Word Order of Numeral Classifiers and Numeral Bases: Harmonization by Multiplication

Abstract

In a numeral classifier language, a sortal classifier (C) or a mensural classifier (M) is needed when a noun is quantified by a numeral (Num). Greenberg (1990b, p. 185) first observes that cross-linguistically Num and C/M are adjacent, either in a [Num C/M] order or [C/M Num]. Likewise, in a complex numeral with a multiplicative composition, the *base* may follow the multiplier as in [*n* × *base*], e.g., *san-bai* 'three hundred' in Standard Mandarin. However, the base may precede the multiplier, thus [*base* × *n*], which is also attested. Greenberg (1990a, p. 293) further observes that [*n* × *base*] numerals appear with a [Num C/M] alignment and [*base* × *n*] numerals with [C/M Num]; base and C/M thus seem to harmonize in word order. This paper first motivates the base-C/M harmonization via the multiplicative theory of classifiers (Her, 2012a, 2017a), and verifies it empirically within six language groups in the world's foremost hotbed of classifier languages: Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai, Tibeto-Burman, and Indo-Aryan. Our survey further reveals two interesting facts: base-initial ([*base* × *n*]) and C/M-initial ([C/M Num]) orders exist only in Tibeto-Burman (TB) within our dataset and so are the few scarce violations to the base-C/M harmonization. We offer an explanation based on Proto-TB's base-initial numerals and language contact with neighboring base-final, C/M-final languages.

Keywords: classifier, multiplication, numeral base, harmonization

1. Introduction

In a numeral classifier language, a sortal classifier (C) or a mensural classifier (M) is required when a noun (N) is quantified by a numeral (Num) (e.g., Aikhenvald, 2000; Allan, 1977; Tai & Wang, 1990). To illustrate, Standard Mandarin Chinese is attested as a canonical classifier language (e.g., Zhang, 2013, pp. 1-2); thus, the enumeration of a countable noun demands the presence of a sortal classifier, while mass nouns rely on mensural classifiers,[1] as shown in (1a) and (1b), respectively.

(1) Classifiers in Standard Mandarin Chinese

a.

三	本	書
san	*ben*	*shu*
three	C_{volume}	book

'three books'

b.

三	箱	書
san	*tong*	*shu*
three	M_{bucket}	water

'three buckets of water'

With regard to their geographical distribution, classifier languages are commonly found in the eastern and south-eastern parts of Asia, from which they spread westward and eastward as far as to the western coasts of the Americas (Gil, 2013). A weighted sample of classifier languages is displayed in Figure 1 via data from the *World Atlas of Language Structures*, with each red dot indicating a classifier language. Though the languages shown do not represent an exhaustive list of classifier languages, the map does offer a general picture of the spatial distribution of classifier languages, which are mostly found in Asia.

[1] Further distinction is made between *mensural classifiers* in numeral classifier languages and *terms of measure* in non-classifier languages such as English (Her, 2012a, p. 1682). For example, *bucket* in *three buckets of water* is not a mensural classifier in the sense of *tong* 'bucket' in (1b) because its syntactic structure is different in that it bears a plural marker and requires the preposition *of*. Hence, it behaves more like a nouns than a mensural classifier (Kilarski, 2014, p. 9).

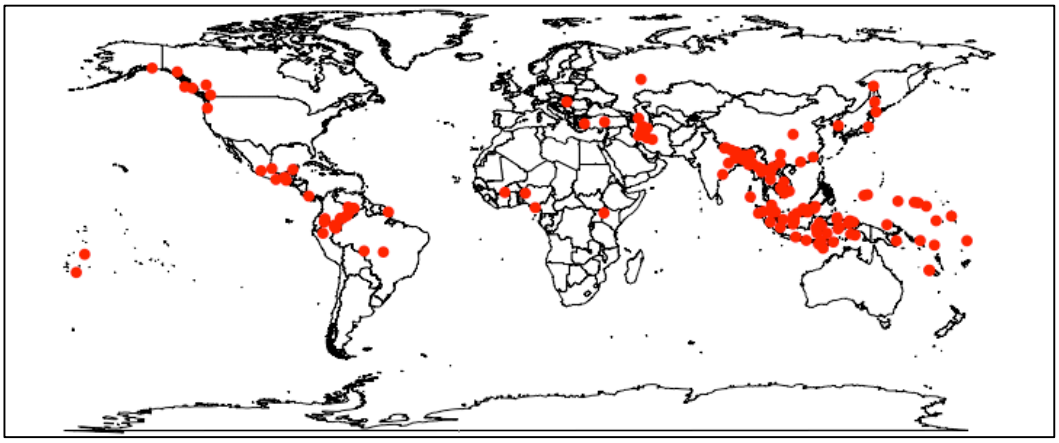

Figure 1. An overview of classifier languages in the world (Gil, 2013)

In terms of word orders formed by the three elements, Num, C/M and N, mathematically there are six possibilities, i.e., 3! = 3×2×1 = 6, as in (2).[2] Mandarin Chinese, for example, throughout its 3,000 years of recorded history, is consistently C/M-final, i.e., Num precedes C/M, as seen in (1), as opposed to languages that are C/M-initial, i.e., Num follows C/M. Crucially, however, it has been observed that cross-linguistically N does not intervene between Num and C/M (Aikhenvald, 2000, pp. 104-105; Greenberg, 1990b, p. 185; Peyraube, 1998; Wu, Feng, & Huang, 2006).

(2) Six possible word orders of Num, C/M, and N
  a. √ [Num C/M N] Many languages, e.g., Mandarin (Sinitic)
  b. √ [N Num C/M] Many languages, e.g., Thai (Tai-Kadai)
  c. √ [C/M Num N] Few languages, e.g., Ibibio (Niger-Congo)
  d. √ [N C/M Num] Few languages, e.g., Jingpho (Tibeto-Burman)
  e. * [C/M N Num] No languages attested
  f. * [Num N C/M] No languages attested

From this four-way typology, two more revealing generalizations, as shown in (3) (Greenberg, 1990a, p. 292), have been derived and supported by theoretical and typological evidence, and are dubbed 'Greenberg's Universal 20A' (Her, 2017a, p. 265).

(3) Greenbergs's Universal 20A (Her, 2017a, p. 298)
- Part 1: Of the three elements Num, C/M, and N, any order is possible as long as Num and C/M are adjacent.[3]
- Part 2: There are many more languages with the C/M-final orders than languages with C/M-initial orders.

The next question is of course why. Why do Num and C/M reject the intervention by N? And why do languages prefer the C/M-final over the C/M-initial order? This paper addresses these questions with a unified mathematical explanation. The underlying mathematics of classifiers further predicts two probabilistic universals concerning the word orders of classifiers and numeral bases. A systematic survey is made in six language groups in the

[2] While sortal classifiers (C) and mensural classifiers (M) are used to designate the two subcategories of numeral classifiers, the overall category is thus referred to as C/M.

[3] Some languages are alleged to have the [C/M N Num] order in (2e) when Num is one. However, this 'one' has been shown to be an indefinite marker and thus not a numeral (Her, 2017a).

world's foremost hotbed of numeral classifier languages, i.e., Sinitic, Miao-Yao (aka Hmong-Mien), Austro-Asiatic, Tai-Kadai, Tibeto-Burman, and Indo-Aryan, or SMATTI as an acronym, to examine to what extent these probabilistic universals are born out empirically. Results are displayed in spatial representation along with statistical analysis.

This paper is organized as follows. Section 2 introduces the multiplicative theory of classifiers (Her, 2012a, 2012b), from which two probabilistic universals are derived. Section 3 explains our methodology and Section 4 presents respectively the validity of the two probabilistic universals in our survey of SMATTI. Section 5 returns to the generalizations in (3) and examines the violations of the proposed probabilistic universals in our dataset, while Section 6 concludes this paper.

