Suantak, Khawlsonkim, Marc Allassonnière-Tang, Eugene Chan, Anthony Chi-Pin Hsu, and One-Soon Her*. 2026. A GIS view of the word orders of numeral bases and numeral classifiers in Kuki-Chin languages. *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqaf133

# A GIS View of the Word Orders of Numeral Bases and Numeral Classifiers in Kuki-Chin Languages

Khawlsonkim Suantak[1] Marc Allassonnière-Tang[2] Eugene Chan[3]
Anthony Chi-Pin Hsu[4] One-Soon Her*[4,5]
[1]Tripura University [2]National Museum of Natural History, Paris [3]Independent researcher, Hong Kong
[4]Tunghai University [5]National Chengchi University

## Abstract

Tibeto-Burman (TB) languages are known for their diversity in numeral systems and classifiers. This paper investigates the Kuki-Chin (KC, also called South-Central Tibeto-Burman) languages in TB's Northeast Indian Areal Group and provides a comprehensive description of the four types of numeral systems: base-final, base-initial, base-split, and no-base, and the four types of classifier systems: CL-final, CL-initial, CL-split, and no-CL. A thorough survey of the literature, aided by fieldwork, enables a GIS view of the distribution of the different types of languages. Also, KC languages conform to Greenberg's Universal 20A: N does not come between Num and CL, and the numeral base and classifier harmonize in word order. We propose the following hypothesis for Proto-KC (PKC) and Proto-TB (PTB) to account for the variation in numeral bases and classifiers in KC, i.e., PKC, like PTB, is base-initial and without numeral classifiers, and the current variation in numeral bases and numeral classifiers in KC is due to horizontal external influence via language contact. Bayesian phylogenetic inference tests, however, show mixed results, as PKC is likely to be base-initial, as expected, but CL-initial, and PTB is likely to be base-final but CL-initial. Thus, we plan to conduct a comprehensive survey of all TB languages and further explore the state of PTB in terms of numeral bases and classifiers.

Keywords: Tibeto-Burman, Kuki-Chin, numeral base, numeral classifier, word order, GIS

## 1. Introduction

The majority of languages have multiplicative complex numerals formed with a multiplier and a numeral base, e.g., three hundred [3×100], where the numeral base hundred serves as the multiplicand (Comrie 2013). However, while most of the modern-day languages of primary cultures have a base-final order, like English and Chinese, most African languages have the opposite base-initial order, e.g., ikie ita [100×3] in Ibibio (Niger-Congo) and all languages in Eurasia are base-final, except some of the Tibeto-Burman (TB) languages[1], e.g., za-thum [100×3] in Vaiphei (Her et al. 2024).

Thus, in the Sino-Tibetan family, while Sinitic languages are entirely uniform in having a base-final decimal numeral system (Her 2017a), TB languages are rather diversified in having both base-final and base-initial complex numerals (Li, Yang, and Her 2018). Interestingly, while Sinitic languages are likewise uniform in having numeral classifiers (NC), which consistently appear after the numeral, hence NC-final, TB languages are again diversified, as some do, and others do not, have NCs and for those that do, both CN-final and NC-initial orders are attested (Her, Tang, and Li 2019).

This study investigates the Kuki-Chin (KC) languages, an understudied TB group where base-initial languages seem to congregate (Her et al. 2024). The group consists of some 54 languages with more than a million speakers, situated mainly in Manipur, Assam, Mizoram, Nagaland, and Tripura in northeastern India, also found in Chittagong Hills District in Bangladesh and Rakhine and Chin states and Sagaing divisions in Myanmar (VanBik 2009). Fig. 1 is a GIS map of the distribution of KC languages. Each point represents a language. The coordinates are extracted from Glottolog (Hammarström et al. 2024). Fig. 2 is the Sino-Tibetan family tree, where KC is under the Northeast Indian Areal Group of the TB branch.

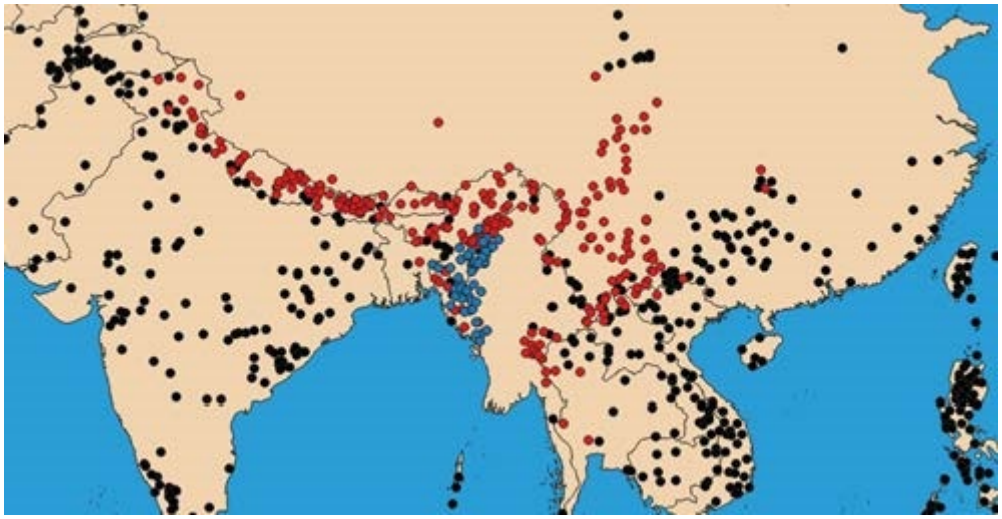


Fig. 1 Distribution of Kuki-Chin languages (blue dots: KC, red dots: TB, black dots: non-TB)

