r.jæᵛp	*rjáv*	𗹦 *jaᵚr¹*	*tʁ-rʑaβ*	*ta-rdʑáp*
axe	*rə̆.pæ	*rə̆.pæ	*rví*	𗼇 *wi̯¹*	*tɯ-rpa*	*ɕə-rpâ*
heart, core	*s.ne	*s.ne	*snê*	𗴲 *nē¹*	*tɯ-sni*	*tə-ɕniɛ*
cloud	*sŭ.ⁿd[ə]m	*s.ⁿd[ə]m	*zdə́m*	𗋒 *dəj²*	*zdɯm*	*zdém*
cook	*sə̆.qæ	*sə̆.qæ	*sʁî*	𗦳 *ʁʕi̯¹*	*sqa*	*skâ*

In general, the three-way preinitial contrast reduced into simple non-syllabic preinitials in modern Gyalrongic languages, and underwent transphonologization in Tangut. The contrast is preserved in Khroskyabs and Tangut in different ways, with Khroskyabs merging *C.C- and *Cŭ.C-, and Tangut merging *Cŭ.C- and *Cə̆.C-. The behavior of preinitials and presyllables in East Gyalrongic is subject to future investigation.

[12] The Japhug data is from Jacques (2021), and the Somang data from Huang and Sun (2002) as well as the comparisons in Jacques (2004).

Although the reconstruction in this section seems valid through the Comparative Method, it may still be necessary to test and evaluate it quantitatively. One may wonder to what extent the reconstruction improves our understanding of the preinitial system, and how we assess the accuracy of a reconstruction in general. In §4, I will explore the quantification of reconstructed systems with information theory based on the results obtained in this section, and further discuss the implications this new method holds for language subgrouping.

4 Predictiveness of the reconstructed system using implicative entropy

In what follows in this section, I will test the explanatory power of our reconstruction using information theory. In §4.1, I present basic concepts about implicative entropy and discuss the appropriateness of its application in historical linguistics. In §4.3, I apply implicative entropy, a method first proposed by Bonami and Beniamine (2016), to measure the predictiveness of our reconstructed PPK, especially with regard to Tangut.

In §4.3, I will propose a way to measure interpredictibility with unary implicative entropy.

4.1 An overview of implicative entropy

Shannon (1948) proposed information entropy to measure the quantity of a given set of information. The entropy of a variable X, noted $H(X)$, represents the uncertainty of this variable in a given piece of information. The basic formula of information entropy is given in (7).

(7) Entropy

$$H(X) = -\sum_{i=1} p(x_i) \log_2 p(x_i)$$

Conditional entropy computes the information needed for Y under a given condition, X, and is noted $H(Y|X)$. See (8) for the formula of conditional entropy.

(8) Conditional entropy

$$\begin{aligned} H(Y|X) &= -\sum_{x \in X} p(x) H(Y|X = x) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x) \end{aligned}$$

Information theory first entered the realm of linguistics through inflectional morphology. Ackerman et al. (2009) raises the *Paradigm Cell Filling Problem* (henceforth PCFP), which appears in the form of a paradigm table with empty cells that requires completion based on the forms already in it. Ackerman and Malouf (2013) address the PCFP by working out the conditional entropy of every pair of cells and their average value. Their results show that missing cells are generally easy to guess with the knowledge of other cells, because the average conditional entropy is relatively low even for morphologically complex languages, hence the "low entropy conjecture". Bonami and Boyé (2014) and Bonami and Beniamine (2016) further confirms this hypothesis with implicative

entropy, a method that enables the inclusion of more than one predictor cell. This method is based on the belief that inflectional paradigms have implicational or implicative structures (Wurzel 1989). Implicative entropy has been applied to various languages since it was first proposed including Latin (Pellegrini 2020) and Siyuewu Khroskyabs (Lai 2021).

The present study makes use of basic unary implicative entropy. A pair of cells in a paradigm table, (*A*, *B*), exhibits an inflectional pattern denoted '$A \sim B$' as a variable. When guessing a particular pattern, one needs to choose from a set of patterns that contains the right one, noted '$A_{A \sim B}$'. The implicative entropy for guessing *B* from *A*, $H(A \Rightarrow B)$, is thus the conditional entropy of $A \sim B$ based on $A_{A \sim B}$, expressed by the formula in (9).[13]

(9) Unary implicative entropy

$$H(A \Rightarrow B) = H(A \sim B | A_{A \sim B})$$

A high entropy value indicates a low predictiveness of the paradigm; the empty cell is harder to guess. A low entropy value indicates a high predictiveness; the empty cell is easier to guess. Usually, an entropy value above 1 is considered high, although one can conduct bootstrap analyses to evaluate the results statistically (Lai 2021).

4.2 Sound correspondences are implicative

Although the PCFP was originally conceived to study inflectional morphology, and implicative entropy has so far been uniquely employed for this same purpose, there are other problems in linguistics that are nearly identical to the PCFP, such as predicting, or "retrodicting" (using Bodt and List's 2019; 2022 term), a sound reflex of a given language based on its correspondences in other related languages.

Sound correspondences across related languages can be metaphorically understood as a type of PCFP wherein a paradigm table is substituted by a correspondence table with empty cells surrounded by known sounds. The task is to guess the actual forms in the empty cells with the knowledge of known cells. Thus, the nature of sound correspondences is implicative.

See Tables 14 and 15 for labial initial correspondences in three Gyalrongic varieties, Siyuewu, Japhug and Bragbar. Table 14 shows an obvious correspondence, *p-* :: *p-* :: *p-* (Pattern I), and Table 15 a slightly less trivial one, *v-* :: *p-* :: *p-* (Pattern II), with Siyuewu leniting the labial plosive to *v-*.

Table 14. Cognates with *p* in Gyalrongic languages

	Siyuewu	Japhug	Bragbar
material	*=spi*	*spa*	*spâ*
frog	*χspî*	*qaɕpa*	*kʰaɕpiɛ̂*
sparrow	*pjezə́*	*pɣʁtɕɯ*	*patɕû*

[13] Implicative entropy allows for the use of several predictor cells to predict the empty cell. The formula is:

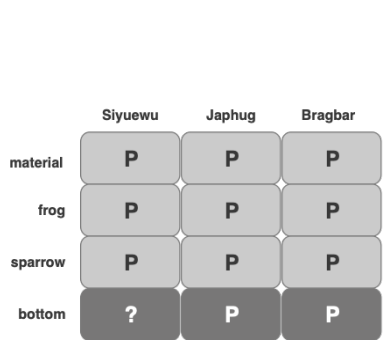
$$H(A^1, A^2, ..., A^n \Rightarrow B) = H(A^1 \sim B, A^2 \sim B, ..., A^n \sim B | A^1{}_{A^1 \sim B}, A^2{}_{A^2 \sim B}, ..., A^n{}_{A^n \sim B}, [A^1 \sim ... \sim A^n])$$

Normally, the higher the number of predictor cells there are, the lower the entropy will be.

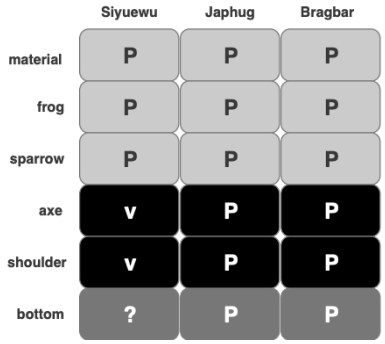
Table 15. Cognates with *p-* and *v-* in Gyalrongic languages

	Siyuewu	Japhug	Bragbar
bottom	*ví*	*ɯ-pa*	*wo-piɛ́*
shoulder	*lváɣ*	*tɯ-rpaʁ*	*tə-rpiɛ̂*
axe	*rví*	*tɯ-rpa*	*ɕʁ-rpiɛ̂*

The "PCFP" of the two correspondences above is shown in Figure 2. In Figure 2a, only Pattern I is known, and the empty cell has no choice but *p-* based on the known pattern. In Figure 2b, both Pattern I and Pattern II are given, and the empty cell can select from *p-* and *v-*. The implicative entropy to guess the Siyuewu forms of the case in Figure 2b is therefore higher than that in Figure 2a, as there is more uncertainty in Figure 2b.



(a) Prediction of the initial of 'bottom' with one single pattern



(b) Prediction of the initial of 'bottom' with two patterns

Figure 2. Prediction of the initial of 'bottom'

We now look at the problem in mirror image: predicting Japhug and Bragbar forms with all Siyuewu forms known, as shown in Figure 3. We find that there is no uncertainty at all in these cases, and that we can safely predict that both 'frog' in Bragbar and 'axe' in Japhug have *p-* as initial. In other words, the entropy value for predicting Bragbar and Japhug forms would be 0.

