Word order of numeral classifiers and numeral bases

One-Soon Her, Marc Tang, Bing-Tsiong Li

▶ **To cite this version:**

One-Soon Her, Marc Tang, Bing-Tsiong Li. Word order of numeral classifiers and numeral bases. STUF, 2019, 72 (3), pp.421-452. 10.1515/stuf-2019-0017 . hal-02529136

HAL Id: hal-02529136
https://hal.science/hal-02529136

Submitted on 4 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

One-Soon Her, Marc Tang* and Bing-Tsiong Li

Word order of numeral classifiers and numeral bases

Harmonization by multiplication

https://doi.org/10.1515/stuf-2019-0017

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). Num and C/M are adjacent cross-linguistically, 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 Mandarin. However, the base may also precede the multiplier in some languages, thus [*base* × *n*]. Interestingly, base and C/M seem to harmonize in word order, i.e., [*n* × *base*] numerals appear with a [Num C/M] alignment, and [*base* × *n*] numerals, with [C/M Num]. This paper follows up on the explanation of the base-C/M harmonization based on the multiplicative theory of classifiers 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. Moreover, the few scarce violations to the base-C/M harmonization are also all in TB and are mostly languages having maintained their original base-initial numerals but borrowed from their base-final and C/M-final neighbors. We thus 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

***Corresponding author: Marc Tang,** Department of Linguistics and Philology, Uppsala University, P.O. Box 635, Uppsala 75126, Sweden, E-mail: marc.tang@lingfil.uu.se.
One-Soon Her, Graduate Institute of Linguistics/Research Center for Brain, Mind, and Learning, National Chengchi University, Taipei City 11605, Taiwan (R.O.C.), E-mail: hero@nccu.edu.tw.
Bing-Tsiong Li, Graduate Institute of Linguistics, National Chengchi University, Taipei City 11605, Taiwan (R.O.C), E-mail: xhoques@gmail.com.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

1 Introduction

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

(1) Classifiers in Mandarin Chinese

a. 三百 本 書
San-bai *ben* *shu*
three-hundred C_{volume} book
'three hundred books'

b. 三千 桶 水
San-qian *tong* *shui*
three-thousand M_{bucket} water
'three thousand buckets of water'

This paper investigates the harmonization of word order between the numeral classifier, e.g., *ben* and *tong* in (1), and the numeral base, e.g., *bai* and *qian* in (1), in classifier languages, as recent studies claim that such harmonization is the consequence of the shared function as multiplicands between the C/M and the numeral base (Her 2017a,b). A language like Mandarin with a C/M-final word order, i.e., [Num C/M], thus tends to have a base-final multiplicative structure, i.e., [*n base*], while a C/M-initial language should have base-initial complex numerals, as in Kilivila, an Austronesian language of the Trobriand Islanders (Senft 1986: 77–80, 2000: 18–21).

Two potential probabilistic universals have been derived from this hypothesis. First, the presence of C/M in a language implies the presence of

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

2 In this paper, a distinction is made between *mensural classifiers* in languages such as Mandarin Chinese and *terms of measure* in languages such as English (Her 2012a: 1682). English has terms of measure such as 'bucket' in *three buckets of water*. However, syntactically such terms are nouns rather than numeral classifiers since they can take plural morphology. Note, however, that some words in English do seem to behave like classifiers in Mandarin Chinese, e.g., 'head' in *three head of cattle*, and does not take plural marking. For further discussion on this theoretical distinction, see Beckwith (2007: 78–79) and Nomoto (2013: 10–11).

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

multiplicative numerals in the same language. Second, given their common underlying function as multiplicands, the word order of base and C/M is expected to be harmonized, both being either head-final or head-initial. While qualitative analyses have been performed from a theoretical approach on a small set of languages, no quantitative data has been provided to verify these two probabilistic universals. This paper aims at filling this gap through a systematic survey of six language families in which languages commonly use numeral classifiers, i.e., Sinitic, Miao-Yao (aka Hmong-Mien), Austro-Asiatic, Tai-Kadai, Tibeto-Burman, and Indo-Aryan, dubbed SMATTI.

The two potential universals, if indeed proven to be statistically significant, have critical consequences for classifier word order typology and the formal structure of the classifier phrase. If base and C/M are harmonized in word order, then indeed Num, which can be of the multiplicative [*n base*] composition, and C/M must be adjacent, thus excluding exactly the two unattested classifier word orders (Her 2017a). This fact then supports a left-branching structure, where Num and C/M form a constituent first before merging with N, i.e., [[Num C/M] N], over the right-branching constituency, i.e., [Num [C/M N]].