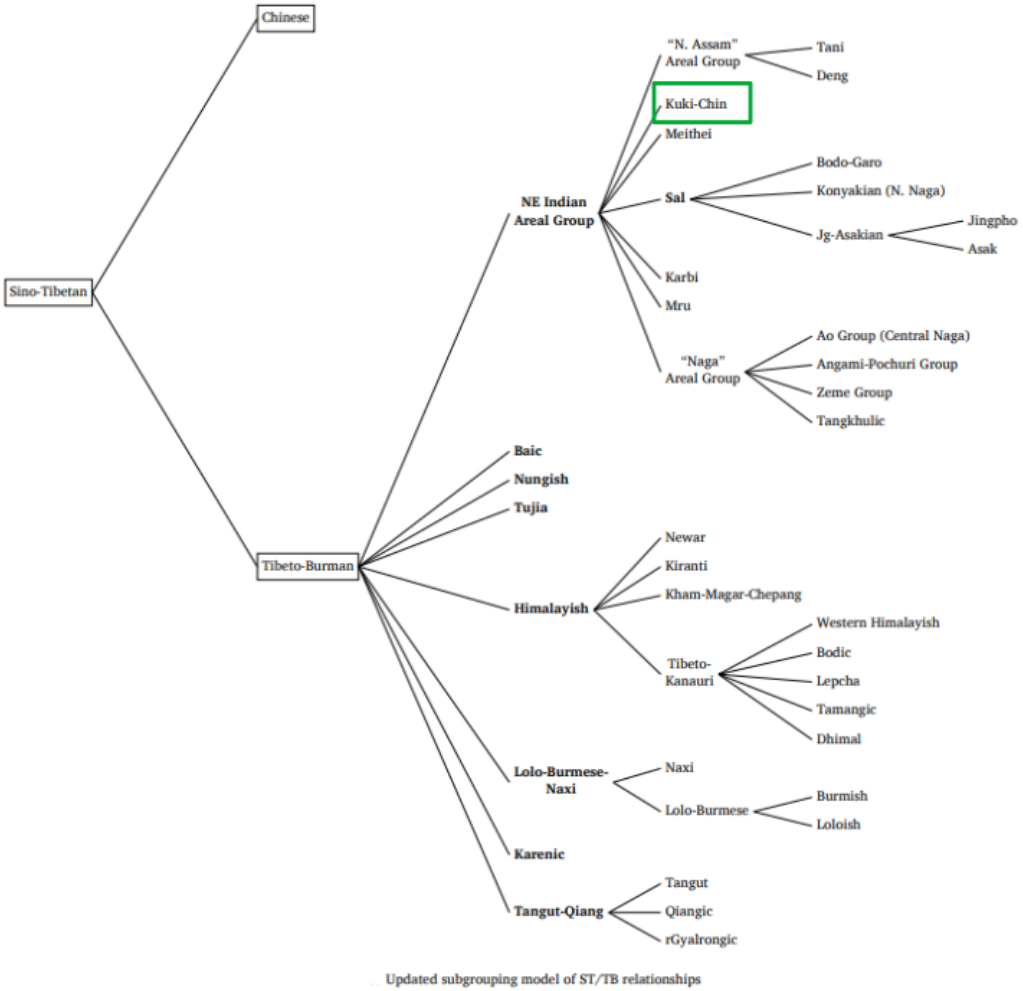



Fig. 2 Sino-Tibetan family tree (Matisoff 2015: xxxii). The Kuki-Chin branch is highlighted in green.

The goal of the study is to conduct a comprehensive survey of KC languages in terms of base orders and, where NCs are present, NC orders, and then provide an account for their distribution within the TB group. This paper is organized as follows. Section 2 gives the background of the study and our motivation. Section 3 focuses on the results of the extensive survey on KC numeral bases and numeral classifiers. We then discuss their distribution in Section 4, offering a qualitative analysis and a

quantitative account of Bayesian phylogenetic inference tests. Section 5 presents discussions, conclusions, and plans for future research.

## 2. Numeral Bases and Numeral Classifiers

In this section, we demonstrate first that numeral bases function as multiplicands in the mathematical formation of a multiplicative numeral. We then illustrate the advances in the study of numeral classifiers by recognizing that the classifier likewise functions as a multiplicand, just like the numeral base. The typology of the classifier word order is thus crucially linked to the typology of base word order. We present a GIS map for the worldwide distribution of the four types of base orders in 2.3. The intriguing patterns of base types and numeral classifiers in TB are summarized in 2.4, which motivates this study on Kuki-Chin languages specifically.

### 2.1 Numeral bases as multiplicands

In the study of numeral systems and numeral classifiers in the world's languages, significant advances have been made in the last decade in terms of both theoretical accounts and the construction of databases. As an informal part of Max Planck Institute's (MPI) *The World Atlas of Language Structures Online* (WALS, https://wals.info/) project, the engineer Eugene Chan's (2024) website, named Numeral Systems of the World's Languages (https://lingweb.eva.mpg.de/channumerals/), contains the limited raw data of the numeral systems of some 4,000 languages. Contrary to common misconception, not all numeral systems in human languages employ multiplication, as some have addition only and some only have simple numerals, and a small number of languages are even without exact numerals. However, it is true that most of the languages thus far documented employ addition and multiplication in the general pattern in (1) (Comrie 2013); *b* refers to a *numeral base*, and *n* stands for *multiplier*.

(1) General Pattern of Numerals
For base *b*: $(n \times b) + m$, where $m < b$ and $n \leq b$
E.g., *sān-bǎi èr-shí yī* [[3×100] + [2×10] + [1]] 'three hundred twenty-one'

The Mandarin example in (1) shows *bǎi* 'hundred' and *shí*' ten' as numeral bases, serving as multiplicands, with *sān* 'three' and *èr* 'two' as the respective multiplier. The Commutative Law of Multiplication recognizes the reversible order between *n* and *b*. While most of the familiar dominant modern languages, e.g., Mandarin, English, Japanese, and Spanish, have a base-final, i.e., $(n \times b)$ numeral system, a large number of languages employ a base-initial, i.e., $(b \times n)$, numeral system, e.g., Kilivila (Austronesian).

(2) Base-initial Numerals in Kilivila (Senft 1986:77-80)

| | | |
|---|---|---|
| luwa-tala | (10×1) | '10' |
| luwa-yu | (10×2) | '20' |
| luwa-tolu | (10×3) | '30' |
| lakatu-tala | (100×1) | '100' |
| lakatu-yu | (100×2) | '200' |
| lakatu-tolu | (100×3) | '300' |

The significance and implications of this word order parameter were largely overlooked in the literature. For example, Comrie (2013) states: 'The order of elements

is irrelevant, as are the particular conventions used in individual languages to indicate multiplication and addition.' There thus had not been any large-scale surveys of languages in terms of this word order parameter, hence no studies on the relevance and significance of this important typological parameter either, until recently (e.g., Her 2017a; 2017b).

Not surprisingly, the debate then extends to the controversy over the constituency of additive complex numerals, e.g., *shí yī* [ten + one] in Mandarin, and multiplicative complex numerals, e.g., *sān-bǎi* [three×hundred]. The dominant view in formal semantics and formal syntax is that such complex numerals do not form a constituent to the exclusion of the noun, which can be overt or covert (e.g., Ionin and Matushansky 2006; Ionin and Matushansky 2018). We have likewise followed through the mathematical evidence and argued that complex numerals are indeed constituents syntactically (He et al., 2017; He and Her 2022). Interestingly, Wągiel and Caha (2020) propose a universal account for all numerals, where the classifier feature is universal, numerals are constituents, and [Num CL] is also a constituent.

The WACL project (Her et al. 2022) was inspired by Greenberg's (1990a[1972], p. 172) insight that numeral classifiers function as a multiplicand and his observation in Greenberg (1990b[1978], p. 293) that the numeral base and the numeral classifier harmonize in word order, formalized in Her (2017a; 2017b) as an implicational universal in (3).

(3) Word order harmonization between base and CL (Her 2017a: 280)
a. CL-final order, or [Num Cl] ⇨ base-final numerals, or ($n \times b$)
b. CL-initial order, or [Cl Num] ⇨ base-initial numerals, or ($b \times n$)

The multiplicative theory of numeral classifiers predicts the CL-base harmonization, thus also in effect predicting that a classifier language should have multiplicative bases in its numeral system. This is stated as a probabilistic implication universal as well, as in (4).

(4) Co-occurrence of numeral bases and classifiers in languages:
Presence of classifiers ⇨ Presence of multiplicative numerals

These predictions have been verified with statistical significance in 219 classifier languages from six language groups, i.e., Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai, Tibeto-Burman, and Indo-Aryan (Her, Tang, and Li 2019). Further studies explored the synchronization of word order between numeral bases, numeral classifiers, and nouns (Allassonnière-Tang and Her 2020), showing that morphosyntactic plural markers and sortal numeral classifiers are alike in encoding countness and functioning mathematically as a multiplicand with the precise numerical value of *one* (Her and Tang 2020).