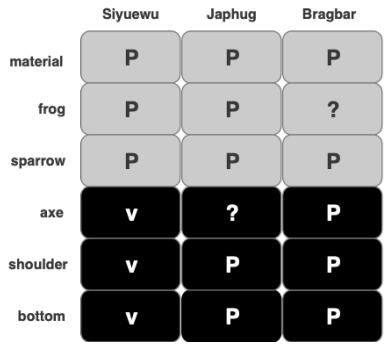


Figure 3. Predicting forms in Japhug and in Bragbar

It is very easy to predict Japhug and Bragbar forms on the basis of Siyuewu forms, but the reverse is not true: one has to choose between *p-* and *v-*. This observation implies that

Japhug and Bragbar have merged two distinct phonemes into *p-*, while Siyuewu preserves an earlier contrast. In this sense, Siyuewu is more conservative than Japhug and Bragbar. This implication can be translated to reconstruction as the positing of two proto-labials, say *p_1- and *p_2-, deriving *p-* and *v-* respectively in Siyuewu, and merging to *p-* in the other two varieties.[14]

4.3 Mutual predictiveness from PPK and Siyuewu to Tangut

Let us now apply implicative entropy to see how well PPK, reconstructed in §3.3, performs in regard to Tangut. Among 148 cognate sets involving Tangut transphonologization, 123 are reconstructible in PPK. For the purposes of this section, the data is presented in a spreadsheet as in Table 16. The "tgt" column contains Tangut syllable types, "R" represents rhotic vowels, "T" represents tense vowels and "L" long vowels.

Table 16. Spreadsheet presentation of raw data

	A	B	C	D	E	F	G
1	Gloss	Tangut	tgt	Siyuewu	syw	PPK	ppk
2	be spicy	za^{ɤ}r^{1}	R	rdzáv	r	*rə̌.tsæ^{ɣ}p	*rə̌.C-
3	star	gẹ^{1}	T	zgrə̂	s	*ɕŭ.^{n}grə	*ɕŭ.C-
4	bile	kə̄r^{2}	L	sk^{h}rə́	ɕ	*ɕŭ.k^{h}rə	*ɕŭ.C-
5	throat	qo^{ɤ}r^{1}	R	rqê	r	*rŭ.qo	*rŭ.C-
6	lung	tsə̣^{ɤ1}	T	rts^{h}óz	r	*rŭ.ts^{h}o^{ɣ}s	*rŭ.C-
7	dream	mẽ^{1}	L	rmô	r	*r.mo^{ɣ}	*r.C-

Table 17 shows the results of implicative entropy computation.

Table 17. Comparison of the predictivenesses from PPK/Siyuewu preinitials to Tangut transphonologization

Siyuewu ⇒ Tangut	Tangut ⇒ Siyuewu	PPK ⇒ Tangut	Tangut ⇒ PPK
1.111	1.891	**0.597**	2.722

The left part of Table 17 concerns the predictiveness between Siyuewu and Tangut in terms of preinitials and transphonologization. The entropy value from Siyuewu to Tangut is 1.111, and that from Tangut to Siyuewu is 1.891. Both values can be considered high, which means it is neither straightforward to predict Tangut transphonologization from Siyuewu preinitials, nor to predict Siyuewu preinitials from Tangut transphonologization. However, Siyuewu ⇒ Tangut is easier by far than the other way around, which means that Siyuewu preserves more contrasts than Tangut.

The right part of Table 17 concerns the predictiveness between PPK (the ancestor of Siyuewu) and Tangut in terms of preinitials and transphonologization. The predictiveness from PPK to Tangut largely increases with regard to that from Siyuewu to Tangut, as evidenced by a much lower entropy, 0.597. On the other hand, Tangut ⇒ PPK is significantly higher than Tangut ⇒ Siyuewu, which means that the reconstructed system

[14] See also List (2019) for a graph-theoretical method for automatic inference of sound correspondence patterns.

has even more contrasts than Siyuewu, making it even harder to guess from Tangut, a phonologically eroded language, which is indeed the case.

Mutual implicative entropy In order to carry out the network analysis of §§6.3 and 6.4, which is based on distance matrices, we need to use a value that represents the distance between two related language varieties to form a symmetrical matrix. This value must reflect the level of interpredictability between two varieties, and is proposed here as achievable through a mutual implicative entropy between varieties A and B.

Let us now return to our tiny sample dataset in Tables 14 and 15 and study the MIE between Siyuewu and Japhug. In the toy dataset, Siyuewu exhibits full predictability toward Japhug; the implicative entropy is thus 0, meaning that no matter what the Siyuewu initial is, there is only $2^0 = 1$ possible correspondence in Japhug, that is, *p-*. In the other direction, the implicative entropy from Japhug to Siyuewu is 1, meaning that Japhug *p-* has $2^1 = 2$ possible correspondences in Siyuewu, namely *p-* and *v-*. On average, Siyuewu predicts all the forms in Japhug, and Japhug guesses only half of the forms in Siyuewu. For n forms in the dataset, the percentage of forms that the two varieties predict correctly for each other are thus $\frac{(n/1+n/2)}{2n} = \frac{1+1/2}{2} = \frac{3}{4} = 0.75$. That is, the pair has a mutual guess rate of 0.75. Letting this guess rate be p_{guess}, and the MIE between Japhug and Siyuewu be H_m, the MIE can thus be calculated as in Equation (1).

Equation 1: MIE between Japhug and Siyuewu (sample dataset)

$$\begin{aligned} p_{guess} &= \frac{1}{2^{H_m}} = 0.75 \\ \therefore 2^{H_m} &= \frac{1}{0.75} \\ \therefore H_m &= log_2 \frac{1}{0.75} \\ &\approx 0.415 \end{aligned} \tag{1}$$

The formula to obtain the MIE between Variety A and Variety B, noted $H(A \Leftrightarrow B)$, can be expressed in Equation (2).

Equation 2: Mutal implicative entropy

$$H(A \Leftrightarrow B) = log_2 \frac{2 \cdot 2^{H(A \Rightarrow B)} \cdot 2^{H(B \Rightarrow A)}}{2^{H(A \Rightarrow B)} + 2^{H(B \Rightarrow A)}} \tag{2}$$

Regarding the actual dataset, using our formula, the MIE between Siyuewu and Tangut is 1.449044, and the MIE between PPK and Tangut is 1.299425. The mutual predictiveness is higher between PPK and Tangut than between Siyuewu and Tangut.

5 Implications of mutual predictiveness for historical linguistics

In §4, I have shown that sound correspondence patterns are a type of PCFP, and can be computed using implicative entropy. Most intuitively, implicative entropy measures the complexity of a set of sound correspondences, or the difficulty of guessing the sounds of a given language on the basis of correspondences in related languages. I have also proposed a

formula to compute the mutual predictiveness, or MIE, of correspondence patterns between two related languages.

The inherent complexities of predicting sounds based on correspondence patterns has implications for determining the relation between two languages. As far as cognate forms are concerned, it is natural that closely related languages show higher mutual predictability, and remotely related languages show lower predictability: it is easy to guess a word's pronunciation between British and American English, but it is harder to guess between English and German. Furthermore, it is easier to predict sounds in languages with contrast mergers from the correspondences in languages with more phonemic contrasts than the reverse direction: even within varieties of English, a rhotic speaker can guess the majority of non-rhotic pronunciations by just removing the feature of rhoticity, while a non-rhotic speaker, without the help of orthography, would have few clues whether a form is rhotic or not. Each sound pattern is governed by one or a set of sound laws, the condition(s) of which may or may not be obvious. If two languages are very close to each other, their sound patterns will exhibit almost one-to-one correspondences. In other words, every sound in one language follows a minimum number of sound laws to reach its corresponding sound in the other, resulting in high mutual predictiveness. In the case that two languages are intelligible to some extent, mutual predictiveness may have implications for the degree of intelligibility.[15] On the other hand, if two languages exhibit complex correspondence patterns wherein one sound in one language has several correspondences in the other, there must be multiple sets of sound laws which have led to these different reflexes, resulting in low mutual predictiveness. Such messy correspondences requiring multiple sound laws usually occur in remotely related languages, both having independently undergone different splits and mergers. Speakers of an Indo-European *centum* language would find it difficult to track correspondences between *k- and *ḱ- in a *śatem* language because of the *centum* merger of plain and palatal velars, while *śatem* speakers would also be similarly confused about plain and rounded velars in *centum* languages, given that these are merged in their mother tongue. Splits and mergers are phonological innovations, and shared phonological innovations are a fundamental criterion for language subgrouping in the Comparative Method. A high number of shared innovations between two languages will distance them from languages that do not exhibit those innovations, resulting in a low MIE between two such languages and a high MIE between such a pair and a third, more remote language.