2. Multiplication theory of C/M and numeral base

Greenberg (1990b, p. 172) first considers the operation between Num and C as multiplication, thus [Num C] = [Num × *1*]. Following this lineage of thought, several studies (e.g., Au Yeung, 2005, 2007, Her, 2012a, 2012b, Yi, 2009, 2011) propose that the difference between C and M is therefore that the value of a C is necessarily *1*, while the value of an M is not necessarily *1*. C/M's different types of mathematical values are summarized in Table 1, with examples from Mandarin Chinese (Her, Chen, & Yen, 2017).

Table 1. Types of C/M Based on Mathematical Values

Numerical or Not	Fixed or Not	Examples		C/M Type
Numerical	Fixed	***1***	個 *ge* $C_{general}$, 隻 *zhi* C_{animal}, 條 *tiao* C_{long}	**C**
		¬1	*2* 雙 *shuang* 'pair', 對 *dui* 'pair'; *12* 打 *da* 'dozen'	**M_1**
	Variable	***>1***	排 *pai* 'row, 群 *qun* 'group, 幫 *bang* 'gang'	**M_2**
Non-numerical	Fixed	***¬n***	斤 *jin* 'catty', 升 *sheng* 'liter', 碼 *ma* 'yard'	**M_3**
	Variable	***¬n***	滴 *di* 'drop', 節 *jie* 'section', 杯 *bei* 'cup'	**M_4**

Cs carry the necessarily fixed numerical value of *1*, as in *san zhi gou* (three C_{animal} dog) 'three dogs', where the quantity of the referents is precisely *3* × *1*, with *zhi* also serving to highlight the animacy of the following noun. Ms, on the other hand, can have various kinds of values, numerical or non-numerical, fixed or variable. For instance, an M may have a fixed numerical value: *san da bi* (three M_{dozen} pen) 'three dozens of pens'. The quantity of the pens is precisely *3* × *12*. Variable numerical value is also possible with an M, as in *san pai shu* (three M_{row} tree) 'three rows of trees'. One row may contain a variable number of trees, making the total number unspecified. An M could also have a fixed non-numerical value, as in *san gongjin shui* (three M_{kilo} water) 'three kilos of water', or a variable non-numerical value, as in *san bei shui* (three M_{cup} water) 'three cups of water'. Thus, while C and M both have a multiplicative relation with the preceding numeral, Cs bear a necessarily numerical value of *1*, while Ms apply all sorts of other values.

While this concept of multiplication in C/M is opaque, it is rather transparent within the numeral systems of most languages in the world. Comrie (2013) conducts an extensive survey of the numeral system in 196 languages, of which 172 (87.75%) employ both addition and multiplication. Comrie (2006) offers a concise formulation for the internal composition of such multiplicative numerals, as in (4).

(4) $(x \times base) + y$, where $y < base$

Taking the Chinese numeral system for example, 三百二十一 *san-bai er-shi yi* (three-hundred two-ten one) '321' has the internal relation of [(*3*×*100*)+(*2*×*10*)+*1*]. In this

numeral system, exponentials of 10 (i.e., *shi* '10', *bai* '100', *qian* '1000', among others) are numeral bases and function as multiplicands to the respective preceding number (*n*). The order of *n* and base is irrelevant and the two possible orders between *n* and base, i.e., base-final [*n* base] and base-initial [base *n*], are both attested in the world's languages. Chinese numerals, once more, have been consistently base-final throughout its 3,000 years of recorded history. See the example in (5). For a more extensive set of examples, see (Her, 2017b; Peyraube, 1998).

(5) Word order of numerals in Archaic Chinese

獲	鳥	二百	十	二
huo	*niao*	*er-bai*	*shi*	*er*
capture	bird	two-hundred	ten	two

'captured 212 birds'

It thus seems that Chinese has always been base-final as well as C/M-final, as shown in (5) and (1). Such harmonization in word order between numeral bases and C/M, as shown in (6), is not only found in Chinese but also to a large extent in classifier languages of the world. This was first noted by Greenberg (1990a, p. 293) and recently further developed by Her (2017a, 2017b).

(6) Harmonization between base and C/M (Her, 2017a, p. 280)
a. C/M-final order ⇔ base-final numerals
b. C/M-initial order ⇔ base-initial numerals

The motivation behind this probabilistic universal is the unification of numeral bases and classifiers under the concept of multiplicand (Au Yeung, 2005, 2007; Her, 2012a, 2017a). In essence, elements that function as multiplicands should naturally follow the same word order in a language (Her, 2012a, p. 279), (6) thus obtains. The underlying force that keeps Num and C/M from being interrupted by N is likewise the multiplicative function that requires Num as the multiplier and C/M as the multiplicand. Furthermore, the multiplicative theory also predicts that a language with C/M must also have multiplicative bases in its numerals. This can be stated as a probabilistic universal as well, as in (7).

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

It is important to highlight the directional differences in (6) and (7). Even though the two universals are both probabilistic implicational universals, the harmonization observed in (6) is bidirectional, i.e., the existence of C/M-final order implies that numerals should also be base-final and vice-versa. However, the co-occurrence of numeral bases and classifiers in a language is expected to be unidirectional. In other words, the presence of classifiers implies the existence of multiplicative numerals. Nevertheless, the existence of multiplicative numerals does not imply the presence of classifiers. The motivation for such statement is purely empirical, since there are numerous languages with multiplicative numerals but not classifiers, such as English. The entailment in (7) equivalently provides a possible explanation for the fact that more C/M-final languages are attested. The reason why C/M-final languages outnumber C/M-initial languages could be due to the fact that base-final classifier languages outnumber base-initial classifier languages, i.e., the presence of multiplicative numerals is the precondition for a language to have classifiers but the word order of C/M must be harmonized according to (6).

To summarize, two potential probabilistic universals are proposed based on the multiplication relation between numerals and C/M. First, the presence of C/M in a language implies the ability of multiplication of its speakers, and thus the presence of multiplicative numerals as described in (7). Second, by reason of their common underlying function of multiplicand, the word order of base and C/M are expected to be harmonized, both being final or initial as in (6). From such hypotheses, the prevalence of C/M-final order, as those in (2a) and (2b), should result from the prevalence of base-final order. What actually gave rise to the base-final numeral system is nevertheless beyond the scope of this paper. We speculate that it is related to the ordering of syntactic head, but we leave this assumption for future studies.

3. Methodology

To assess the veracity of the two proposed probabilistic universals, we perform a systematic survey of classifier languages in six language groups in the hotbed of the world's classifier languages: Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai, Tibeto-Burman, and Indo-Aryan, dubbed SMATTI. In a database of 491 classifier languages from the Syntax and Lexicon Lab at National Chengchi University, SMATTI accounts for 45.41% (223/491) of the data points, making these six language groups a suitable target for our preliminary study. We are aware that such choice may limit the geographical and phylogenetic diversity of our samples; yet, in order to assure a certain level of quality, we narrow the scope in a way that a sufficient amount of data is available to test our hypothesis, while each data point can be cross-checked. To avoid confounding probabilistic universals and areal features (Sinnemaki, to appear), we do plan to include languages in other regions of the world in future analyses.

Figure 2 displays all 969 languages in SMATTI according to *Ethnologue* (Lewis, Simons, & Fennig, 2009). Each point represents one language, and each language is represented once in the map. For those languages which have multiple habitats and are recorded with multiple coordinates in *Ethnologue*, the most iconic coordinates are chosen based on number of speakers or the place of origin. The six language groups investigated in our study are distinguished by six different colors.