## 2.2 Typology of numeral base orders

The open-access WACL database covers 3338 languages with information on numeral classifiers and the relevant base orders (Her et al. 2022), and the open-access database reported by Her et al. (2024) expanded to 4099 languages with information on base orders. The current database has 4104 languages, where each language is marked with information on the presence/absence of numeral classifiers, classifier word order, presence/absence of numeral bases, and base order, along with information on the language's genealogical position, its location in terms of geographical coordinates, the

population size of its native speakers, and some of the important references in relation to this language. See Fig. 3, a GIS map of the 4104 languages showing the worldwide distribution of numeral systems with different base orders.

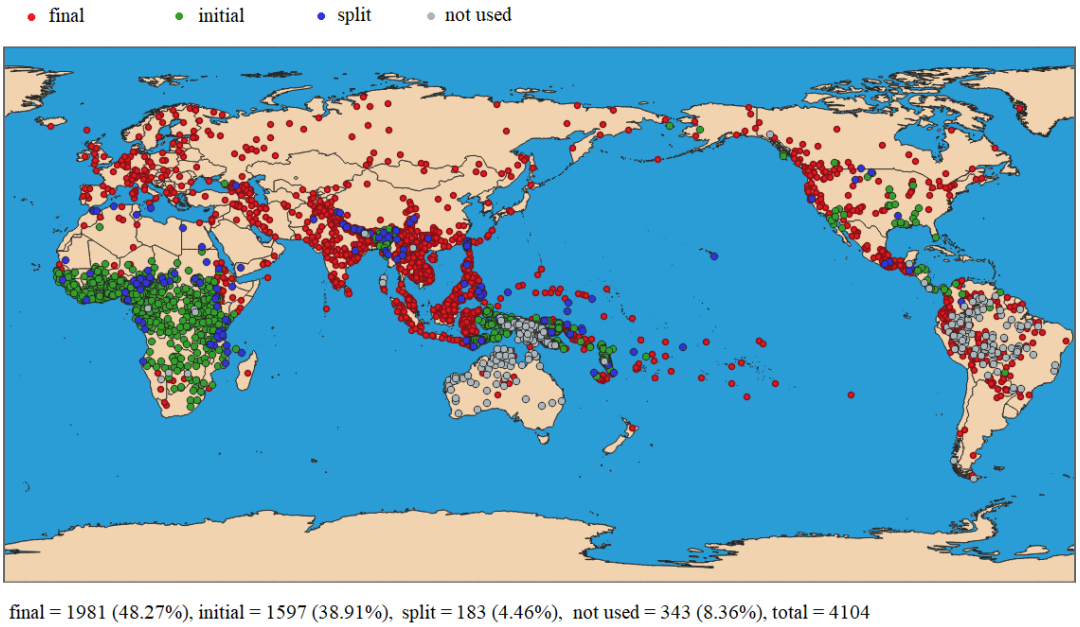



Fig. 3 Worldwide distribution of numeral systems in 4104 languages

The distribution pattern of base-final systems, base-initial systems, base-split systems, and no-base systems does not seem random at all, and Her et al. (2024) propose a set of six hypotheses to account for it. However, what especially caught our attention is the fact that on the Asia continent, the only hotbed of the green dots, i.e., base-initial languages, seem to be within the TB family.

## 2.3 Base-initial numeral systems in TB

Fig. 4 shows the geographical distribution of the 312 TB languages (red dots). Note that while nearly all other languages (gray dots) on the Asian continent are base-final, the TB language group sandwiched in between is the only exception.

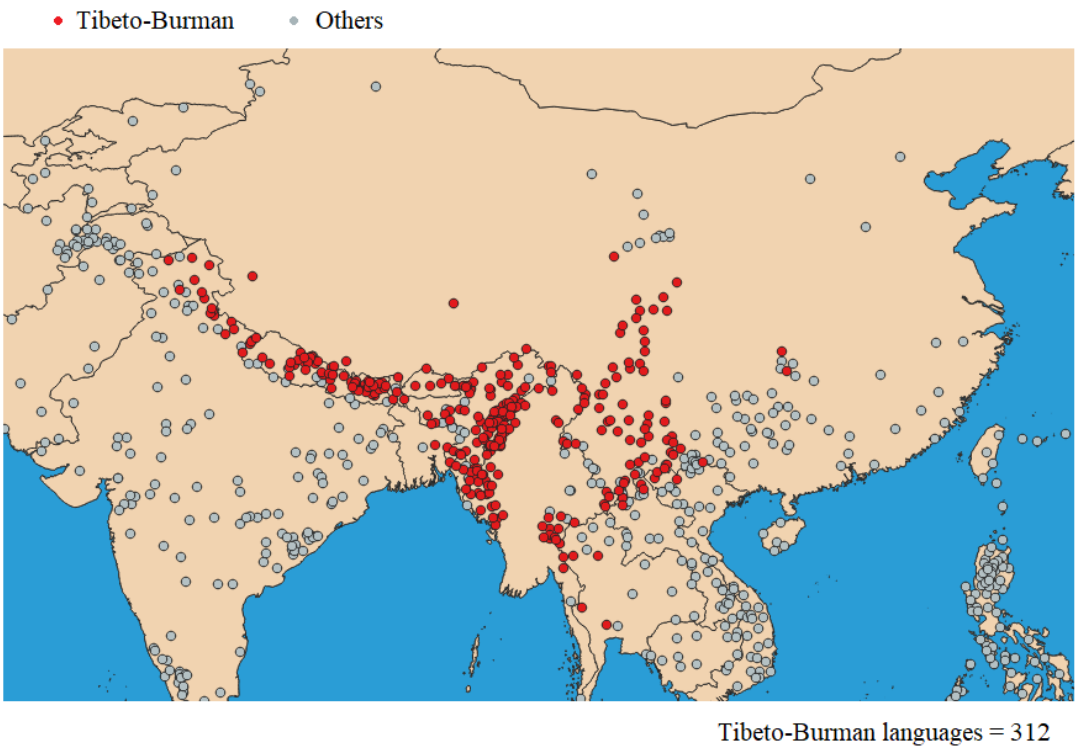
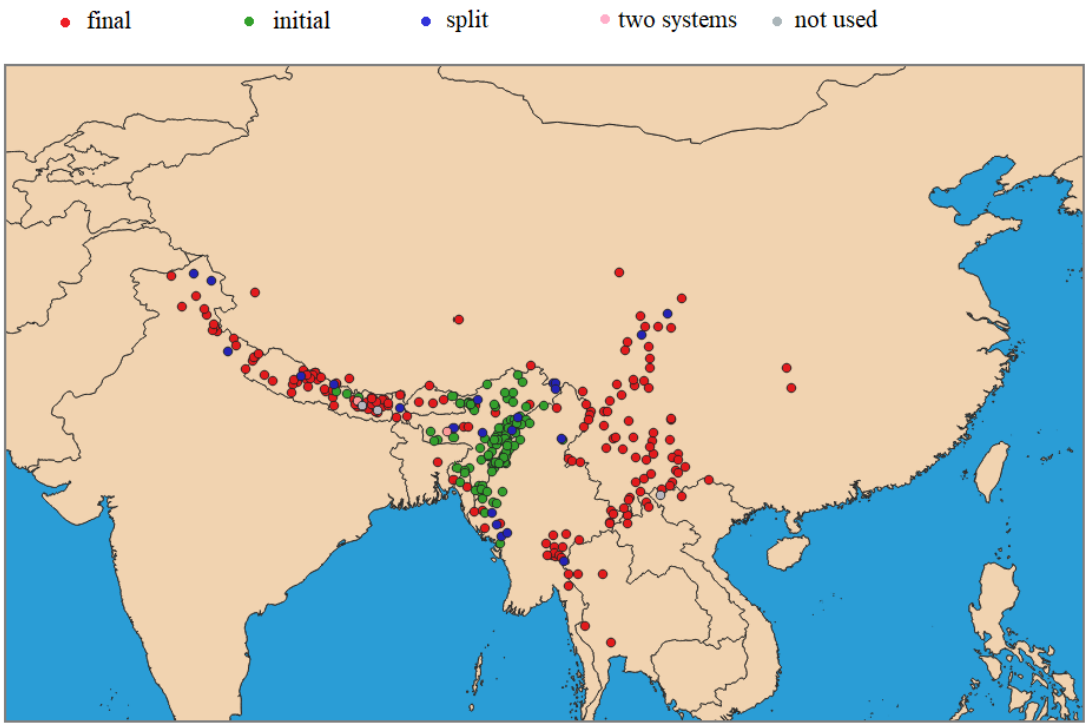