Returning to Siyuewu and Tangut, why is it easier to guess a Tangut syllable type from a Siyuewu preinitial ($H = 1.111$) than the other way around ($H = 1.891$)? The answer is that Siyuewu preserves fourteen contrasts in the preinitial domain, which merged into only four syllable types in Tangut. And why does PPK ($H = 0.597$) predict Tangut forms much better than modern Siyuewu does ($H = 1.111$)? Because one more syllabic contrast (resulting in eight more concrete phonemic contrasts) is reconstructed for PPK, significantly increasing the predictability of Tangut. The MIE between PPK and Tangut ($H = 1.299425$) is lower than that between Siyuewu and Tangut ($H = 1.449044$), which means that PPK has a closer relation with Tangut than Siyuewu. This should not be surprising, as PPK is the direct ancestor of Siyuewu and would have undergone many fewer changes. Given that modern Siyuewu has been separate from Tangut for at least two thousand years (the assumption is based on Lai and List 2020), there has been far greater opportunity for many more independent sound changes to occur.

In essence, closely related languages usually have fewer sound laws to pass from one to

[15] See Gooskens et al. (2007) for a successful case study of mutual intelligibility among Scandinavian languages with conditional entropy.

another, hence fewer and less messy sound correspondence patterns, while the opposite is true for remotely related languages. The messiness of correspondence patterns can be measured using MIE, thus, MIE should be regarded as a valid indicator of language relations.

6 Gyalrongic subgrouping based on MIE of preinitial correspondences

As can be seen from the analyses presented in §§4 and 5, MIE of sound correspondences presented is predictive of genetic closeness between languages. In this section, I will focus on this hypothesis and explore the significance of MIE in language subgrouping.

6.1 Data presentation

As well as Tangut, I selected nine modern Gyalrongic varieties with reliable data to test the method. They are Siyuewu and Wobzi (Khroskyabs varieties), Geshiza and Bawang (Horpa varieties), Bragbar and Somang (Situ varieties), Japhug, Tshobdun and Zbu.

As is briefly mentioned in §3.2, there are 469 known cognates between Tangut and modern Gyalrongic languages, 148 of which involve transphonologization.[16] Not all cognate sets cover all the varieties under investigation. There are 58 cognate sets in the entire dataset that have forms in every Gyalrongic variety in this study, which is 39.2 per cent of the cognate sets with transphonologization, and 12.4 per cent of all cognates. I will thus focus on two subsets of data.

1. Subset 1 ("Unequal Subset"): This subset takes all 148 concepts with Tangut transphonologization into account, and computes MIEs of every pair of varieties. The process will cause unequal comparisons among the varieties, as every pair of varieties has a different overlapping set of cognates, but involves a maximum number of cognates in the analysis.

2. Subset 2 ("Equal Subset"): This subset only considers the 53 cognate sets shared among all Gyalrongic varieties in our data. Its advantage is that it guarantees that the cognates involved in the analyses are equal in all varieties. Though this data set it limited in size, I consider the 53 cognate sets as sufficient for the purpose of the present study.

The choice of cognate sets depends largely on chance availability among existing sources. See Table 18 for the exact counts in each Gyalrongic variety.

[16] Character independence plays an important role in selecting the cognates. MIE computation takes type frequency into account (Bonami & Beniamine 2016: 163-164), therefore, etymological duplicates must be avoided as they affect the statistics of type frequency. For example, both *rvî* 'axe' and its bound state *rvæ-* in Khroskyabs are derived from the proto-form **rə.pæ*, only one of which, preferably *rvî*, should be included in the database.

Table 18. Tangut cognate counts

Total cognate count	469
Common cognates	58
Tangut	148
Siyuewu	132
Wobzi	128
Japhug	123
Geshiza	119
Bawang	118
Tshobdun	104
Zbu	102
Bragbar	98
Somang	96

The data for this study were entered into a spreadsheet in a similar format to Table 16. Note that PPK is no longer included. See Table 19.

Table 19. Data curation of Gyalrongic preinitials in a spreadsheet

	A	B	C	D	E	F	G
1	Gloss	Tangut	tgt	Siyuewu	syw	Japhug	jph
2	be spicy	za^{ɤ}r^{1}	R	rdzáv	r	mɯ-rtsaβ	r
3	star	gẹ^{1}	T	zgrə̂	s	zŋgri	ɕ
4	bile	kə̄r^{2}	L	skʰrə́	ɕ	ɕkrɯt	ɕ
5	throat	qo^{ɤ}r^{1}	R	rqê	r	tɯ-rqo	r
6	lung	tsə̣^{ɤ1}	T	rtsʰóz	r	tɯ-rtsʰʏz	r
7	dream	mẽ^{1}	L	rmô	r	tɯ-jmŋo	j

A Python implementation is included in the Supplementary Material for the computation of MIE.

6.2 Subgrouping of Gyalrongic languages

There are very few studies on the details of Gyalrongic subgrouping. Most previous work operates within top level branching of East and West Gyalrongic languages, and discussion on lower order subgroupings is fragmentary. Sagart et al. (2019) and Lai and List (2020) are two studies involving Gyalrongic subgrouping which apply Bayesian phylogenetic inference. Figure 4 shows the subgrouping proposed by Sagart et al. (2019), and Figure 5 shows that proposed by Lai and List (2020).

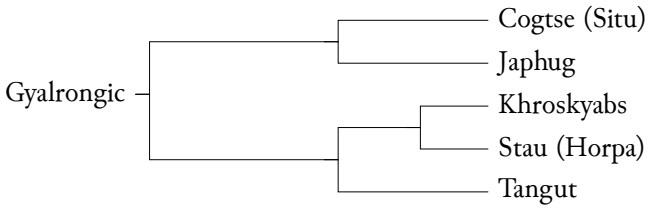


Figure 4. Gyalrongic subgrouping according to Sagart et al. (2019)

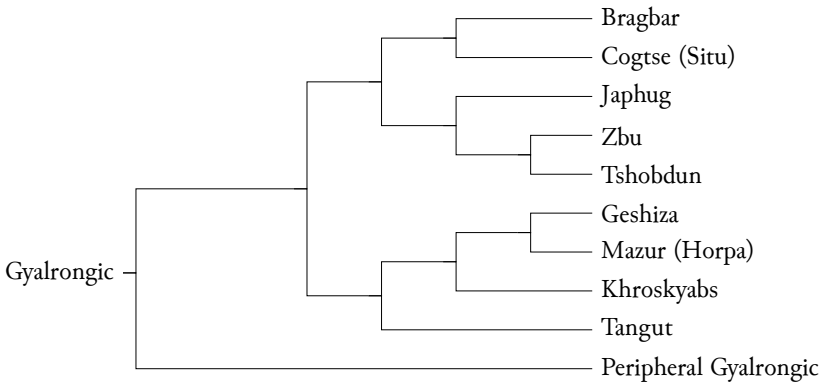


Figure 5. Gyalrongic subgrouping according to Lai and List (2020)

Both trees broadly conform to those produced by the Comparative Method.

The position of Tangut in West Gyalrongic Both Sagart et al. (2019) and Lai and List (2020) position Tangut outside modern West Gyalrongic, albeit confirming the findings of Lai et al. (2020) that Tangut is really West Gyalrongic. However, superficial similarities and the Comparative Method suggest that Tangut is closer to Horpa varieties than to Khroskyabs. The proximity between Horpa and Tangut was initially mentioned by Stein (1966: 285) without substantial linguistic evidence. Subsequently, through a comparison of basic vocabulary, Li and Lin (1983: 305) conclude that Tangut is indeed more closely related to Horpa than to East Gyalrongic and Tibetan, further hypothesizing that the Horpa people are descendants of the Tangut. Recently, Honkasalo (2019: 646) finds that Geshiza and Tangut share a common distribution of cognate modal negators, indicating a probable shared innovation. Lai (2023b: 27) points out shared innovative behavior between Tangut and Horpa in relation to sound changes involving voiceless nasal initials, and Zhang (2023; forthcoming) finds that Tangut and Horpa share innovations in the lexical pairing of certain orientation prefixes and verbs, such as the cases of 'to die' and 'to kill'.