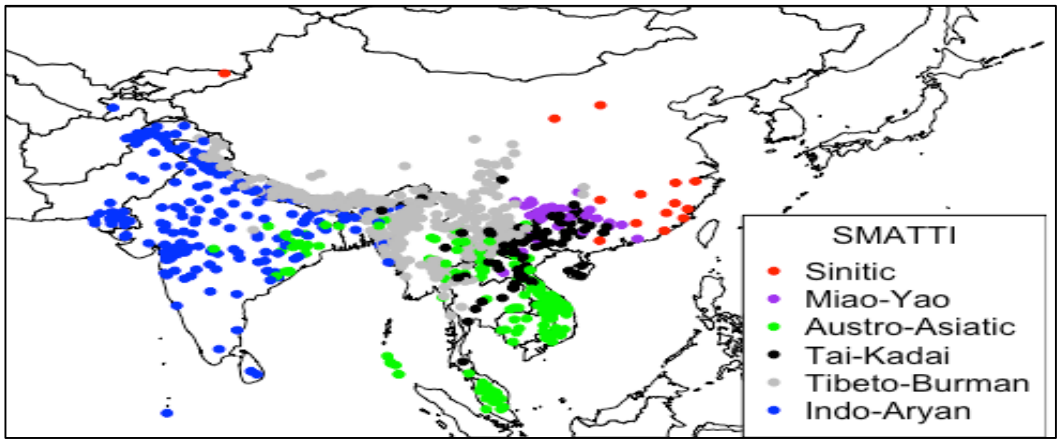

Figure 2. A spatial overview of all languages in SMATTI

Information on classifiers and numeral systems of these languages was collected from the existing literature. Taking Khasi (Austro-Asiatic) for example, the numeral system was provided by Chan (2017), as shown in Table 2. This language is then analyzed and annotated as base-final, since the morpheme of 'ten' [p^{h}u] follows the multipliers from 20 to 90, e.g., in 'thirty' [la:j p^{h}u], the multiplicand (i.e., the numeral base) [p^{h}u] 'ten' follows the multiplier [la:j] 'three'. The same pattern is observed in higher numbers such as hundreds and thousands, e.g., [ʔa:r sp^{h}aʔ] 'two-hundred' is the combination of [ʔa:r] 'two' and [sp^{h}aʔ] 'hundred'.

Table 2. Numeral system of Khasi (Chan, 2017)

1. wej // ʃi	10. ʃi pʰeːu	100. ʃi spʰaʔ
2. ʔaːr	20. ʔaːr pʰu	200. ʔaːr spʰaʔ
3. laːj	30. laːj pʰu	1000. ʃi had͡ʒaːr
4. saːo	40. saːo pʰu	2000. ʔaːr had͡ʒaːr
5. san	50. san pʰu	
6. hnriːu	60. hnriːu pʰu	
7. hnŋeu	70. hnŋeu pʰu	
8. pʰra	80. pʰra pʰu	
9. kʰndaj	90. kʰndaj pʰu	

As for C/M, the literature is rather generous with regard to naming. For example, sortal classifiers as we define them in this paper, may be referred to as individual classifier, numeral classifier, word of measure, quantifier, unit word, numerative, among others. Hence, we apply the methodology of Her (2012a, 2012b) on language data provided by other researchers to maintain a unified and consistent analysis. For instance, one of our references to the Indo-Aryan language Bengali, Bhattacharya (2001), mentions the existence of the following classifiers: the human classifier *jon*, the inanimate count classifier *khana*, the collective classifier *gulo*, and the numeral absorbing human collective classifier *ra*, among others. By reviewing the examples provided and cross-checking different sources, most classifiers do fit the definition of our study, as in *tin-jon chele* (three-C_{human} boy) 'three boys' (Biswas, 2013). However, the collective classifiers are not included in our analysis, since they are viewed as plural markers (Biswas, 2013, p. 2; Dayal, 2014, p. 49).

Furthermore, following the definition of Gil (2013), languages with few and/or optional classifiers are viewed as classifier languages too. For example, Marathi (Indo-Aryan) is attested to have only two numeral classifiers, *jan* and *jani*, for counting masculine and feminine people with numerals higher than four, while these two classifiers are optional from two to four (Aikhenvald, 2000, p. 287; Emeneau, 1956, p. 11). Nevertheless, it is still counted as a classifier language in our database. As a result, we were able to identify both the numeral and classifier systems of 219 classifier languages among the entirety of SMATTI (23.11%). The detailed numbers are displayed in Table 3. The full list of languages is in the Appendix.

Table 3. Overview of language survey

	Languages	Classifier languages
Sinitic	14	14 (100%)
Miao-Yao	38	8 (21.05%)
Austro-Asiatic	169	39 (23.08%)
Tai-Kadai	92	40 (43.48%)
Tibeto-Burman	435	100 (22.99%)
Indo-Aryan	221	18 (8.14%)
Total	969	219 (22.60%)

The observed tendencies are in accordance with the literature. Sinitic languages are expected to be prototypical classifier languages (Bisang, 1999; Zhang, 2013). Thus, every language of the group is expected to be a classifier language. The high ratio of classifier languages within the Tai-Kadai group is also not surprising as most Tai-Kadai languages are expected to employ numeral classifiers (Morev, 2000). With regard to Miao-Yao, the amount of numeral classifier languages is rather low, considering the fact that the literature often

refers to Miao-Yao as a classifier language group. The reason for such divergence is that some of the languages in Miao-Yao actually possess noun classifiers instead of numeral classifiers (Mortensen, 2017, p. 15). Thus, we did not include them in our dataset (the same logic applies for other language groups). As for the Austro-Asiatic group, Bauer (1992, p. 374) states that "numeral classifier systems found in Austroasiatic languages are not an inherited feature, but represent a secondary, or borrowed, system". Moreover, the structure of numeral classifiers in Austro-Asiatic also represent a high level of diversity probably due to different language contact situations (Adams, 1986, pp. 256-257). Such phenomenon results in the fact that some languages were not validated by our formal criteria of classifiers. Classifiers are not a common feature in Tibeto-Burman. They are largely attested in languages in contact with Austro-Asiatic sub-groups and certain other branches of Tibeto-Burman such as Qiang, and Burmish (Fu, 2015, pp. 45-46). Finally, Indo-Aryan languages show a very small ratio of classifier languages, which is also expected since Indo-European languages generally rely on other systems of nominal classification such as grammatical gender (Luraghi, 2011).

We are aware that such general picture is still subject to challenge, as some studies attest that Miao-Yao, Austro-Asiatic, Tai-Kadai, and Tibeto-Burman widely use classifier devices (Xu, 2013, pp. 54-55). Nonetheless, no database known to the authors actually provides detailed references and examples of such statements, i.e., most examples are extracted from languages with a large population of speakers, while much less detailed data is provided for languages with restricted number of speakers. We only included in our dataset classifier languages which are supported by actual examples and theoretical verification. Hence, this may affect the general distribution. We estimate that the general criteria of diversity are matched for the purpose of this paper as every language group is represented in terms of ratio. The same observation is made with regard to spatial distribution in Figure 3. Red points indicate the 219 surveyed languages while gray points symbolize the languages of SMATTI which are not included in our dataset.