Fig. 4 The Tibeto-Burman sandwiched between base-final languages

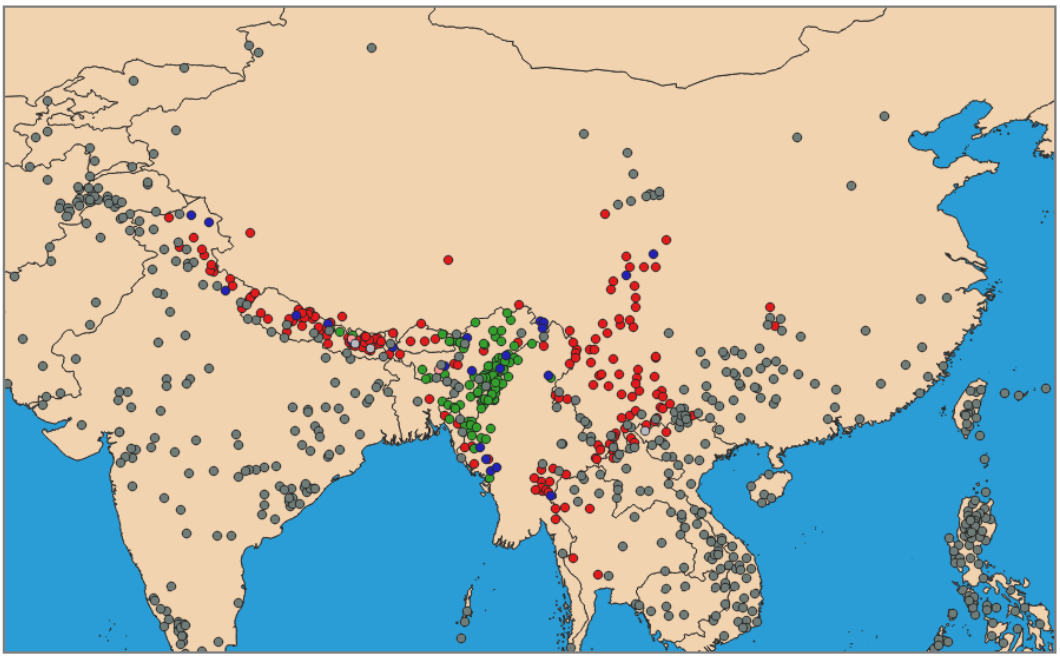
All four types of numeral systems are also found in TB: out of the 312 languages documented in the current database, 100 (32.05%, 100/312) are base-initial, 184 (58.97%, 184/312) are base-final, 23 (7.37%, 23/312) have both orders, and 3 (0.96%, 3/312) have no bases. Fig. 5 shows the distribution of the four types of base orders in TB. Note that the base-initial TB languages (green dots) are sandwiched between base-final TB languages (red dots). In particular, note the concentration of the base-initial feature in the northeastern India region, where most of the KC languages are located.




final = 184 (58.97%), initial = 100 (32.05%), split = 23 (7.37%), two systems = 2 (0.64%), not used = 3 (0.96%), total = 312

Fig. 5 Base-initial TB languages sandwiched between base-final TB languages

Fig. 6 shows the four types of TB languages nestled in the neighboring non-TB languages (gray dots) on the Asian continent, which, again, are nearly all base-final.




final = 184 (58.97%), initial = 100 (32.05%), split = 23 (7.37%), two systems = 2 (0.64%), not used = 3 (0.96%), total = 312

Fig. 6 Distribution of the four types of numeral systems in TB on the Asian continent

It is likely that early humans used base-initial numerals (Her et al. 2024). Such a pattern of distribution strongly suggests that base-initial TB languages are the most conservative and have so far resisted the influence of the base-final force dominant on the Asian continent. It has indeed been claimed that in Proto-TB (PTB) multiplicative numerals are base-initial (Kiryu 2009: 64-66; Her and Li 2023). Note that Kiryu (2009: 65) states: 'The Proto-Tibeto-Burman order of numeral and quantifier was Num-Q.' Crucially, 'Num-Q' here, or base-final in our terminology, is clearly a typo given the previous context. We have contacted the author and confirmed with him that what he meant to say was 'Q-Num', i.e., base-initial, instead. Kiryu (2009: 64-66) thus further suggests that the base-final order in TB is due to contact with neighboring base-final language groups. Her and Li (2023) further cite Fu's (2015: 50) view that the default order between N and Num in TB is [N Num], thus head-initial, as supporting evidence, as base-initial numerals are likewise head-initial. This base-initial PTB hypothesis thus suggests that the base-final order in TB is due to external influences, and the split systems in TB are in a transitional stage.