Tangut was attested in today's Ningxia Province, China, a thousand kilometers to the north of the Gyalrongic homeland. As the official language of the Tangut Empire (1038-1227 AD), it was in frequent contact with the neighboring Chinese, Tibetan, Mongolic, and Tungusic languages, resulting in the absorption of non-native vocabulary from various sources. Tangut's historical circumstances thus create difficulties for determining its genetic affiliation using lexicon-based methods.

Relations among Japhug, Tshobdun and Zbu Japhug, Tshobdun and Zbu are three neighboring languages in the north of the Gyalrongic speaking region. Traditionally, scholars believed that Zbu and Tshobdun formed a separate group, called 'Stodpa dialects' and that they were more closely related to each other than they were to Japhug. This view is also supported by Lai and List (2020) based on lexical data (Figure 5). However, Gong Xun (2018: 24-26) points out that this traditional classification "n'est nullement soutenue par des caractéristiques partagées" (is in no way supported by shared characteristics). Among other morphophonological innovations shared by Japhug and Tshobdun, Gong Xun identifies an important isogloss that helps to refute the Stodpa conjecture. Zbu preserves the rhyme distinction between *tə-rnɑ́* 'ear' and *tə-rvéʔ* 'axe', which is lost in Japhug (*tɯ-rna* 'ear' and *tɯ-rpa* 'axe') and Tshobdun (*tə-rne* 'ear' and *tə-rpe* 'axe'). The distinction must have existed in Proto-Gyalrongic, as all western varieties have it: Tangut

荒 *nu*[1] ‘ear’ / 巍 *wi*[1] ‘axe’, Siyuewu *ɲú* ‘ear’ / *rví* ‘axe’, and it can even be reconstructed to Proto-Sino-Tibetan, as Old Chinese exhibited different reflexes as well: 耳 *C.nəʔ ‘ear’ and 斧 *p(r)a ‘axe’ (Old Chinese is the first branch of Sino-Tibetan according to Sagart et al. 2019 and Zhang et al. 2019a). This piece of evidence shows that Japhug and Tshobdun, along with other East Gyalrongic languages, innovatively merged the ‘ear’-‘axe’ contrast, thereby contradicting traditional belief and the phylogenetic results of Lai and List (2020). The superficial similarity between Zbu and Tshobdun lies largely in their shared retention of old stem alternation systems which Japhug no longer possesses, as well as certain shared vocabulary that Japhug has lost, which might explain why the lexicon-based inference fails to detect relevant differences between Zbu and Tshobdun.

In §6.3, I will use MIE to infer Gyalrongic subgrouping.

6.3 MIE of preinitial correspondences

Table 20 extracts individual MIE values from both the Unequal and Equal subsets between modern Gyalrongic languages and Tangut. The West Gyalrongic varieties are shaded gray. For both subsets, West Gyalrongic varieties occupy roughly the top half of the tables. Although Zbu ends up third in the Unequal subset, it falls to fifth in the Equal subset. On the one hand, Zbu is the most West-Gyalrongic-like East Gyalrongic language (Gong Xun 2018: 25), attesting some sound changes that also occur in most West Gyalrongic languages, such as the merger of *s.C- and *ɕ.C-. On the other hand, I might have not identified all the cognates between Zbu and Tangut, and some forms may be borrowed from West Gyalrongic. I believe that the Equal subset is closer to Zbu’s phylogenetic reality, as the MIEs in both East and West Gyalrongic in the Equal subset show smaller variances – modern varieties in the same branch should hypothetically show similar distances to Tangut.

Table 20. Rankings of mutual predictiveness with Tangut

	MIE (Unequal)		MIE (Equal)
Geshiza	1.044456	Geshiza	0.907200
Bawang	1.135010	Bawang	0.921191
Zbu	1.400072	Siyuewu	1.148636
Siyuewu	1.4449044	Wobzi	1.148636
Wobzi	1.492394	Zbu	1.213984
Japhug	1.494951	Japhug	1.236797
Tshobdun	1.523162	Somang	1.253783
Bragbar	1.578684	Tshobdun	1.255415
Somang	1.678849	Bragbar	1.328332

Tangut and Horpa Observing Table 20 in more detail, Geshiza and Bawang, two Horpa varieties, exhibit the highest mutual predictiveness with Tangut, no matter which subset is taken into account. This result shows that Horpa languages are indeed closer to Tangut as far as preinitial correspondences are concerned. This conclusion agrees with the early studies such as Stein (1966) and Li and Lin (1983), the phonological comparative study of Lai (2023b) and the morphological innovations proposed by Zhang (2023; forthcoming).

Japhug and Tshobdun Table 21 extracts mutual predictiveness among Japhug, Zbu and Tshobdun. Regardless of the subset, the values show that the three varieties are very close, with MIE values well under 0.53. However, contrary to traditional analyses and phylogenetic inference, Japhug and Tshobdun show a significantly higher proximity than the other pairs, especially in the Equal subset, with a MIE value of only 0.315498. This result is in line with Gong Xun (2018) who adheres to shared innovations.

Table 21. Mutual predictiveness among Japhug, Zbu and Tshobdun

	MIE (Unequal)	MIE (Equal)
Japhug ⇔ Tshobdun	0.449633	0.315498
Japhug ⇔ Zbu	0.516944	0.527636
Tshobdun ⇔ Zbu	0.458435	0.507556

6.4 Neighbornet networks inferred from MIE

Information theory has been used for various purposes in phylogenetics in biology (Criscuolo & Gribaldo 2010; Batista et al. 2011). But its use in the subgrouping of language families or branches is still rare. An early attempt was made with typological features over two decades ago (Juola 1998), and Zhou and Bowern (2015) use entropy to measure the uncertainty of ancestral state reconstruction in languages. As already mentioned, most phylogenetic linguistic analyses make use of lexical data. Though these analyses have been successful in a broad sense, they have also resulted in subgroupings which diverge from the principles of the Comparative Method, such as regular sound correspondences and shared innovations.

In this section, I combine the present study's MIE computation with Neighbornet, a method for constructing phylogenetic networks based on mutual distances among taxa (Bryant & Moulton 2004). Neighbornet is a convenient visualization of taxon clustering in a given distance matrix, and thus a preferred choice for exploratory data analyses (Morrison 2010). Neighbornet networks constructed from lexical data have been widely applied to the visualization of language families (Gray et al. 2010), and even recently for Tibeto-Burman (Gao 2020). Neighbornet networks in this paper were generated with the software *SplitsTree 4* (Huson & Bryant 2006).

Neighbornet networks for Gyalrongic preinitial correspondences Figure 6 shows the Neighbornet network inferred from the Unequal subset. The branch lengths indicate the the depth of evidence available to support each proposed split.

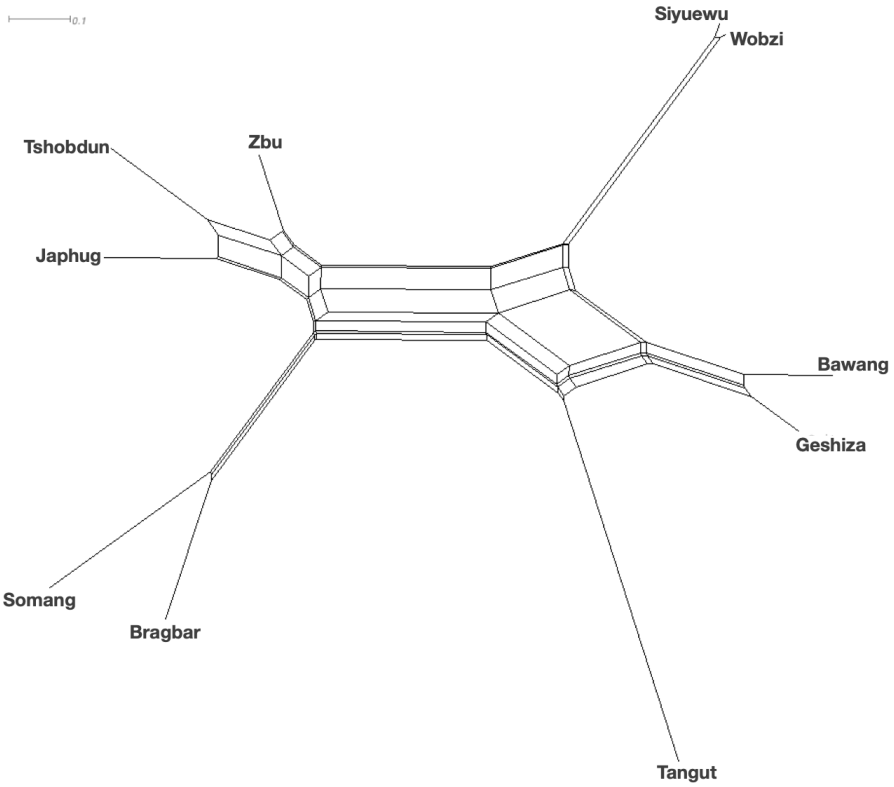


Figure 6. Neighbornet inferred from preinitial correspondences (Unequal subset)

The network shows a separation between the east and west sub-branches, and clear evidence for lower clades. Khroskyabs varieties (Siyuewu and Wobzi), Horpa varieties (Bawang and Geshiza) and Situ varieties (Bragbar and Somang) are unsurprisingly close to each other, consistent with previous analyses.