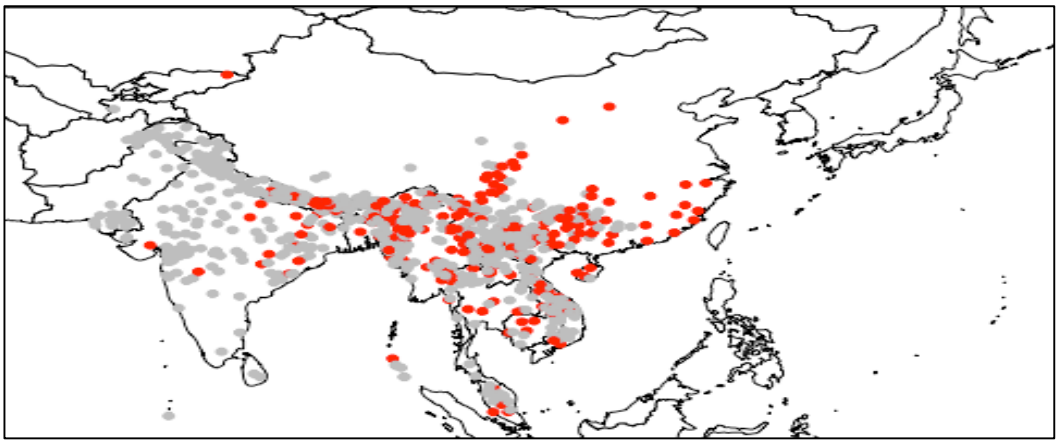

Figure 3. Spatial overview of the 219 surveyed languages

Furthermore, our current aim is to verify the harmonization of multiplicative numerals and numeral classifiers. Thus, we only require classifier languages in our dataset. We do not attempt to provide a full phylogenetically diverse set of samples for an empirical reason, i.e., classifiers may be a feature of certain sub-branches of a language group rather than an over-the-board property across an entire language family. To illustrate, only a few branches of Tibeto-Burman display the use of classifiers; therefore, it is only natural that our dataset only includes these specific branches. Inclusion of these other languages may reveal more than the current study, but we leave it here for future studies.

Another important disclaimer concerns the definition of the proposed probabilistic universals. As reflected in the term itself, a probabilistic universal refers to an observation which "holds for most, but not all, languages", as opposed to absolute universals, where no exceptions are allowed (Dryer, 1998; Velupillai, 2012, p. 31). We do not claim that the two universals under proposal are absolute universals for two reasons. First, it has been shown that cross-linguistic analyses are rarely statistically justified to be absolute and without exceptions (Piantadosi & Gibson, 2014, p. 736). Both Bayesian and frequentist statistical methods would require an unachievable amount of data to support the existence of absolute universals, and our dataset does not contain an exhaustive list of languages of the world. Second, even if we do have data on all current languages, theoretically an observation in the data does not justify that it applies to all languages, as there is no way of knowing whether there are languages contradicting the universal, either the ones that were no longer spoken or hypothetically possible human languages that have not emerged due to historical accident (Dryer, 1998).

However, under the proposal of probabilistic universals, it is quite possible to falsify the null hypothesis. In our case, the alternative hypotheses are manifested via the two proposed probabilistic universals, while the null hypotheses are 1) there is no harmonization between the base-parameter and the C/M-parameter, 2) the presence of classifiers is not a reliable factor to predict the existence of multiplicative numerals in a language. Thus, our study may not prove the alternative hypotheses theoretically, but our cross-linguistic analysis can possibly reject the null hypotheses quite convincingly within the observed dataset. Such probabilistic approach is hence "explored in the same theory-hypothesis-statistics triangle that characterizes most sciences" (Bickel, 2014, p. 119).

4. Results

In this section, we scrutinize our data with regard to the two probabilistic universals. A two-tailed exact binomial test is applied for the probabilistic universal related to the co-occurrence of classifiers and multiplicative numerals. With regard to the harmonization between the base-parameter and the C/M-parameter, we first display the overall distribution of tokens via bar plots. We then calculate the odd ratio of the variables to obtain a preliminary statement. Third, we measure the probability of the alternative hypothesis in terms of statistical significance via the Chi-square test of independence, which is further supported by the Fisher's exact test. Finally, we apply the φ coefficient to generate the effect size of the alternative hypothesis.

4.1 Co-occurrence of numeral bases and classifiers

All 219 SMATTI classifier languages have been confirmed to employ multiplicative numerals, as shown in Table 4.

Table 4. Numeral systems and numeral classifiers in SMATTI

	With classifiers
With multiplication	219
Without multiplication	**0**

The data required to testify the probabilistic universal only involves a binomial variable, i.e., with/without multiplication. Thus, we apply the two-tailed exact binomial test, which assesses whether the proportion of success on the nominal variable significantly differs from a hypothesized value. Generally, this hypothesized value is determined by chance, e.g., the probability of tossing a coin 10 times and getting tail is 10/2=5 times. Nevertheless, the presence of multiplicative numerals in languages of the world does not follow such pattern. As mentioned in Section 2, the survey of Comrie (2013) attests that 87.75% (172/196) of the

surveyed languages employ multiplication. Hence, we formulate the null hypothesis as follow: the number of observed languages with multiplicative numerals is expected to represent 87.75% of the dataset. On the other hand, the alternative hypothesis suggests that the observed data are different from the hypothesized distribution. By applying such criterion, we can demonstrate whether the 100% ratio of languages with multiplicative numerals within classifier languages is statistically significant or not. The detailed equation of an exact binomial test is shown in (8). While n represents the total quantity of tokens and k indicates the number of expected observations, p incarnates the probability of success.

(8) Formula of the Exact Binomial Test

$$\mathrm{P}(\mathrm{X} = \mathrm{k}) = C_k^n p^k (1-p)^{n-k}$$

Thus, an exact binomial two-tailed test with 95% confidence interval is performed to assess the probability of the null hypothesis that the co-occurrence of classifiers and multiplicative numerals is not correlated. The result is at the level of high significance, $p < 0.001$, thus allowing us to reject the null hypothesis of no association. The proportion of languages with multiplicative numerals significantly exceeds the hypothesized value of 87.75% and supports our first probabilistic universal.

We are aware that a more extensive survey of languages of the world is required to further support such a probabilistic universal, as the association of multiplicative numerals and classifiers may be due to coincidence, i.e., most languages of the world have multiplicative numerals and, due to phylogenetic or areal influence, our dataset may not include languages without multiplicative numerals. It would be necessary to cross-check the association between the existence/absence of multiplicative numerals and classifiers in a phylogenetically weighted sample of languages. However, such an approach would require additional data and is beyond the scope of the current paper. For the purpose at hand, we proceed to examine whether the second probabilistic universal, stated in (6), also applies to the languages in SMATTI.

4.2 Harmonization between numeral bases and classifiers

Within the 219 languages for which we have information on numeral bases and classifiers, the harmonization between the base-parameter and the C/M-parameter is attested in 213 languages (97.26%), with only 6 exceptions. As shown in Table 5, most of the observed languages are base-final and C/M-final.[4] We do not discuss here the distribution of each category per language family, since it does not influence the verification of the probabilistic universal. Nevertheless, that subject is developed in Section 5. For now, we focus on testing the null hypothesis of no harmonization between numeral bases and classifiers.

Table 5. Observation on the base-parameter and the C/M parameter in SMATTI

	C/M-final	C/M-initial	Total Languages
Base-final	187 (85.39%)	**5 (2.28%)**	192
Base-initial	**1 (0.46%)**	26 (11.87%)	27
Total Languages	188	31	219

[4] Two languages from the Tibeto-Burman family require some explanation. Sunwar is attested to have two numeral systems. However, it is counted as C/M-final and base-final, since base-initial numerals do not co-occur with classifiers in the language. Furthermore, in Rabha, all four attested C/M word orders in (1) are found. Nevertheless, Rabha is also counted as C/M-final and base-final due to the prominence of these word orders over the residual C/M-initial and base-initial orders (see also further discussions in Section 5).

The information encoded in Table 5 is equivalently shown via a bar plot of two-dimensional table in Figure 4. The *x*-axis represents the two categories of C/M word order, whereas the *y*-axis indicates the frequency of base-final (black) and base-initial (gray) languages, respectively. The plot clearly shows that the proportion of base-final languages is greater in the C/M-final than in the C/M-initial, and vice-versa.