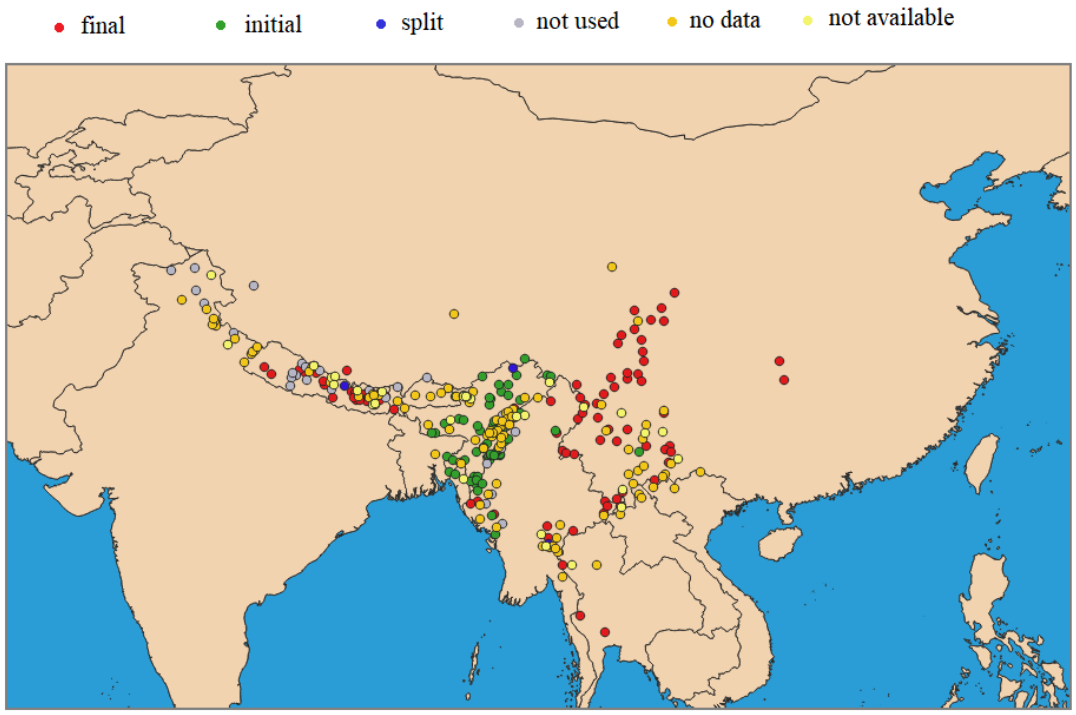
## 2.4 Numeral classifiers in TB

To the best of our knowledge, no TB specialist has ever claimed that the classifier feature is indigenous in any TB language or in PTB, except perhaps Matisoff (2003: 609). In his *Handbook of Proto-Tibeto-Burman*, Matisoff proposes a single classifier **ra* in Index of Proto-Forms. Yet, as aptly demonstrated by Her and Li (2023) with several pieces of evidence, instead of a classifier for humans, **ra* is a noun with the meaning of 'person'. Evidence available thus strongly suggests that PTB does not have the classifier feature and all TB classifier languages acquired this feature via contact with neighboring classifier languages. The word order distribution of classifiers in TB is thus very intriguing.

Out of the 312 TB languages, the database has information on the classifier feature in 180 languages; see Table 1 for a summary in terms of the classifier word order. Fig. 7 shows the geographical distribution of the 312 TB languages.

Table 1 Distribution of classifier languages in TB

| CL feature | No. of languages | Percentage |
|---|---|---|
| CL-final | 87 | 27.88% |
| CL-initial | 62 | 19.87% |
| CL-split | 4 | 1.28% |
| no-CL | 27 | 8.65% |
| No Data | 102 | 32.69% |
| Not Available | 30 | 9.62% |
| Total | 312 | 99.99% |




final = 87 (27.88%), initial = 62 (19.87%), split = 4 (1.28%), not used = 27 (8.65%),
no data = 102 (32.69%), not available = 30 (9.62%), total = 312

Fig. 7 Distribution of classifier languages in TB

Given that PTB is base-initial and without CLs, it follows that TB languages without CLs are most conservative, followed by those with classifiers also in a CL-initial order, thus observing the base-CL harmonization. Those with the CL-split orders are in a transitional state, likely to further develop and join the CL-final TB classifiers languages. Note that the CL-initial feature, like the base-initial feature, again appears to concentrate in the northeastern India region.

## 3. Numeral Bases and Numeral Classifiers in Kuki-Chin

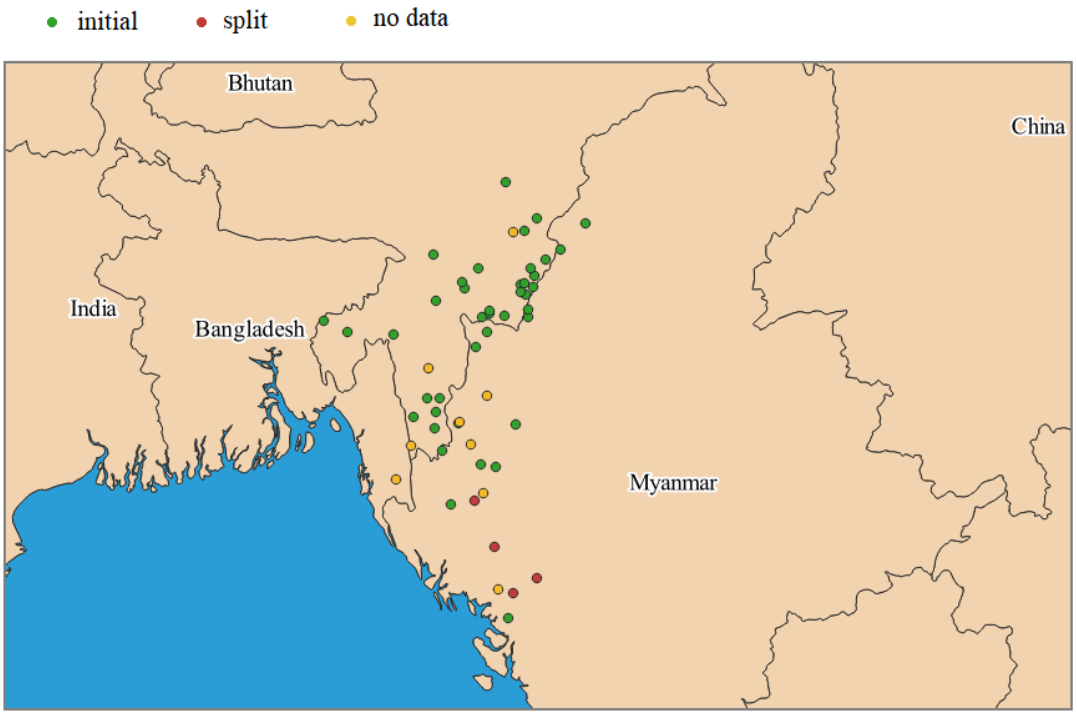
Upon a careful deliberation of the intriguing patterns of the distribution of numeral bases and numeral classifiers in TB and the enticing concentration of base-initial and CL-initial languages in the northeastern India region, we have decided that a thorough and targeted investigation of the South-Central Tibeto-Burman languages, *aka* Kuki-Chin, is the best starting point, and indeed a necessary first step, to a full understanding of how the TB languages came to have this pattern of distribution of the four types of numeral bases and numeral classifiers.

In the current database, 54 KC languages have been recorded with information on their base feature, while still 9 languages lack reliable data. Table 2 is a summary, and Fig. 8 shows the geographical distribution of the 54 KC languages.

Table 2 Distribution of the base feature in KC languages

| Base feature | No. of languages | Percentage |
|---|---|---|
| base-final | 0 | 0% |
| base-initial | 41 | 75.93% |
| base-split | 4 | 7.41% |
| No data | 9 | 16.67 |

| | | |
|---|---|---|
| total | 54 | 100.01% |




initial = 41 (75.93%), split = 4 (7.41%), no data = 9 (16.67%), total = 54

Fig. 8 Distribution of the base feature in TB

Notably, all KC languages documented are still base-initial, except the four languages with both base-final and base-initial numerals, known as 'split' systems. As for the classifier feature, 39 KC languages have been recorded with information on the classifier feature, while still 15 languages require data. Table 3 is a summary, and Fig. 9 shows the geographical distribution of the known KC classifier languages and non-classifier languages among the known TB classifier languages and non-classifier languages.