The Tangut-Horpa clade, as supported by historical linguistic analysis such as Lai (2023b) and Zhang (2023; forthcoming), is marginally confirmed in the network. The Tangut-Horpa clade has a weight of 0.1437, slightly higher than the Horpa-Khroksyabs clade, with a weight of 0.1225. The network also clearly supports the proximity between Japhug and Tshobdun, proposed by Gong Xun (2018).

Delta score (δ score) and Q-residual are used to measure the treelike-ness of the network. For both values, 0 means a perfect tree structure (Holland et al. 2002; Gray et al. 2010). The average δ score of the network in Figure 6 is 0.1941, and the average Q-residual is about 0.01265, indicating a moderate treelike structure.[17]

Figure 7 shows the Neighbornet network inferred from the Equal subset. The topology of this network is similar to the one for the Unequal subset. Though it shows a clearer relation between Tangut and Horpa languages, it still supports the Japhug-Tshobdun clade. The δ score for this network is 0.1805, and the Q-residual is 0.01063.

[17]Tables of detailed δ scores and Q-residuals for all networks are in the Supplementary Material.

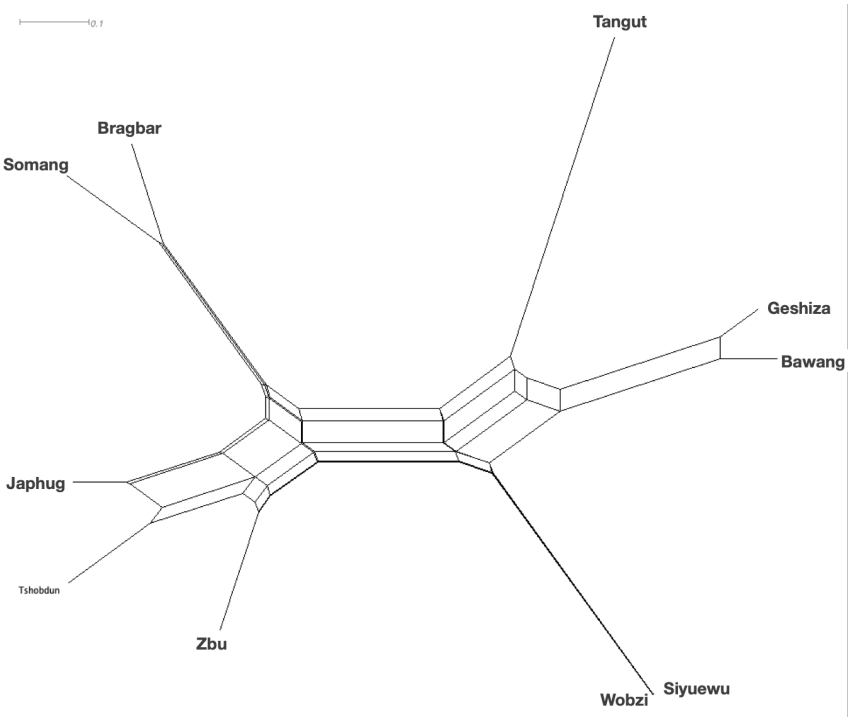
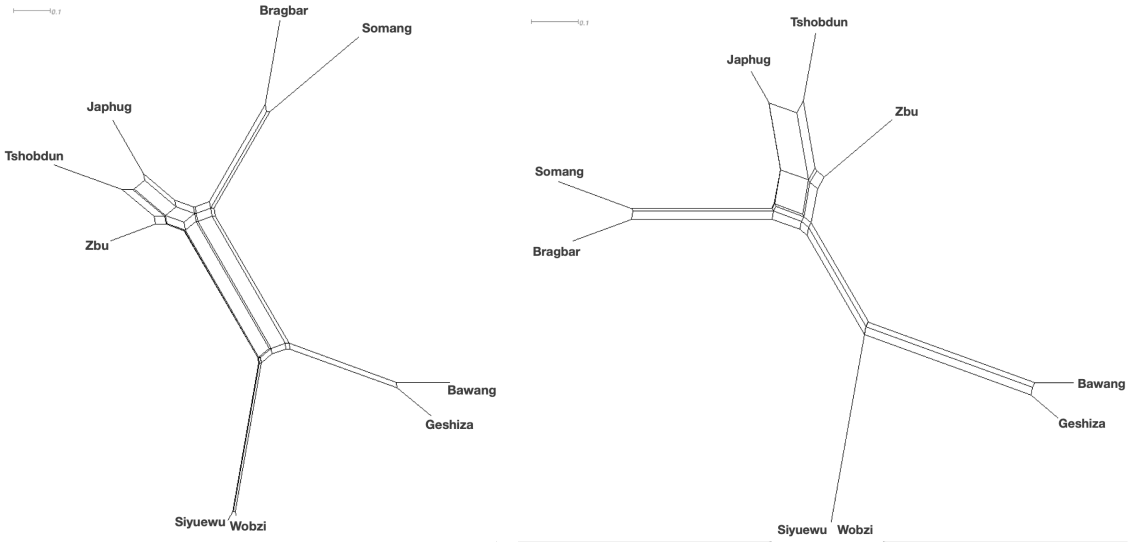


Figure 7. Neighbornet inferred from preinitial correspondences (Equal subset)

Regarding individual δ scores in Figure 6 and Figure 7, Tangut has the highest value, 0.27918 and 0.23929 respectively, which is unsurprising as Tangut phonology is the most eroded of all Gyalrongic languages. The two networks in Figure 8 show cases with Tangut taken off Unequal and Equal subsets. The subgrouping is expected, and the δ scores, 0.1374 and 0.1413 respectively, are reduced significantly from the subsets with Tangut. Given the reduced δ scores, these networks show that modern Gyalrongic languages have a relatively strong tree structure.



(a) Unequal subset **(b)** Equal subset

Figure 8. Neighbornet inferred from preinitial correspondences without Tangut

Neighbornet networks for Gyalrongic initial correspondences One might wonder if the result would be different with correspondences of other parts of the syllable as input data. I thus provide the same analysis with initial consonants in the same dataset. The network generated is illustrated in Figure 9.

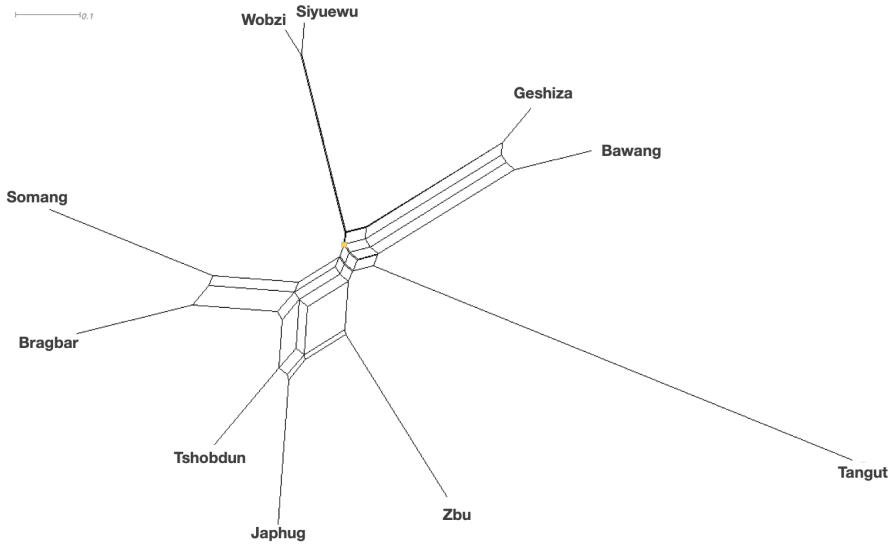


Figure 9. Neighbornet inferred from initial correspondences

The topology of the network for initial correspondences is consistent with the one in Figure 6, although the East-West split is less well supported, with much shorter edges. The average δ score is 0.2679 and the average Q-residual is 0.0098. The tree-ness of this network based on initials is moderate. The correspondences of initials in Gyalrongic languages are less diversified than those of preinitials, thus providing less information about subgrouping. Nevertheless, the topology of this network shows little difference from the preinitial networks in Figures 6 and 7, and still supports the Japhug-Tshobdun and Horpa-Tangut clades.