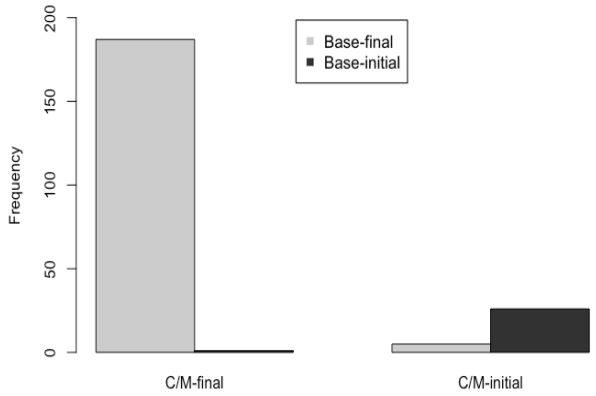

Figure 4. Bar plot of the two-dimensional table

Thus, the proportions of the two base-parameter tokens are clearly different in the two different C/M-parameter groups. We then need to investigate the strength of the effect size and its statistical significance. Effect size is not discussed in detail in the previous probabilistic test due to the different types of data. It is necessary to do so here since we now have one more variable. The effect size represents the magnitude of the difference between groups, while the statistical significance demonstrates the probability that the observed difference across two groups is due to chance (Sullivan & Feinn, 2012, p. 279). For instance, a smaller *p*-value shows that the probability that the divergence between the two groups is less likely to be caused by chance. However, the *p*-value does not tell us the strength of the association between the variables. Thus, it is preferred to analyze the effect size as well.

We first calculate the statistical significance of our observations by carrying out a Pearson's Chi-square (χ^2) test of independence with Yates' continuity correction. We formulate the null hypothesis as the absence of association between the variables (i.e., the base-parameter and the C/M-parameter), while the alternative hypothesis points toward the correlation of the variables. The Chi-square test is based on the comparison of observed and expected frequencies. The former refers to the actual observations in the data, i.e., the actual numbers in our contingency table; the latter indicates the anticipated frequencies resting on the assumption that the variables are independent, i.e., if the null hypothesis is true. The expected frequencies are generated by dividing the product of the marginal frequency of a row and the marginal frequency of a column by the total number of observations. With regard to our data, the expected frequencies are displayed in Table 6. To be more precise, if the null hypothesis is true and there is no association between the base-parameter and the C/M-parameter, the distribution of languages in our dataset should be as shown in Table 6.

Table 6. Expected frequencies of contingency table

	C/M-final	C/M-initial
Base-final	165 (75.26%)	27 (12.41%)
Base-initial	23 (10.58%)	4 (1.75%)

We then apply the Chi-square test to verify if the divergence between our observations in Table 5 and the statistically expected distribution in Table 6 is statistically significant. The formula of the Chi-square test is given in (9). The output of the evaluation is equal to the sum of the square of the differences between the observed (O) and expected values (E) divided by the expected values.

(9) Formula of the Chi-square test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The resulting $\chi^2(1) = 163.38$ and $p < 0.001$, below the level of high significance, permits us to reject the null hypothesis of no association between the two variables.

However, note that one of the values in the expected frequencies is lower than 5 (C/M-initial and base-initial) and may have influenced the result of the Chi-square test. Hence, we apply a two-tailed Fisher's exact test to verify the statistical significance of our observation. The Fisher's exact test calculates the probability of obtaining the values via the hypergeometric sampling distribution of the hypergeometric-likelihood measure. In other words, we divide the product of the factorial of the sum of each row and column via the product of the factorial of the value in every cell along with the factorial of the total amount of observations. As a mean of clarification, the formula of the Fisher's exact test is shown in (10). C_1, C_2, R_1, R_2 indicates the sum of each row and column from the contingency table, whereas V_1, V_2, V_3, V_4 represents the individual value of every cell in the contingency table. Finally, n equals to the sum of all the observations in the data.

(10) Formula of the Fisher's exact test

$$\mathrm{P} = \frac{C_1!\, C_2!\, R_1!\, R_2!}{V_1!\; V_2!\; V_3!\; V_4!\, n!}$$

The resulting $p < 0.001$ indicates that the Fisher's exact test, like the Chi-square test of independence, also rejects the null hypothesis.

Then, we need to calculate the effect size to measure the strength of association between the variables. A simple way to measure effect size is the *odds ratio*. We divide the odds of observing a base-final numeral system in a C/M-final language by the odds of noticing a base-final numeral system in a C/M-initial language, i.e., (187/1)/(5/26)=972.4. This number means that the odds of having a base-final numeral system in a C/M-final language are 972.4 times greater than those in a C/M-initial language. Nevertheless, such a method merely offers a preliminary observation. To measure the effect size in a more appropriate statistical way, we apply *the φ coefficient* (also named mean square contingency coefficient), which is similar to the *Pearson correlation coefficient* and is used to calibrate the degree of association between two binary variables. As shown in (11), the φ coefficient is obtained by the square of the Chi-squared statistic of our contingency table divided by the total number of subjects.

(11) Formula of the φ coefficient

$$\varphi = \sqrt{\chi^2/_n}$$

The obtained φ coefficient varies between zero and one. Thus, the closer the φ coefficient to one, the stronger the association. More specifically, a φ coefficient smaller than 0.3 represents a small effect size; between 0.3 and 0.5 indicates a moderate effect; bigger than 0.5 displays a strong effect. Based on our data, the generated φ coefficient equals to 0.884. Thus, the results of the correlation between the base-parameter and the C/M-parameter in our data not only show a statistically significant association but also a strong effect size.

5. Discussions

While base and C/M are indeed harmonized in word order in most of the languages surveyed (97.26%, 213/219), the harmonized base-final and C/M-final parameter is again the majority and accounts for 85.39% (187/219) of the languages, which is in line with the observation in (2). By way of illustration in (12), Standard Mandarin is a strictly base-final and C/M-final language. Within the numeral structure (12a), the numeral base (e.g., 'hundred') is positioned after the numeral (e.g., 'three'). Similarly, C/Ms follow Num, mimicking the the numeral structure. As shown in (12b), C/Ms are located after the numeral construction 'five-hundred', whereas the noun 'book' comes afterward and does not intervene between Num and C/M. Most cases in our survey follow this pattern.

(12) Base-final and C/M-final word order in Standard Mandarin

a.	三-百	二-十	一	b.	五-百	本/箱	書
	san-bai	*er-shi*	*yi*		*wu-bai*	*ben/xiang*	*shu*
	three-hundred	two-ten	one		five-hundred	C_{volume}/M_{box}	book
	'three hundred twenty-one'				'five hundred (/boxes) of books'		

The second largest type of word order is base-initial combined with C/M-initial, which is likewise harmonized according to the proposed probabilistic universals. Examples from Garo, a Tibeto-Burman language, are given in (13). In (13a), the numeral base 'twenty' precedes the multiplier 'three' within the numeral structure. In (13b), the classifier *ak* also precedes the numeral 'four'. Interestingly, the noun in Garo also appears before C/M and Num, which is the opposite of what we observed in some base-final and C/M-final languages, e.g., Chinese as shown in (12). This word order may thus be interpreted with regard to syntactic heads, i.e., the order within a phrase can be head-final or head-initial, which is expected to be reflected in the general structure of the language. Such hypothesis is in accordance with the two probabilistic universals in terms of numeral base and C/M. However, we leave this for future studies.