Table 3 Distribution of the classifier feature in KC languages

| CL feature | No. of languages | Percentage |
|---|---|---|
| CL-final | 0 | 0% |
| CL-initial | 32 | 59.26% |
| CL-split | 0 | 0% |
| no-CL | 7 | 12.96% |
| No data | 15 | 27.78% |
| total | 54 | 100% |

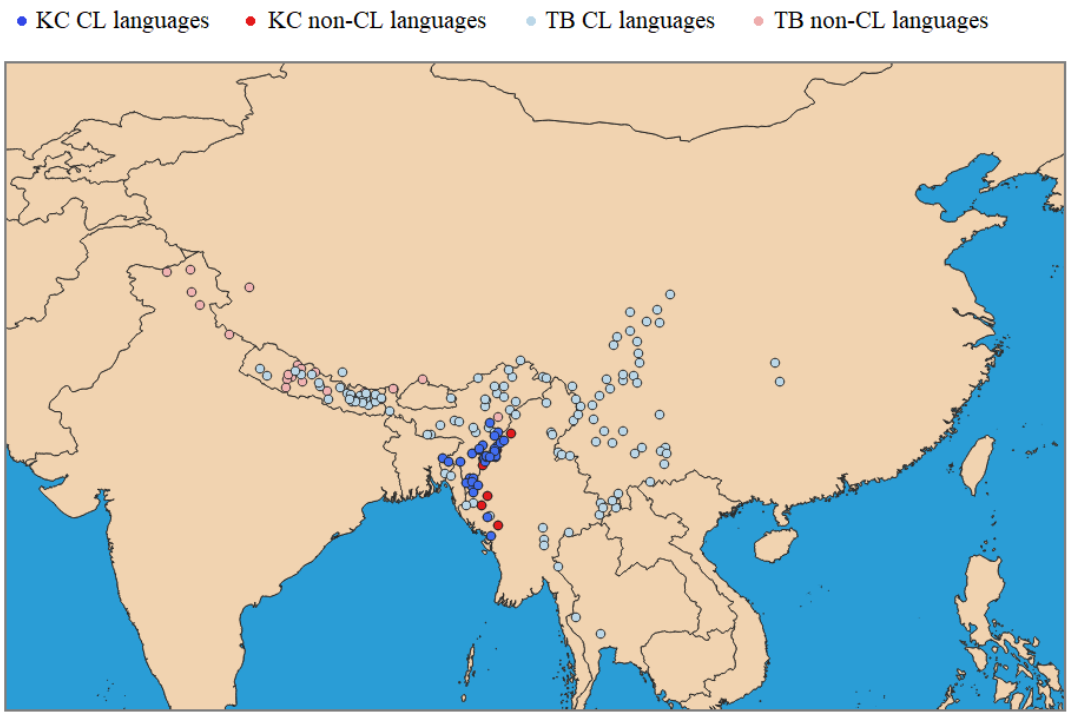



Fig. 9 Distribution of the classifier feature in KC languages and TB classifier languages

Table 4 The base feature and classifier feature in KC languages

| Language Name | Base order | CL order |
|---|---|---|
| Aimol | base-initial | CL-initial |
| Anal | base-initial | CL-initial |
| Asho Chin | base-split | no-CL |
| Bawm Chin | base-initial | CL-initial |
| Biete | base-initial | CL-initial |
| Bualkhaw Chin | no-data | no-data |
| Chinbon Chin | base-split | no-data |
| Chiru | base-initial | CL-initial |
| Chothe | base-initial | CL-initial |
| Daai Chin | base-split | CL-initial |
| Darlong | base-initial | CL-initial |
| Eastern Khumi Chin | base-initial | no-data |
| Falam Chin | base-initial | CL-initial |
| Gangte | base-initial | CL-initial |
| Haka/Hakha Chin | base-initial | CL-initial |
| Halam | base-initial | CL-initial |
| Hmar | base-initial | CL-initial |
| Hrangkhol/Hrangkhaw l | base-initial | CL-initial |
| Kaang Chin | no-data | no-data |
| Kharam Naga | no-data | no-CL |
| Khumi Chin | base-initial | CL-initial |
| Koireng | base-initial | CL-initial |
| Kom (India) | base-initial | CL-initial |
| Lamkang | base-initial | CL-initial |

| | | |
|---|---|---|
| Laamtuk Thet | base-initial | no-data |
| Lautu | no-data | no-data |
| Mizo | base-initial | CL-initial |
| Mara Chin | base-initial | no-CL |
| Monsang Naga | base-initial | no-data |
| Moyon | base-initial | no-CL |
| Mro Chin | base-initial | CL-initial |
| Mün Chin | base-split | no-CL |
| Nga La | no-data | no-data |
| Ngawn Chin | base-initial | no-data |
| Paite Chin | base-initial | CL-initial |
| Pankhu (Pangkhu, Pangkhua, Paang) | base-initial | CL-initial |
| Purum | base-initial | no-CL |
| Ralte | base-initial | CL-initial |
| Rengmitca | no-data | no-data |
| Sakachep | base-initial | CL-initial |
| Senthang Chin | base-initial | no-CL |
| Shendu | no-data | no-data |
| Simte | base-initial | CL-initial |
| Sizang Chin | base-initial | CL-initial |
| Sorbung | base-initial | no-data |
| Sumtu Chin | no-data | no-data |
| Syriem | base-initial | no-data |
| Tarao | base-initial | CL-initial |
| Tedim Chin/Thadou | base-initial | CL-initial |
| Thado Chin | base-initial | CL-initial |
| Vaiphei | base-initial | CL-initial |
| Zotung Chin | no-data | no-data |
| Zou | base-initial | CL-initial |
| Zyphe/Zophei | base-initial | CL-initial |

## 4. Hypotheses and A Quantitative Account

To account for the variation in numeral bases and classifiers in KC, we shall recruit two of Greenberg's hypothetical universals and propose a hypothesis regarding Proto-KC. First, Greenberg's Universal 20A is about the classifier word order in a nominal phrase (Greenberg 1990b:185; Her 2017a), as stated in (5).

(5) Greenberg's Universal 20A
Of the three elements Num, CL, and N, any order is possible as long as N does not come between Num and CL.

Thus, out of the six mathematically possible orders among three different elements in (6), only four are allowed in languages.

(6) Six Possible Word Orders of [Num, CL, N]
a. √ [Num CL N] (many languages, e.g., Chinese)
b. √ [N Num CL] (many languages e.g., Thai)
c. √ [CL Num N] (few languages e.g., Ibibio [Niger-Congo])

d. √ [N CL Num] (few languages e.g., Jingpho [Tibeto-Burman])
e. * [CL N Num] (no languages)
f. * [Num N CL] (no languages)

A more revealing taxonomy of the four legal word orders in (6) can be the results of the interaction between two binary parameters: CL-initial/CL-final and N-initial/N-final, as shown in Table 5.

Table 5 Taxonomy of the universal CL word orders

| | N-FINAL | N-INITIAL |
|---|---|---|
| CL-FINAL | (A) [Num CL N] | (B) [N Num CL] |
| CL-INITIAL | (C) [CL Num N] | (D) [N CL Num] |

Recall that since Greenberg (1990b) this universal has received a mathematical interpretation under the multiplicative theory of numeral classifiers, i.e., with CL as a multiplicand and Num as the multiplier, the two form a multiplicative unit that cannot be interrupted. We anticipate that all KC classifier languages will have the less common orders of (6c) and (6d).

Next, Greenberg (1990[1978]:292) observes that the order between Num and C/M and the order between *n* and *base* should 'harmonize'. This implicational universal is stated in (7), repeated from (3).

(7) Word order harmonization between base and CL (Her 2017a: 280)
a. CL-final order, or [Num Cl] ⇨ base-final numerals, or ($n \times b$)
b. CL-initial order, or [Cl Num] ⇨ base-initial numerals, or ($b \times n$)

Thus far all KC classifier languages turn out to be base-initial and all attested cases of (7b). Consequently, KC languages all confirm the proposed universal that a language with numeral classifiers also have numeral bases; see (8), repeated from (4).