The paper is organized as follows. Section 2 introduces the multiplicative theory of classifiers, from which two probabilistic universals are derived. Section 3 explains how the six language families have been selected and how the data has been gathered. Section 4 presents respectively the validity of the two probabilistic universals in our survey. Section 5 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

Six possibilities are expected mathematically in terms of word orders formed by Num, C/M and N, i.e., $3! = 3 \times 2 \times 1 = 6$, as in (2). For example, Chinese is consistently C/M-final throughout its 3,000 years of recorded history. Num precedes C/M, as seen in (1); as opposed to languages that are C/M-initial, i.e., C/M precedes Num. Crucially, however, it has been observed that cross-linguistically N does not intervene between Num and C/M (Aikhenvald 2000: 104–105; Greenberg 1990[1972]: 185; Peyraube 1998; Wu et al. 2006).

(2) Six possible word orders of Num, C/M, and N
[Greenberg 1990[1972]: 185]
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)

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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, 1990[1978]: 292), have been derived and supported by theoretical and typological evidence. They are dubbed 'Greenberg's Universal 20A' (Her 2017a: 265).

(3) Greenberg's Universal 20A [Her 2017a: 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 interesting 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? Greenberg (1990[1972]: 172) first considered the operation between Num and C as multiplication, i.e., [Num C] = [Num × *1*]. With this knowledge, several studies (Au Yeung 2005, 2007; Her 2012a,b; Yi 2009, 2011) proposed 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*. The different types of mathematical values for C/M are summarized in Table 1, with examples from Mandarin Chinese (Her et al. 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

3 Some languages such as Ejagham (Niger-Congo) and Nung (Tai-Kadai) are alleged to have the [C/M N Num] order in (2e). However, it has been shown that these are not genuine classifier constructions. Her (2017a) argue that the alleged classfiers in Ejagham are nouns and the alleged numeral *one* in Tai-Kadai putative violations is not a numeral but a singular indifinite article, like *a/an* in English.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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, non-numerical, fixed, and variable. For instance, an M may have a fixed numerical value: *san da bi* (three M_{dozen} pen) 'three dozen 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 necessarily bear a numerical value of *1*, while Ms apply all sorts of other values.

The concept of multiplication is found within the numeral systems of most languages in the world. Comrie (2013) conducted an extensive survey of numeral systems 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 $[(3 \times 100) + (2 \times 10) + 1]$. 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 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 [Hu 1983 [41802]]

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

'captured 212 birds'

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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 (1990[1978]: 293) and recently further developed by Her (2017a,b).

(6) Harmonization between base and C/M [Her 2017a: 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). In essence, elements that function as multiplicands should naturally follow the same word order in a language (Her 2012a: 279), as in (6). 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 a 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 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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

language to have classifiers, but the word order of C/M must be harmonized according to (6).[4]

To summarize, two potential probabilistic universals based on the multiplication relation between numerals and C/M have been proposed in the literature. First, the presence of C/M in a language implies the use of multiplication in that language, and thus the presence of multiplicative numerals. Second, by reason of their common underlying function of multiplicands, the word order of base and C/M is expected to be harmonized, both being final or initial.

3 Methodology

In terms of geographical distribution, classifier languages are commonly found in the eastern and south-eastern parts of Asia, while their presence is also sporadically attested in Europe, Africa, and 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.

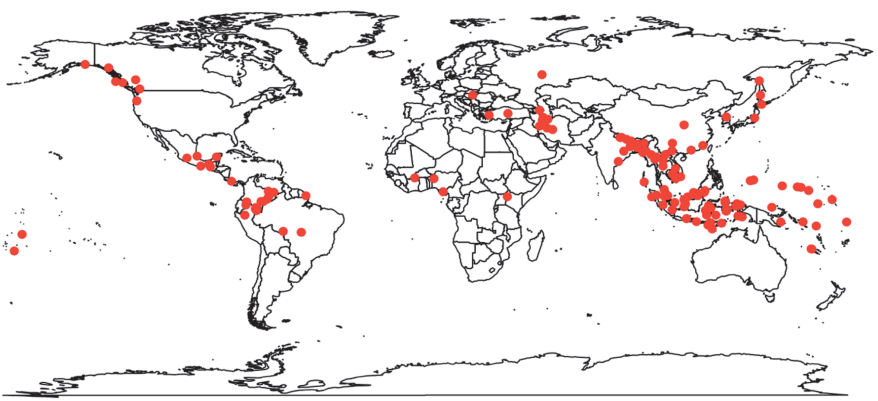

Figure 1: An overview of classifier languages in the world (Gil 2013).