(13) Base-initial and C/M-initial word order in Garo

a.	*kolchaŋ-gittam*	b.	*meʔchik*	***ak**-bri*
	twenty-three		woman	C-four
	'sixty'		'four women'	

An overview of the distribution of the base-parameter and the C/M-parameter across language families is shown in Table 7. First, note that the 27 base-initial and C/M-initial languages are exclusively Tibeto-Burman, even though the Tibeto-Burman family contains a majority of base-final (73%, 73/100) and C/M-final (69%, 69/100) languages. Second, the six observed exceptions are also exclusively Tibeto-Burman. This suggests the influence of language phylogeny and language contact.

Table 7. Distribution of the base-parameter and the C/M-parameter in SMATTI by language families

	Base-final	Base-initial	C/M-final	C/M-initial
Sinitic	14	0	14	0
Miao-Yao	8	0	8	0
Austro-Asiatic	39	0	39	0
Tai-Kadai	40	0	40	0
Tibeto-Burman	73	27	69	31
Indo-Aryan	18	0	18	0
Total	219		219	

Therefore, we display the distribution of the base-parameter via spatial representation in Figure 5. The gray dots represent base-final languages, and blue dots, base-initial languages, whereas red circles indicate the languages that violate the base and C/M harmonization. Due to the high ratio of harmonization within our data, the same map can also be interpreted in terms of the C/M-parameter. It shows a picture of base-initial and C/M-initial Tibeto-Burman languages being sandwiched between base-final and C/M-final languages, while the six exception cases, highlighted in red circles, are located on the fringe between the two harmonized word orders.

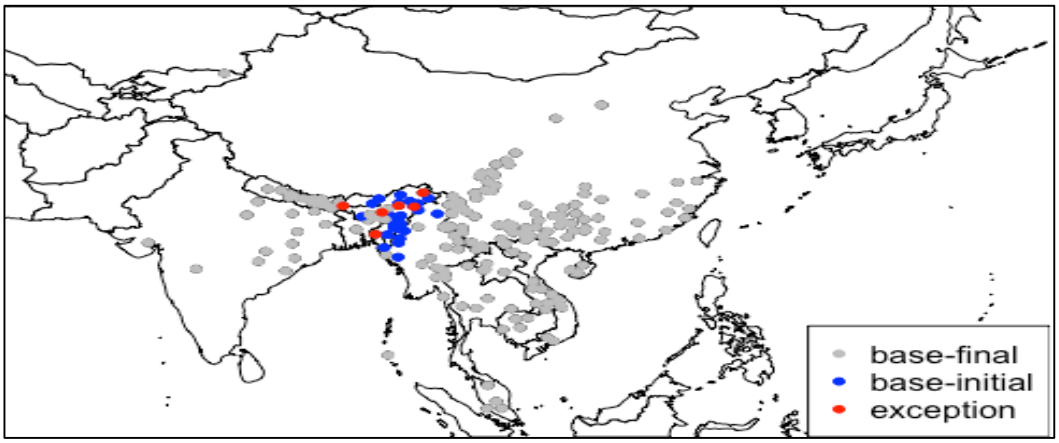

Figure 5. Spatial overview of base-parameter in SMATTI

In fact, the situation of the entire Tibeto-Burman family, as depicted in Figure 2, shows a similar pattern, i.e., this family with various combinations of base and C/M word orders is surrounded by strictly base-final language groups, Sinitic and Tai-Kadai on one side and Indo-Aryan on the other side. SMATTI thus presents an interesting typological sandwich, and we speculate that the middle part, the Tibeto-Burman languages, was initially base-initial and C/M-initial (Matisoff, 1995) but received influence from the base-final and C/M-final languages in the East and the West (Benedict, 1987; Gvozdanovic, 1999; Mazaudon, 2009; Peyraube, 1991), and eventually evolved into today's distribution. While the actual development process of these languages requires additional studies, we present a case study of Rabha from the Tibeto-Burman family to shed some light on this issue of contact-induced language change.

Rabha is spoken in the Indian state of Assam, with around 50,000 speakers according to *Ethnologue* (Lewis et al., 2009). Unlike most other languages in our dataset, which have either an initial or a final base and C/M word order, Rabha uses all four attested word orders of C/M, as illustrated in (14) with data from Joseph (2007) and Chan (2017). The first observed word order is C/M-final (14a) and base-final (14c). However, it may also be C/M-initial (14b) and base-initial (14d).

(14) Base-parameter and C/M-parameter in Rabha

a. *pan* *phaŋ-atham*
tree C-tree
'three trees'

b. *sari-jon* *kai*
four-C person
'four people'

c. *gota-antham*
hundred-three
'three hundred'

d. *dui-so*
two-hundred
'two hundred'

At first glance, Rabha behaves like a drastic violation to our probabilistic universal of harmonization, since the orders of base and C/M can vary freely. There are rules, however, underlying the use of these word orders. Two sets of numerals are attested in Rabha. The predominant version in use is a base-final system. As illustrated in Table 8, the numeral bases of this scheme are consistently positioned after the multiplier numeral.

Table 8. Base-final numerals in Rabha (Chan 2017)

ek so	*dui so*	*ek hajar*	*dui hajar*
one hundred	two hundred	one thousand	two thousand
'100'	'200'	'1000'	'2000'

This system is borrowed from Assamese (Indo-Aryan), which is a dominant language in this area, enjoying an enormous population of 16,000,000 speakers and the prestigious status as one of the official languages in the state of Assam (Lewis et al., 2009). Due to such unbalance in terms of population and use, it is therefore quite common for Rabha to borrow linguistic elements from Assamese, which may gradually replace the indigenous vocabulary and linguistic subsystems, and in this case, the numerals. A sample of numerals from Assamese is shown in Table 9. It not only demonstrates that Assamese is a strictly base-final and C/M-final language, but also shows the phonetic similarity between Rabha and Assamese in terms of numerals.

Table 9. Base-final numerals in Assamese (Chan, 2017)

exa	*duxa*	*ehezar*	*duhezar*
[ɛxɔ]	[duxɔ]	[ɛhɛzar]	[duhɛzar]
'100'	'200'	'1000'	'2000'

The second set of numerals in Rabha is indigenous, but only the numerals *sa* 'one', *niŋ* 'two', and *tham* 'three' are still in use. However, the overall numeral system is still remembered by elder speakers and documented in the literature. Hence, we are able to identify it unmistakably as a base-initial system. In Table 10, *gota-aniŋ* 'two hundred' is composed of the base *gota* 'hundred' and the numeral *aniŋ* 'two', with the latter in the second position, showing a base-initial pattern.

Table 10. The original numeral system of Rabha (Joseph, 2007)

1. *sa*	11. *goda-sa*	199. *gota-sa pinsip-pindas*
2. *niŋ*	20. *rikha*	200. *gota-aniŋ*
3. *tham*	30. *siri*	300. *gota-antham*
4. *ari*	40. *arli*	400. *gota-ari*
5. *campa*	50. *phala*	500. *gota-campa*
6. *hes*	60. *hesti*	600. *gota-hes*

7. *sorta*	70. *sorto*	700. *gota-sorta*
8. *parta*	80. *arsi*	800. *gota-parta*
9. *pindas*	90. *pinsip*	900. *gota-pindas*
10. *goda*	100. *gota-sa*	1000. *hajar-sa*

Intriguingly, C/Ms in Rabha can also be divided into two groups, depending on their word order and respective origin. The C/M-final classifiers are Assamese loans and can only be used with base-final loaned numerals likewise from Assamese. The C/M-initial indigenous classifiers can only appear with the three surviving indigenous numbers, which are part of a base-initial system no longer in use. Rabha thus provides strong evidence for the base-C/M harmonization, which is not only confirmed in a cross-language manner, but also observed language-internally.