(8) Co-occurrence of numeral bases and classifiers in languages:
Presence of classifiers ⇨ Presence of multiplicative numerals

Notably, in the extensive data we have collected, all KC languages have the base-initial order, only four have the base-final order alongside the base-initial order, and none have the CL-final order. All KC classifier languages have multiplicative numeral bases, and all KC classifier languages comply with base-CL harmonization. We thus hypothesize that PKC is base-initial, which is consistent with the existing hypothesis that PTB is base-initial (Kiryu 2009: 64-66; Her and Li 2023) and PKC's base-initial feature is thus inherited from PTB. Furthermore, based on the data in Matisoff's (2003) monumental work *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*, Her and Li (2023: 133) argue that PTB is without classifiers. We thus presume that, given the relative isolation of KC languages in the TB branch of the Sino-Tibetan family, PKC remained without classifiers, which is a borrowed feature in some of the KC languages after PKC. In summary, we put forth the following four hypotheses, two for PKC and two for PTB.

(9) Two hypotheses for PKC
a. PKC has base-initial, not base-final, numerals.
b. PKC does not have numeral classifiers.

(10) Two hypotheses for PTB
   a. PTB has base-initial, not base-final, numerals.
   b. PTB does not have numeral classifiers.

We use phylogenetic comparative methods to test the four hypotheses above. Such methods are considered appropriate as they consider the non-independence of features from evolutionary processes (Mace and Holden 2005, Macklin-Cordes and Round 2022). We extract the tree from the Kuki-Chin and Tibeto-Burman languages included in our data set from a world tree of languages proposed by Bouckaert et al. (2022). This results in a sample of 902 trees that show possible histories of the languages in the data set. This sample can be summarized with a maximum clade credibility tree, in which every tree of the sample is scored by the product of the likelihood of the splits observed in each individual tree. The tree with the highest score is taken to represent the sample. The maximum clade credibility tree extracted from the tree sample is shown in Fig. 10, along with the variables included in our data.

Fig. 10. The maximum clade credibility tree of Kuki-Chin and Tibeto-Burman languages based on the world tree from Bouckaert et al. (2022). The heatmap visualization displays the variables included in our data set. The clade in green represents Kuki-Chin languages.

Based on the sample of trees, we infer the probability of classifiers and base order at the root of Kuki-Chin and Tibeto-Burman languages. We use a reverse jump hyperprior (RJHP) (Green 1995; Gowri-Shankar and Rattray 2007). The RJHP is based on a Continuous Time Markov Chain process. It considers scenarios of reversed change between different states of a variable. For example, it scores the probability that classifiers are lost, acquired, re-acquired and re-lost. The analyses are conducted in R (R-Core-Team, 2021) with the following packages: ape (Paradis and Schliep 2019), diagram (Soetaert 2020), Ggally (Schloerke et al. 2021), ggpubr (Kassambara 2020), phangorn (Schliep 2011), phytools (Revell 2012), tidyverse (Wickham et al. 2019). The code and the data used for the analysis are available at the following repository: [LINK to be provided]. Within Kuki-Chin languages, the results show that the base-initial is the most likely state at the root of Kuki-Chin languages (the median probability at the root is 0.95 for RJHP). Furthermore, classifiers are likely to be present at the root of Kuki-Chin languages (the median probability at the root is 0.86 for RJHP). When considering Tibeto-Burman languages, base-final is the most likely state at the root (the median probability at the root is 0.97 for RJHP). However, CL-initial is the most likely state at the root of Tibeto-Burman languages (the median probability at the root is 0.59 for RJHP).

## 5. Discussions and Conclusions

We have demonstrated in this paper that numeral bases function as multiplicands in the mathematical formation of a multiplicative numeral and that numeral classifier likewise functions as a multiplicand, just like the numeral base. As a result, the typology of the classifier word order is thus closely related to the typology of base word order. The results of a comprehensive survey of base orders and classifier word orders in the Kuki-Chin languages within the Tibeto-Buran group of the Sino-Tibetan family are summarized in Table 6.

Table 6 Distribution of the base feature in 54 KC languages

| Base-initial 41 | | | Base-split 4 | | | No-data 9 | | | Total 54 |
|---|---|---|---|---|---|---|---|---|---|
| Cl-initial 31 | No-Cl 4 | No-data 6 | Cl-initial 1 1 | No-Cl 2 | No-data 1 | Cl-initial 0 | No-Cl 1 | No-data 8 | 54 |

We have also verified that all KC languages conform to Greenberg's Universal 20A, in that nouns do not intervene between numerals and classifiers. Furthermore, all KC languages with numeral classifiers have the base and the classifier in harmonious order, i.e., base-initial and Cl-initial, as seen in Table 2. Based on the data and the existing views in the literature, we propose four tentative hypotheses, two for PKC in (11), repeated from (9), and two for PTB in (12), repeated from (10), respectively.

(11) Two hypotheses for PKC
PKC has base-initial, not base-final, numerals.
PKC does not have numeral classifiers.

(12) Two hypotheses for PTB
PTB has base-initial, not base-final, numerals.
PTB does not have numeral classifiers.

The results of phylogenetic comparative analyses show the following:

1. **PKC has base-initial, not base-final, numerals.**
The preliminary results match our hypothesis that base-initial numerals are the most likely state at the root of Kuki-Chin languages.

2. **PKC has numeral classifiers.**
The preliminary results do not match our hypothesis, as classifiers were more likely at the root of Kuki-Chin. We thus may need to revise the hypothesis and acknowledge the existence of classifiers in PKC already. Further research on KC's sister groups and the root of their mother node, the Northeast Indian Areal Group will shed light on this issue.

3. **PTB numerals are more likely base-final.**
The preliminary results do not match our hypothesis. We suspect this is due to the large number of Tibeto-Burman languages switching to base-final order due to contact with neighboring base-final languages. Additional qualitative analyses and more comprehensive quantitative data on TB languages, especially those of underrepresented languages, will be needed for a clearer picture.

4. **PTB may have numeral classifiers.**
The preliminary results do not match our hypothesis, as CL-initial is the most likely state at the root of Tibeto-Burman languages. Again, we suspect this is due to the effect of language contact between Tibeto-Burman languages and neighboring classifier languages. Additional qualitative analyses and more comprehensive quantitative data on TB

languages, especially those of underrepresented languages, will be needed for a clearer picture.

For further research, we plan to conduct a comprehensive survey of all TB languages and find additional qualitative and quantitative evidence for the state of PTB in terms of numeral bases and classifiers.

## Notes

1. Recent studies by Zhang et al. (2019) and Zhang et al. (2020) have provided robust evidence supporting a dichotomy between Sinitic and TB languages.

## Funding

This work was supported by the Indian Council of Social Science Research (File No: ICSSR-NSTC(Taiwan)/JRP-3/2023-IC) and Taiwan's National Science and Technology Council (112-2923-H-029-001-MY2, 111-2410-H-029-009-MY3, 111-2811-H-029-002, 110-2811-H-029-507, 111-2811-H-029-003).

## References

Allassonnière-Tang, M. and Her, O. (2020) 'Numeral base, numeral classifier, and noun: Word order harmonization', *Language and Linguistics*, 2020, 21(4): 511–56.

Bouckaert, R., et al. (2022) 'Global language diversification is linked to socio-ecology and threat status'. <https://doi.org/10.31235/osf.io/f8tr6> accessed 20 June 2024.

Chan, E. (2024) *Numeral Systems of the World's Languages* <https://lingweb.eva.mpg.de/channumerals/TypNum_Latest_21ho.pdf> accessed 6 June 2024.