In summary, as long as the dataset contains few errors in cognate identification and annotation, we do not expect fundamental differences to emerge with different phonological properties as input data. Our example also shows that preinitials are more appropriate than initials to infer the subgrouping in this particular dataset.

6.5 Test with Sinitic varieties

In this section, I use a sample dataset containing eight Sinitic varieties to test the method proposed in this paper. The choice of the varieties is based on my personal knowledge of them.[18] The Sinitic varieties selected are shown in Table 22.

[18] I am a speaker of six varieties to at least an intermediate level, except Nankinese and Fuzhou. I am however familiar with the historical phonology of the latter two.

Table 22. Sinitic varieties in the dataset

Variety	Branch
Putonghua	Mandarin
Sichuanese	Mandarin
Nankinese	Mandarin
Cantonese	Yue
Meixian	Hakka
Zhangzhou	Southern Min
Chaozhou	Southern Min
Fuzhou	Eastern Min

The expected subgrouping of these Sinitic varieties should look more or less like the one in Figure 10 (see also Baxter 2006 and Sagart 2011, among others).

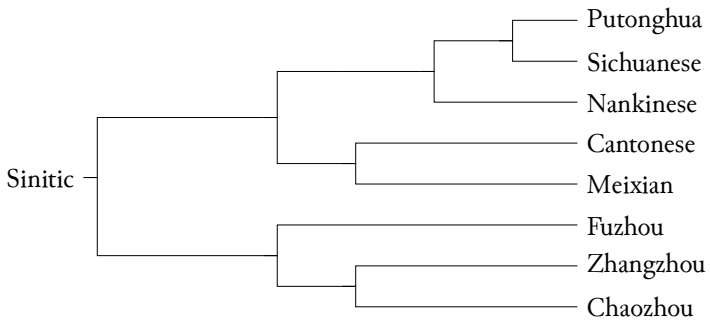


Figure 10. Approximate Sinitic subgrouping

A slightly modified Swadesh list was used in creating the present Sinitic dataset, which compares the initial consonant of a shared Chinese character in each variety. The dataset contains 91 cognate sets, all of which are included in the Supplementary Material.

The Neighbornet network is shown in Figure 11. The average δ score of the Sinitic network is 0.1399, and the average Q-residual is 0.01036, indicating the network shows a relatively strong tree structure. The topology of the Neighbornet network is exactly in line with the expected subgrouping shown in Figure 10.

This brief demonstration with Sinitic suggests that the methods proposed here are suitable for language branches beyond Gyalrongic.

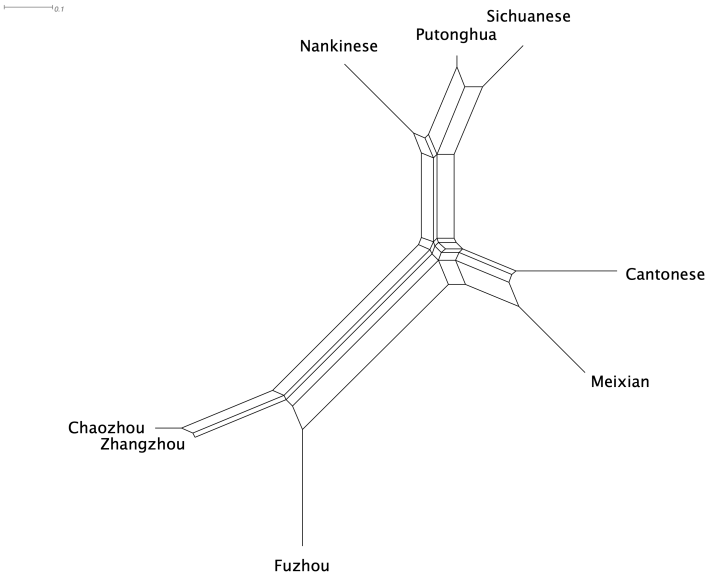


Figure 11. Neighbornet inferred from Sinitic initial correspondences

7 Conclusion

This paper has examined the correspondences of preinitials in Gyalrongic languages from two perspectives. In the first place, it proposed a three-way syllabicity contrast in Proto-Gyalrongic, which helps to explain transphonologization in Tangut, an important problem in Sino-Tibetan historical linguistics. Secondly, it used information theory to test reconstruction and explore the subgrouping of Gyalrongic languages.

The results show that implicative entropy, a method previously applied to morphological complexity, can be efficiently employed to measure the regularity of sound correspondence. A plausible reconstruction should have high predictiveness toward its daughter languages; in the reverse direction, the daughter languages should have lower predictiveness toward their ancestor. Varieties with a high genetic proximity should be rather mutually predictable, reflected by low MIE. On the contrary, varieties that are genetically remote are expected to have high MIE. Thus, MIE can be used to infer language subgrouping, as demonstrated by Gyalrongic and Sinitic in the present paper.

The method detailed here is highly relevant to the traditional Comparative Method of historical linguistics because it focuses on sound correspondence patterns as the material for proto-language reconstruction. One universal feature of language change is the change of phonological shape or structure, with one phoneme splitting into two or more separate phonemes, or two or more phonemes merging into one. Different varieties of any given language may undergo different splits and mergers, resulting in different phonological features from the proto-language and uncertainty in sound correspondences. Independent phonetic changes do not necessarily indicate the closeness or remoteness of sister languages, as such changes are often influenced by homoplasy, where the same sound changes occur in different branches. Both Bawang and Japhug underwent the change of *l.C- to *j.C-*, but Bawang is undoubtedly much closer to Khroskyabs, which did not undergo this change, than it is to Japhug. More often than not, sound change brings about change in phonological shape. Notably in this case, it is not just the sound change which is meaningful *per se*, but that Bawang also attests the change *r.C- > *l.C- > *j.C-* (which may

go back to a different rhotic preinitial), making *jvi* 'axe' and *jvi* 'snow' homophonous and indistinguishable (c.f. Japhug *tɯ-rpa* 'axe' vs. *tʏ-jpa* 'snow'). MIE takes such change in phonological shape into account, detecting the group of patterns *j.C-* :: *j.C-* and *j.C-* :: *r.C-* that is responsible for the split between Bawang and Japhug. In combination with the principles of the Comparative Method, the present method enables us to solve problems beyond the reach of lexicon-based phylogeny, such as the relation between Japhug and Tshobdun, and the position of Tangut.

MIE has its own limits. For example, it does not take the variance between $H(A \Rightarrow B)$ and $H(B \Rightarrow A)$ into account. It only measures the overall guess rate between two varieties and neutralizes the difference in variance. The same guess rate can result from two varieties with high variance between their implicative entropies, or from exactly the same entropy value. However, such cases are unlikely to occur with significant frequency in actual data. As our analyses show, MIE gives expected results with different datasets and language branches. MIE can thus serve as a handy and effective way to infer language subgrouping alongside other more time-consuming methods, such as Bayesian inference in phylogeny. It can also be used to examine classifications of lower branches where other types of data may encounter difficulties. However, it should be noted that this method cannot take the place of dated phylogeny, as it does not include dating at its present stage. The correlation between MIE and branch age is a subject requiring future work.

Funding

This research is funded by the Irish Research Council under the SFI-IRC Pathway Programme (Project ID: 21/PATH-A/9374, Gyalrongic unveiled: Languages, Heritage, Ancestry; awardee: Yunfan Lai) and the Nanyang Assistant Professorship (NAP 2024), Nanyang Technological University, Singapore (Project ID: #024576-00001, Language evolution through trees, entropies, reconstruction and networks).

Acknowledgements

I would like to express my gratitude to Shuya Zhang, Nathan Hill, Agnes Conrad, Guillaume Jacques, the two anonymous reviewers, and the editor, Claire Bowern, for their valuable and insightful comments and corrections.

References

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 54–82. Oxford: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464. doi: https://doi.org/10.1353/lan.2013.0054.

Arcodia, Giorgio Francesco & Bianca Basciano. 2020. Morphology in Sino-Tibetan languages. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*, Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.530.