The observed violations to the probabilistic universal in Tibeto-Burman can thus receive a preliminary explanation. Rabha demonstrates a developmental stage of a contact-induced process of a systematic change of grammatical features. Violations found in this survey may be due to a similar reason, given that they are all distributed along the edge between base-final Indo-Aryan languages and base-initial Tibeto-Burman languages. Tiwa, one of the languages of violation, for example, is C/M-initial but base-final (Emeneau, 1956). It is thus not surprising that, except the numerals for one and two, Tiwa numerals are also borrowed from Assamese (Chan, 2017). Another Tibeto-Burman language, Kok Borok, which is also C/M-initial (Jacquesson, 2007), shows an even more complex state with both base-final and base-initial numerals (see Table 11), a result of influence from the base-final Bengali. A reverse kind of violation is found in Konyak Naga, which is C/M-final and base-initial (Chan, 2017; Emeneau, 1956).

Table 11. The numeral system of Kok Borok (Chan, 2017)

1. *ʂa*	10. *tʃi*	100. *ra ʂa*
2. *nuĵ*	20. *tʃinuĵ*	200. *nuĵ ra*
3. *tʰam*	30. *tʰamtʃi*	1000. *hadʒar ʂa*
4. *buruĵ*	40. *buruĵtʃi*	2000. *nuĵ hadʒar*
5. *ba*	50. *batʃi*	
6. *doǰk*	60. *doǰktʃi*	
7. *ʂɪni*	70. *ʂɪnitʃi*	
8. *tʃar*	80. *tʃatʃi*	
9. *ʃiku*	90. *ʂikutʃi*	

The cases above show different degrees of contact-induced change. It is then possible that the violations to harmonization are the results of such language contact. Another piece of evidence is the geographic distribution of base-final and base-initial languages, shown in Figure 9. Although not all gray dots (base- and C/M-final languages) are located on the plains, blue dots (base-initial and C/M-initial languages) are concentrated in the mountainous areas between India and Myanmar, while a few are located sporadically at the southern edge of the Tibetan Plateau. A scenario involving a gradual process of contact-induced change is therefore reasonable. Tibeto-Burman languages, originally base-initial and without classifiers, have long been under the pressure from the neighboring base-final classifier languages, which are politically, socially, and economically more powerful. Tibeto-Burman languages have thus gradually adopted their numeral systems and also acquired their classifier feature, and evolved into today's different degrees of language change. The more isolated languages in the mountainous areas have been more protected by the geographic barriers; they have retained

more of the original systems. We acknowledge that such a hypothesis requires additional data and analysis based on a phylogenetic and statistical approach.

6. Conclusion

In this paper, a multiplicative theory uniting numeral bases and classifiers is proposed to explain two probabilistic linguistic universals, which were first observed by Greenberg (1990b). These probabilistic universals are further derived from the multiplicative theory and tested in the world's foremost hotbed of classifier languages, namely SMATTI (Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai, Tibeto-Burman, Indo-Aryan). The two probabilistic universals are: 1) the presence of classifiers entails the existence of multiplicative numerals in a language, 2) the base-parameter and the C/M-parameter are harmonized within a language. The results of our typological analysis show that we can reject the null hypotheses of no-association with high statistical significance. Moreover, we measured a strong effect size of harmonization between the base-parameter and the C/M-parameter. The encountered exceptions are exclusively from the Tibeto-Burman family, and are tentatively explained by different stages of contact-induced language change.

The limitations of our study include a lack of phylogenetically weighted sample of languages, as we have selected languages from six specific language groups. The methodology employed may be affected by the modifiable areal unit problem. Moreover, we have only involved classifier languages in our survey. Additional evidence may be obtained by running the statistical tests on a phylogenetically weighted sample of languages across the world that include both classifier languages and attested non-classifier languages. Furthermore, we have demonstrated that the base-parameter and the C/M-parameter are harmonized. Yet, we did not provide a concrete theoretical foundation as to why the base-final and C/M-final word orders are more frequent. Hence, additional research is likewise needed in this regard.

Acknowledgements

To be added

References

Adams, K. L. (1986). Numeral classifiers in Austroasiatic. In C. Craig (Ed.), *Noun classes and categorization* (pp. 241–262). Amsterdam: John Benjamins.

Aikhenvald, A. Y. (2000). *Classifiers: A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.

Allan, K. (1977). Classifiers. *Language*, *53*(2), 285–311.

Au Yeung, W. H. B. (2005). *An interface program for parameterization of classifiers in Chinese* (PhD dissertation). Hong Kong University of Science and Technology, Hong Kong.

Au Yeung, W. H. B. (2007). Multiplication basis of emergence of classifiers. *Language and Linguistics*, *8*(4), 835–861.

Bauer, C. (1992). Adams Karen Lee: Systems of numeral classification in the Mon-Khmer, Nicobarese and Aslian subfamilies of Austroasiatic. (Pacific Linguistics, Series B, no. 101) xiii, 219 pp. Canberra: Australian National University, Research School of Pacific Studies, 1989 [pub.1990]. *Bulletin of the School of Oriental and African Studies*, *55*(2), 374–378.

Benedict, P. K. (1987). Early MY/TB loan relationships. *Linguistics of the Tibeto-Burman Area*, *10*(2), 12–21.

Bhattacharya, T. (2001). Numeral/quantifier-classifier as a complex head. In N. Corver & H. C. van Riemsdijk (Eds.), *Semi-lexical categories: The function of content words and the content of function words* (pp. 191–221). Berlin: Mouton de Gruyter.

Bickel, B. (2014). Linguistic diversity and universals. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The Cambridge handbook of linguistic anthropology* (pp. 101–124). Cambridge: Cambridge University Press.

Bisang, W. (1999). Classifiers in East and Southeast Asian languages: counting and beyond. In J. Gvozdanovic (Ed.), *Numeral Types and Changes Worldwide* (pp. 113–186). Munchen: Walter de Gruyter.

Biswas, P. (2013). Plurality in a classifier language: two types of plurals in Bangla. *Proceedings of Generative Linguists of the Old World in Asia (GLOW in Asia)*, 1–14.

Chan, E. (2017). *Numeral systems of the world languages*. Retrieved from https://mpi-lingweb.shh.mpg.de/numeral/.

Comrie, B. (2006). Numbers, language, and culture. Presented at the Jyvaskyla Summer School, Jyvaskyla.

Comrie, B. (2013). Numeral bases. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dayal, V. (2014). Bangla Plural Classifiers. *Language and Linguistics*, *15*(1), 47–87. https://doi.org/10.1177/1606822X13506151

Dryer, M. S. (1998). Why statistical universals are better than absolute universals. *Papers from the 33rd Regional Meeting of the Chicago Linguistic Society*, 1–23.

Emeneau, M. B. (1956). India as a Lingustic Area. *Language*, *32*(1), 3–16. https://doi.org/10.2307/410649

Fu, J. (2015). The status of classifiers in Tibeto-Burman languages. In D. Xu & J. Fu (Eds.), *Space and quantification in languages of China* (pp. 37–54). Dordrecht: Springer.

Gil, D. (2013). Numeral classifiers. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Greenberg, J. H. (1990a). Generalizations about numeral systems. In K. Denning & S. Kemmer (Eds.), *On language: Selected writings of Jeseph H. Greenberg* (pp. 271–309). Stanford: Stanford University Press.

Greenberg, J. H. (1990b). Numeral classifiers and substantival number: Problems in the genesis of a linguisitc type. In K. Denning & S. Kemmer (Eds.), *On language: Selected writings of Joseph H. Greenberg* (pp. 166–193). Stanford: Stanford University Press.

Gvozdanovic, J. (1999). *Numeral types and changes worldwide*. Berlin: Mouton de Gruyter.