Comrie, B. (2013) 'Numeral Bases'. In Dryer M S & Haspelmath M (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/131> accessed 6 August 2021.

Gowri-Shankar, V. and Rattray, M. (2007) 'A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model', *Molecular Biology and Evolution*, 24(6): 1286–99.

Green, P. J. (1995) 'Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika*, 82: 711–32.

Greenberg, J. H. (1990a[1972]) Numerical classifiers and substantival number: Problems in the genesis of a linguistic type. In: K. Denning & S. Kemmer (Eds.), *On language. Selected writings of Joseph H. Greenberg* (pp. 166-193). Stanford, CA: Stanford University Press. [First published 1972 in *Working Papers on Language Universals* 9: 1–39.]

Greenberg, J. H. (1990b[1978]) Generalizations about numeral systems. In Denning, K., & Kemmer, S. (Eds.), *On language. Selected writings of Joseph H. Greenberg* (pp. 271-309). Stanford, CA: Stanford University Press. [First published 1978 in Greenberg, J H, Ferguson, C A, and Moravcsik, E A (Eds.), *Universals of the Human Language, Vol. 3 Word structure*, pp. 249–95. Stanford, CA: Stanford University Press.]

Hammarström, H. et al. (2024) *Glottolog 5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://doi.org/10.5281/zenodo.10804357> accessed 20 June 2024.

He, C. et al. (2017) ‘Overt coordination in additive numerals of minority languages in South China’. *Syntax*, 20/3: 292-316.
He, C and Her, O.-S. (2022) ‘Complex numerals in English: Constituents or not?’ *English Language and Linguistics*, 26/4: 723-49. doi:10.1017/S1360674321000447
Her, O.-S. (2017a) ‘Structure of numerals and numeral classifiers in Chinese: Historical and typological perspectives and cross-linguistic implications’. *Language and Linguistics*, 18/1: 26-71.
Her, O.-S. (2017b) ‘Deriving classifier word order typology, or Greenberg’s Universal 20A and Universal 20’. *Linguistics*, 55/2: 265-303.
Her, O.-S., Hammarström H., and Allassonnière-Tang, M. (2022) ‘Defining numeral classifiers and identifying classifier languages of the world’. *Linguistics Vanguard*, 8.1: 151-64. <https://wacl.clld.org/> accessed 20 June 2024.
Her, O.-S. and Li, B.-T. (2023) ‘A Single origin of numeral classifiers in Asia and Pacific: A hypothesis.’ In: Allassonnière-Tang M and Kilarski M (eds.), *Nominal Classification in Asia: Functional and Diachronic Perspectives*, 113-60. Amsterdam: John Benjamins.
Her, O.-S. and Tang, M. (2020) ‘A statistical explanation of the distribution of sortal classifiers in languages of the world via computational classifiers’. *Journal of Quantitative Linguistics*, 27/2: 93-113.
Her, O.-S., Tang, M., and Li B.-T. (2019) ‘Word order of numeral classifiers and numeral bases: Harmonization by multiplication’. *STUF-Language Typology and Universals*, 3: 1-32.
Her, O.-S. et al. (2024) ‘Early humans out of Africa had only base-initial numerals’. *Humanities and Social Sciences Communications*, 11/254: 1-7. <https://doi.org/10.1057/s41599-023-02506-z> accessed 20 June 2024.
Ionin, T. and Matushansky, O. (2006) ‘The composition of complex cardinals’. *Journal of Semantics*, 23/4: 315–360.
Ionin, T. and Matushansky, O. (2018) *Cardinals: The Syntax and Semantics of Cardinal-containing Expressions*. Cambridge, MA: MIT Press.
Kassambara, A. (2020) Ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. <https://CRAN.R-project.org/package=ggpubr>
Kiryu, K. (2009) ‘On the rise of the classifier system in Newar’. In: Nagano Y (ed.) *Issues in Tibeto-Burman Historical Linguistics* (*Senri Ethnological Studies 75*), pp. 51-69. Senri Ethnological Studies (75). Osaka, Japan: National Museum of Ethnology.
Li, B.-T., Yang, W.-C., and Her, O.-S. (2018) ‘On the variation of numeral systems and classifier systems under ethnic contact: A case study of Sunwar and Rabha in Tibeto-Burman’. *Ethnologia*, 42: 39-80.
Mace, R. and Holden, C. J. (2005) ‘A phylogenetic approach to cultural evolution’. *Trends in Ecology & Evolution*, 20/3: 116–121.
Macklin-Cordes, J. L., and Round, E. R. (2022) ‘Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast’. *Linguistic Typology*, 26(3): 533–72. <https://doi.org/10.1515/lingty-2021-0025> accessed 20 June 2024.
Matisoff, James A. (2003) *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. Berkeley, CA: University of California Press.
Matisoff, James A. (2015) The Sino-Tibetan Etymological Dictionary and Thesaurus, Berkeley: STEDT. <https://stedt.berkeley.edu/dissemination/STEDT.pdf> accessed 20 June 2024.

534 Paradis, E. amd Schliep, K. P. (2019) ‘Ape 5.0: an environment for modern
535 phylogenetics and evolutionary analyses in R’. *Bioinformatics* 35, 526–28.
536 R Core Team. (2021) *R: A language and environment for statistical computing*. Vienna:
537 R Foundation for Statistical Computing. <https://www.R-project.org/>
538 Revell, Liam J. (2012) Phytools: An R package for phylogenetic comparative biology
539 (and other things). *Methods in Ecology and Evolution*, 3. 217–23. doi:
540 10.1111/j.2041-210X.2011.00169.x.
541 Schliep, K. P. (2011) ‘Phangorn: phylogenetic analysis in R’. *Bioinformatics*, 27/4:
542 592–593.
543 Schloerke, B., et al. (2021) Ggally: Extension to 'ggplot2'. R package version 2.1.2.
544 <https://CRAN.R-project.org/package=GGally>
545 Senft, G. (1986) *Kilivila: The Language of the Trobriand Islanders*. Berlin: de Gruyter
546 Mouton.
547 Soetaert, K. (2020) *Diagram: Functions for visualising simple graphs (networks),*
548 *plotting flow diagrams. R package version 1.6.5*. <https://CRAN.R-
549 project.org/package=diagram>
550 VanBik, K. (2009) *Proto-Kuki-Chin: A reconstructed ancestor of the Kuki-Chin*
551 *languages*. Berkeley: Sino-Tibetan Etymological Dictionary and Thesaurus
552 Project, University of California.
553 <https://stedt.berkeley.edu/pubs_and_prods/STEDT_Monograph8_Proto-Kuki-
554 Chin.pdf> accessed 20 June 2024.
555 Wągiel, M., and Caha, P. (2020) ‘Universal semantic features and the typology of
556 cardinal numerals’. *Catalan Journal of Linguistics*, 19: 199-229.
557 Wickham, H. et al. (2019) Welcome to the tidyverse. *Journal of Open Source*
558 *Software*, 4/43: 1686. <https://doi.org/10.21105/joss.01686> 20 June 2024.
559 Zhang, M. (2019) ‘Phylogenetic evidence for Sino-Tibetan origin in northern China in
560 the Late Neolithic’. *Nature* 569, 112–5. <https://doi.org/10.1038/s41586-019-
561 1153-z> accessed 20 June 2024.
562 Zhang, H. et al. (2020) ‘Dated phylogeny suggests early Neolithic origin of Sino-
563 Tibetan languages’. *Scientific Reports* 10, 20792.