Batista, Marcus VA, Tiago AE Ferreira, Antonio C Freitas & Valdir Q Balbino. 2011. An entropy-based approach for the identification of phylogenetically informative genomic regions of papillomavirus. *Infection, Genetics and Evolution* 11(8). 2026–2033. doi: https://doi.org/10.1016/j.meegid.2011.09.013.

Baxter, William H. 2006. Mandarin dialect phylogeny. *Cahiers de linguistique Asie orientale* 35(1). 71–114. doi: https://doi.org/10.1163/19606028-03501005.

Baxter, William H. & Laurent Sagart. 2014. *Old Chinese: A new reconstruction*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199945375.001.0001.

Bodt, Timotheus A & Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in western kho-bwa languages. *Papers in Historical Phonology* 4. 22–44. doi: https://doi.org/10.2218/pihph.4.2019.3037.

Bodt, Timotheus A & Johann-Mattis List. 2022. Reflex prediction: A case study of Western Kho-Bwa. *Diachronica* 39(1). 1–38. doi: https://doi.org/10.1075/dia.20009.bod.

Bonami, Olivier & S. Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182. doi: https://doi.org/10.3366/word.2016.0092.

Bonami, Olivier & Gilles Boyé. 2014. De formes en thèmes. In Florence Villoing, Sophie David & Sarah Leroy (eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, 17–45. Paris: Presses Universitaires de Paris Ouest.

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

Criscuolo, Alexis & Simonetta Gribaldo. 2010. BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology* 10. 1–21. doi: https://doi.org/10.1186/1471-2148-10-210.

Czekanowski, Jan. 1927. *Wstęp do historii Słowian: Perspektywy antropologiczne, etnograficzne, prehistoryczne i językoznawcze*, vol. 3. Jakubowskij.

DeLancey, Scott. 2014. Second person verb forms in Tibeto-Burman. *Linguistics of the Tibeto-Burman Area* 37(1). 3–33. doi: https://doi.org/10.1075/ltba.37.1.01lan.

Gao, Tianjun. 2020. Reconstruction and analysis of phylogenetic network on Tibeto-Burman languages in China. *Journal of Chinese Linguistics* 48(1). 257–293. doi: https://doi.org/10.1353/jcl.2020.0006.

Gates, Jesse P. 2021. A grammar of Mazur Stau. Paris: Écoles des Hautes Études en Sciences Sociales dissertation.

Gong, Hwang-cherng. 1999. Xixiayu de jinyuanyin jiqi qiyuan 西夏語的緊元音及其起源 [Tense vowels in Tangut and their origins]. *Bulletin of the Institute of History and Philology* 70.2. 531–558.

Gong, Hwang-cherng. 2003. Tangut. In Randy LaPolla & Graham Thurgood (eds.), *The Sino-Tibetan Languages*, 602–620. London: Routledge.

Gong, Xun. 2017. Grade II in Tangut and Hexi Late Middle Chinese. Paper presented at Recent Advances in Tangut Studies, 24 January, 2017.

Gong, Xun. 2018. *Le rgyalrong zbu, une langue tibéto-birmane de Chine du Sud-ouest : Une étude descriptive, typologique et comparative*. Paris: Institut National des Langues et Civilisations Orientales dissertation.

Gong, Xun. 2020. Uvulars and uvularization in tangut phonology. *Language and Linguistics* 21(2). 175–212. doi: https://doi.org/10.1075/lali.00060.gon.

Gong, Xun. 2021. Nasal preinitials in Tangut Phonology. *Archiv orientální* 89(3). 443–482. doi: https://doi.org/10.47979/aror.j.89.3.443-482.
Gooskens, Charlotte, John Nerbonne, Nathan Vaillette et al. 2007. Conditional entropy measures intelligibility among related languages. *LOT Occasional Series* 7. 51–66.
Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435–439. doi: https://doi.org/10.1038/nature02029.
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. doi: https://doi.org/10.1098/rstb.2010.0162.
Greenhill, Simon J, Paul Heggarty & Russell Gray. 2020. Bayesian phylolinguistics. In Richar D. Janda, Brian D. Joseph & Barbara S. Vance (eds.), *The handbook of historical linguistics, Volume II*, vol. 2, 226–253. New Jersey: Wiley Blackwell. doi: https://doi.org/10.1002/9781118732168.ch11.
Hill, Nathan W. 2019. *The historical phonology of Tibetan, Burmese, and Chinese*. Cambridge: Cambridge University Press.
Holland, Barbara R, Katharina T Huber, Andreas Dress & Vincent Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular biology and evolution* 19(12). 2051–2059. doi: https://doi.org/10.1093/oxfordjournals.molbev.a004030.
Honkasalo, Sami. 2019. A grammar of Eastern Geshiza: A culturally anchored description. Helsinki: University of Helsinki dissertation.
Huang, Bufan & Qingxia Dai (eds.). 1992. *Zangmian yuzu yuyan cihui* 藏缅语族语言词汇 [Vocabulary of Tibeto-Burman languages]. Zhongyang Minzu Daxue 中央民族大学 Minzu University of China.
Huang, Liangrong & Hongkai Sun. 2002. *Han-Jiarong cidian* 汉嘉戎词典 [Chinese-Gyalrong dictionary]. Beijing: Minzu chubanshe 民族出版社 Publishing House of Minority Nationalities.
Huson, Daniel H & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23(2). 254–267. doi: https://doi.org/10.1093/molbev/msj030.
Jacques, Guillaume. 2004. *Phonologie et morphologie du japhug (Rgyalrong)*. Paris: Université Paris Diderot (Paris 7) dissertation.
Jacques, Guillaume. 2012. Agreement morphology: The case of Rgyalrongic and Kiranti. *Language and Linguistics* 13.1. 83–116.
Jacques, Guillaume. 2014. *Esquisse de phonologie et de morphologie historique du tangoute*. Leiden: Brill.
Jacques, Guillaume. 2021. *A grammar of Japhug*. Berlin: Language Science Press. doi: https://doi.org/10.5281/zenodo.4548232.
Jacques, Guillaume & Zhen Chen. 2004. Chapuhua de chongdie xingshi 茶堡话的重叠形式 [reduplication in japhug]. *Minzu Yuwen* 民族语文 Minority languages of China 4. 7–11.
Jacques, Guillaume & Jade d'Alpoim Guedes. 2023. Sichuan peppercorn and the birth of numbing spices in East Asia. *Ethnobiology Letters* 14(1). 10–23. doi: 10.14237/ebl.14.1.2023.1842. https://ojs.ethnobiology.org/index.php/ebl/article/view/1842.
Juola, Patrick. 1998. Cross-entropy and linguistic typology. In D.M.W. Powers (ed.), *New methods in language processing and computational natural language learning*, 141–149.
Kroeber, Alfred L & C Douglas Chrétien. 1937. Quantitative classification of