Her, O.-S. (2012a). Distinguishing classifiers and measure words: A mathematical perspective and implications. *Lingua*, *122*(14), 1668–1691. https://doi.org/10.1016/j.lingua.2012.08.012

Her, O.-S. (2012b). Structure of classifiers and measure words: A lexical functional account. *Language and Linguistics*, *13*, 1211–1251.

Her, O.-S. (2017a). Deriving classifier word order typology, or Greenberg's Universal 20A and Universal 20. *Linguistics*, *55*(2), 265–303.

Her, O.-S. (2017b). Structure of numerals and classifiers in Chinese: Historical and typological perspectives and cross-linguistic implications. *Language and Linguistics*, *18*(1), 26–71.

Her, O.-S., Chen, Y.-C., & Yen, N.-S. (2017). Mathematical values in the processing of Chinese numeral classifiers and measure words. *PLOS ONE*, *12*(9), 1–9.

Jacquesson, F. (2007). Kok-Borok, A short analysis. In *Hukumu, 10th anniversary volume* (pp.

109–122). Agartala: Kokborok Tei Hukumu Mission.
Joseph, U. V. (2007). *Rabha*. Leiden: Brill.
Kilarski, M. (2014). The Place of Classifiers in the History of Linguistics. *Historiographia Linguistica*, *41*(1), 33–79. https://doi.org/10.1075/hl.41.1.02kil
Lewis, P., Simons, G. F., & Fennig, C. D. (2009). *Ethnologue: Languages of the world*. Dallas: SIL International.
Luraghi, S. (2011). The origin of the Proto-Indo-European gender system: Typological considerations. *Folia Linguistica*, *45*(2), 435–464.
Matisoff, J. A. (1995). Sino-Tibetan numerals and the play of prefixes. *Bulleting of the National Museum of Ethnology (Osaka)*, *20*(1), 105–251.
Mazaudon, M. (2009). Number-building in Tibeto-Burman languages. In S. Morey & M. W. Post (Eds.), *North East Indian linguistics* (pp. 117–148). Cambridge: Foundation Books.
Morev, L. (2000). Some afterthoughts on classifiers in the Tai languages. *Mon-Khmer Studies*, *30*, 75–82.
Mortensen, D. R. (2017). Hmong-Mien languages. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics* (pp. 1–25). Oxford: Oxford University Press.
Peyraube, A. (1991). Some remarks on the history of Chinese classifiers. *Santa Barbara Papers in Linguistics*, *3*, 106–126.
Peyraube, A. (1998). On the history of clsasifiers in Archaic and Medieval Chinese. In B. K. T'sou (Ed.), *Studia linguistica serica* (pp. 131–145). Hong Kong: City University of Hong Kong.
Piantadosi, S. T., & Gibson, E. (2014). Quantitative Standards for Absolute Linguistic Universals. *Cognitive Science*, *38*(4), 736–756. https://doi.org/10.1111/cogs.12088
Sinnemaki, K. (to appear). On the distribution and complexity of gender and numeral classifiers. In F. Di Garbo & B. Wälchli (Eds.), *Grammatical gender and linguistic complexity*. Berlin: Language Science Press.
Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the *P* Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1
Tai, J., & Wang, L. (1990). A semantic study of the classifier Tiao. *Journal of the Chinese Language Teachers Association*, *25*(1), 35–56.
Velupillai, V. (2012). *An introduction to linguistic typology*. Amsterdam: John Benjamins.
Wu, F., Feng, S., & Huang, C. T. J. (2006). Hanyu shu+lianhg+ming geshi de laiyuan [On the origin of the construction of 'numeral+classifier+noun'in Chinese]. *Zhongguo Yuwen [Studies of the Chinese Language]*, *5*, 387–400.
Xu, D. (2013). *Plurality and classifiers across languages in China*. Munchen: Walter de Gruyter.
Yi, B. U. (2009). Chinese classifiers and count nouns. *Journal of Cognitive Science*, *10*, 209–225.
Yi, B. U. (2011). What is a numeral classifier? *Philosophical Analysis*, *23*, 195–258.
Zhang, N. N. (2013). *Classifier structures in Mandarin Chinese*. Berlin: Mouton de Gruyter.

Appendix - Classifier languages in SMATTI

Sinitic

Dungan	Mandarin Chinese	Pu-Xian Chinese
Gan Chinese	Min Bei Chinese	Wu Chinese
Hakka Chinese	Min Dong Chinese	Xiang Chinese

Huizhou Chinese	Min Nan Chinese	Yue Chinese
Jinyu Chinese	Min Zhong Chinese	

Miao-Yao		
Biao-Jiao Mien	Hmong Njua	Pa-Hng
Bu-Nao Bunu	Jiongnai Bunu	She
Hmong daw	Northern Qiandong Miao	

Austro-Asiatic		
Blang	Khasi	Parauk Wa
Bondo	Khmu	Pear
Bugan	Lave	Pnar
Car Nicobarese	Lyngngam	Prai
Central Khmer	Mae Hong Son Lawa	Ruching Palaung
Chong	Mah Meri	Samre
Chrau	Mal	Samtao
Eastern Bru	Man Met	Santali
Eastern Katu	Mang	Sapuan
Jah Hut	Mon	Sedang
Jeh	Northern Khmer	Semelai
Jehai	Nyahkur	So
Kharia	Pacoh	Vietnamese

Tai-Kadai		
Ahom	Lingao	Sui
Baha Buyang	Liujiang Zhuang	Tai Daeng
Biao	Lu	Tai Dam
Bouyei	Mak	Tai Don
Chadong	Maonan	Tai Nua
Cun	Mulam	Ten
Dai Zhuang	Nong Zhuang	Thai
Gelao	Northern Dong	White Gelao
Guibei Zhuang	Nung	Yang Zhuang
Guibian Zhuang	Qabiao	Yongbei Zhuang
Hlai	Red Gelao	Youjiang Zhuang
Lachi	Saek	Zuojiang Zhuang
Lakkia	Shan	
Lao	Southern Dong	

Tibeto-Burman		
Achang	Haka Chin	Rawang
Adi	Hani	Sangkong
Akha	Hmar	Sani
Angami Naga	Horpa	Sgaw Karen
Anu	Idu-Mishmi	Shixing
Apatani	Inpui Naga	Simte
Atong	Jiarong	Southern Bai
Axi	Jingpho	Southern Pumi

Azhe	Kado	Southern Qiang
Baima	Karbi	Southern Tujia
Bantawa	Katso	Sunwar
Bhujel	Kok Borok	Tawang Monpa
Bisu	Konyak Naga	Thado Chin
Bodo	Lahu	Thangmi
Burmese	Lashi	Thulung
Camling	Leinong Naga	Tiwa
Central Bai	Lhao Vo	Tshangla
Chak	Lisu	Ugong
Chantyal	Miri	Usoi
Chhintange	Mizo	Vaiphei
Daai Chin	Muya	Wambule
Deori	Namuyi	Wayu
Dhimal	Newar	Western Gurung
Dimasa	Nocte Naga	Western Kayah
Drung	Northern Bai	Xiandao
Dumi	Northern Pumi	Yakha
Eastern Kayah	Northern Qiang	Yamphu
Ersu	Northern Tujia	Youle Jinuo
Falam Chin	Nung	Zaiwa
Galo Adi	Paite Chin	Zauzou
Gangte	Pela	Zhaba
Garo	Puma	Zou
Geba Karen	Queyu	
Guiqiong	Rabha	

Indo-Aryan

Assamese	Chhattisgarhi	Maithili
Awadhi	Darai	Marathi
Balkan Romani	Fiji Hindi	Nepali
Bengali	Gujarati	Oriya
Bhojpuri	Halbi	Rajbanshi
Bishnupriya	Hindi	Sadri