Indo-European languages. *Language* 13(2). 83–103.
Lai, Yunfan. 2013. Erehua de fuyin chongdie 俄热话的辅音重叠 [The consonantal reduplication in Wobzi]. *Minzu Yuwen* 民族语文 Minority languages of China 2013(6). 12–18.
Lai, Yunfan. 2017. *Grammaire du khroskyabs de Wobzi*. Paris: Université Sorbonne Nouvelle (Paris 3) dissertation.
Lai, Yunfan. 2021. The complexity and history of verb-stem ablauting patterns in Siyuewu Khroskyabs. *Folia Linguistica* 55(1). 75–126. doi: https://doi.org/10.1515/folia-2020-2071.
Lai, Yunfan. 2022a. On the origins of Tangut long vowels, Invited talk at the Hong Kong Polytechnic University, 28 October 2022.
Lai, Yunfan. 2022b. On the origins of Tangut "long vowels", Invited talk at Trinity College Dublin, 23 November 2022.
Lai, Yunfan. 2022c. When internal reconstruction goes further: Proposing the vowel system of Proto-Khroskyabs through examining bound state apophony. *Folia Linguistica Historica* 56(s43-s1). 213–261. doi: https://doi.org/10.1515/flin-2022-2015.
Lai, Yunfan. 2023a. Lenition alternation in West Gyalrongic and its implication for Sino-Tibetan sound change typology. *Diachronica* 40(3). 341–383. doi: https://doi.org/10.1075/dia.21016.lai.
Lai, Yunfan. 2023b. On plosive-nasal correspondences and alternations in Gyalrongic and their possible solutions. *Cahiers de Linguistique Asie Orientale* 52(1). 1 – 39. doi: https://doi.org/10.1163/19606028-bja10027.
Lai, Yunfan. 2023c. Xixiayu de shengmu ruanhua yu Xibu Jiarongyuzu de bijiao 西夏语的声母软化与西部嘉绒语组的比较 A comparison between initial lenition between Tangut and West Gyalrongic. In *Dibajie Xixiaxue guoji xueshu luntan huiyi lunwenji* 第八届西夏学国际学术论坛会议论文集 [Proceedings of the 8th International Conference on Tangut Studies], 114–128. Ningxia: University of Ningxia.
Lai, Yunfan. forthcoming in 2025. Long vowels and the origin of stem alternation in Khroskyabs. *Bulletin of Chinese Linguistics (A collection of essays by 2022 Outstanding Young Scholars of the Li Fang-Kuei Society)* .
Lai, Yunfan, Xun Gong, Jesse P. Gates & Guillaume Jacques. 2020. Tangut as a West Gyalrongic language. *Folia Linguistica Historica* 41(1). 171–203. doi: https://doi.org/10.1515/flih-2020-0006.
Lai, Yunfan & Johann-Mattis List. 2020. Phylogeny of Gyalrongic languages: How to do and what to expect. Paper presented at the 53rd International Conference on Sino-Tibetan languages and linguistics.
Leskien, August. 1876. *Die Declination im Slavisch-Litauischen und Germanischen*. Leipzig: Hirzel.
Li, Fanwen. 1997. *Xia-Han zidian* 夏漢字典 [Tangut-Chinese dictionary]. Beijing: China Social Sciences Press.
Li, Fanwen & Xiangrong Lin. 1983. Shilun Jiarongyu yu daofuyu de guanxi – Jianlun Xixiayu yu Daofuyu, Jiarongyu, Zangyu de guanxi 試論嘉戎語與道孚語的關係 – 兼論西夏語與道孚語、嘉戎語、藏語的關係 [An exploration of the relationship between Gyalrong and Stau Languages – Along with a discussion on the relationship between Tangut, Stau, Gyalrong, and Tibetan]. In *Xixia yanjiu lunji* 西夏研究論集 [Collection of Tangut Studies], 279–305. Ningxia Renmin Chubanshe 宁夏人民出版社 Ningxia people's Publishing House.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45(1). 137–161. doi: https://doi.org/10.1162/coli_a_00344.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38(2). 210–258. doi: https://doi.org/10.1075/dia.20004.mac.

Miyake, Marc Hideo. 2012. Complexity from compression: A sketch of Pre-Tangut. In Irina Popova (ed.), *Тангуты в Центральной Азии: Сборник статей в честь 80-летия проф. Е.И.Кычанова* [Tanguts in Central Asia: a collection of articles marking the 80th anniversary of Prof. E. I. Kychanov], Moscow: Oriental Literature.

Morrison, David A. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. *Molecular Biology and Evolution* 27(5). 1044–1057. doi: https://doi.org/10.1093/molbev/msp309.

Osthoff, Hermann & Karl Brugmann. 1878. *Morphologische untersuchungen auf dem gebiete der indogermanischen sprachen*. Leipizg: S. Hirzel.

Pellegrini, Matteo. 2020. Patterns of interpredictability and principal parts in Latin verb paradigms: An entropy-based approach. *Journal of Latin Linguistics* 19(2). 195–229. doi: https://doi.org/10.1515/joll-2020-2014.

Sagart, Laurent. 2011. Classifying Chinese dialects/Sinitic languages on shared innovations. Séminaire Sino-Tibétain du CRLAO.

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: https://doi.org/10.1073/pnas.1817972116.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.

Stein, Rolf A. 1966. Nouveaux documents tibétains sur le Mi-Ñag/Si-Hia. In *Mélanges de sinologie offerts à Monsieur Paul Demiéville i*, 281–289. Paris: Presses Universitaires de France.

Sun, Jackson T.-S. & Blogros Bstan'dzin. 2019. *Tshobdun Rgyalrong spoken texts with a grammatical introduction*. Taipei: Academia Sinica.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contact. *Proceedings of the American Philosophical Society* 96. 452–63.

Wurzel, Wolfgang Ulrich. 1989. *Inflectional morphology and naturalness*. Dordrecht: Kluwer.

Yang, Chih-Fan. 2021. *The morpho-syntax of tense, aspect, evidentiality and modality in Bawang Horpa*: National Taiwan Normal Univeristy dissertation.

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

Zhang, Shuya. 2020. *Le rgyalrong situ de brag-bar et sa contribution à la typologie de l'expression des relations spatiales: L'orientation et le mouvement associé*: Institut National des Langues et Civilisations Orientales dissertation.

Zhang, Shuya. 2023. Towards a new generalisation of the tri-axial orientation system in Situ Rgyalrong. *Transactions of the Philogical Society* 121(2). 203–225. doi: https://doi.org/10.1111/1467-968X.12264.

Zhang, Shuya. forthcoming. Xixiayu quxiangqianzhui yu dongci de quxiang keyixing 西夏

語趨向前綴與動詞的趨向可易性 [Directional prefixes and verbal orientability in Tangut]. *Language and Linguistics* .
Zhang, Shuya, Guillaume Jacques & Yunfan Lai. 2019b. A study of cognates between Gyalrong languages and Old Chinese. *Journal of Language Relationship* 17(1). 73–92. doi: https://doi.org/10.31826/jlr-2019-171-210.
Zhou, Kevin & Claire Bowern. 2015. Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B: Biological Sciences* 282(1815). 20151278. doi: https://doi.org/10.1098/rspb.2015.1278.

Appendix

The reader is invited to download the Supplementary Material via the following link: https://doi.org/10.5281/zenodo.10004559
For any additional information and data, please contact the author directly.

Résumé

Cet article utilise les langues gyalronguiques, une branche conservatrice du sino-tibétain, pour illustrer une nouvelle méthode d'évaluation des reconstructions de proto-langues dans la linguistique historique générale, ainsi que pour mener des analyses exploratoires dans la phylogénie des langues. Il commence par reconstruire un proto-système des préinitiales gyalronguiques, calcule et compare les entropies implicatives entre les systèmes reconstruits et modernes. Dans une deuxième étape, l'entropie implicative mutuelle (Mutual implicative entropy, MIE) est utilisée pour mesurer la distance génétique entre les langues apparentées et générer des réseaux Neighbornet pour visualiser la sous-classification des langues gyalronguiques. Les réseaux résultants sont en accord avec les analyses linguistiques historiques qualitatives et permettent des ajustements par rapport aux sous-classifications précédentes obtenues par l'inférence phylogénétique bayésienne. Ainsi, cette méthode peut être utilisée pour détecter des nuances dans les sous-branches inférieures, parfois négligées par les méthodes basées sur le lexique. L'utilisation de la MIE en linguistique historique est donc un moyen rapide et efficace de vérifier l'efficacité des reconstructions et d'élaborer une forme préliminaire précise de la sous-classification linguistique.

Zusammenfassung

Dieser Artikel verwendet die gyalrongischen Sprachen, ein konservativer Zweig des Sino-Tibetischen, um eine neue Methode zur Bewertung von Rekonstruktionen von Ursprungssprachen in der allgemeinen historischen Linguistik zu illustrieren und um explorative Analysen in der Sprachphylogenie durchzuführen. Zunächst wird ein Proto-System der gyalrongischen Präinitiale rekonstruiert, und die implikative Entropie zwischen rekonstruierten und modernen Systemen wird berechnet und verglichen. In einem zweiten Schritt wird die gegenseitige implikative Entropie (Mutual implicative entropy, MIE) verwendet, um die genetische Distanz zwischen verwandten Sprachen zu messen und Neighbornet-Netzwerke zu generieren, um die Untergruppierung der gyalrongischen Sprachen zu visualisieren. Die resultierenden Netzwerke stimmen mit

qualitativen Analysen in der historischen Linguistik überein und ermöglichen Anpassungen an vorherige Untergruppierungen, die durch die bayessche phylogenetische Inferenz erhalten wurden. Diese Methode kann daher dazu genutzt werden, Nuancen in unteren Teilzweigen zu erkennen, die manchmal von lexikonbasierten Methoden vernachlässigt werden. Die Verwendung von MIE in der historischen Linguistik ist somit ein schnelles und effizientes Mittel, um die Wirksamkeit von Rekonstruktionen zu überprüfen und eine genaue vorläufige Form der Sprachuntergliederung zu erstellen.

Author's address

Yunfan Lai
Trinity Centre for Asian Studies
Trinity College Dublin
D02 PN40 Dublin

yunfan.lai@tcd.ie