4 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.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

To evaluate the two probabilistic universals, we thus performed a systematic survey of classifier languages in Asia. More specifically, we have chosen 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 a choice may limit the geographical and phylogenetic diversity of our samples; yet, in order to assure a certain level of quality, we narrowed 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 (Sinnemäki 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* (Simons and Fennig 2018). Each point represents one language, and each language is represented once on 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 the population size of speakers or the place of origin. The six language groups investigated in our study are distinguished by six different colors.

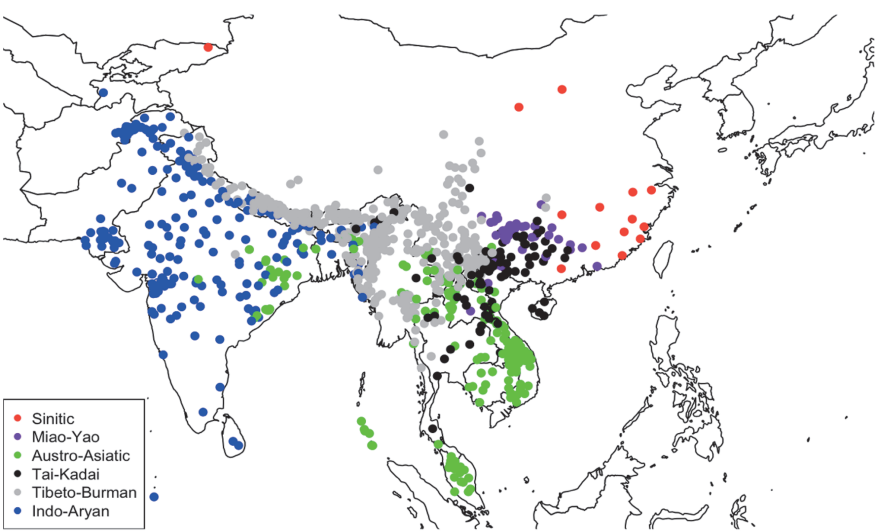

Figure 2: A spatial overview of all languages in SMATTI.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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)[5], as shown in Table 2. This language is then analyzed and annotated as base-final, since the morpheme of 'ten' [pʰu] follows the multipliers from 20 to 90, e.g., in 'thirty' [la:j pʰu], the multiplicand (i.e., the numeral base) [pʰu] 'ten' follows the multiplier [la:j] 'three'. The same pattern is observed in higher numbers such as hundreds and thousands, e.g., [ʔa:r spʰaʔ] 'two-hundred' is the combination of [ʔa:r] 'two' and [spʰaʔ] 'hundred'.

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

1. *wej // ʃi*	10. *ʃi pʰe:u*	100. *ʃi spʰaʔ*
2. *ʔa:r*	20. *ʔa:r pʰu*	200. *ʔa:r spʰaʔ*
3. *la:j*	30. *la:j pʰu*	1000. *ʃi hɑ͡ʤa:r*
4. *sa:o*	40. *sa:o pʰu*	2000. *ʔa:r hɑ͡ʤa:r*
5. *san*	50. *san pʰu*	
6. *hnri:u*	60. *hnri:u 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, or numerative, among others. Moreover, linguistic elements that are not classifiers in our definition may be labeled as classifiers by other publications. Hence, we apply the formal and semantic tests developed in the literature (Her and Hsieh 2010; Her 2012a,b) 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 human collective classifier *ra*, among others. By reviewing the examples provided and cross-checking different sources (Biswas 2013: 2; Dayal 2014: 49), most classifiers fit the definition of our

5 Detailed page numbers from Chan (2017) are not listed since the data is only displayed as an online version without specific page numbers affiliated to each language. However, languages are categorized by language families. Readers are thus encouraged to visit the website mentioned in the reference for further details.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

study. As an example, in *tin-jon chele* (three-C_{human} boy) 'three boys' (Biswas 2013), the sortal classifier *jon* highlights an inherent semantic property (i.e., humanness) of the referent 'boys'. However, the collective classifier *gulo* is not included in our analysis, since its syntactic behavior is not in accord with the canonical behavior of classifiers. For instance, *gulo* is incompatible with numerals (Dayal 2014: 62), which makes it more likely to be a plural marker than a numeral classifier in our judgment, c.f., **boi tin gulo* (book three C) and **tin boi gulo* (three book C).

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: 287; Emeneau 1956: 11). Nevertheless, it is still counted as a classifier language in our database.[6] As a result, we were able to identify both the numeral and classifier systems of 219 classifier languages from SMATTI (22.60%, 219/969). The total quantity of languages and the ratio of classifier languages per family is displayed in Table 3. The full list is in the Appendix.

Table 3: Genealogical distribution of the 219 classifier languages included in our 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

6 The actual usage of classifiers in modern Marathi is subject to discussion as some Marathi speakers tend to use these classifiers as nouns (Pär Eliasson p.c.). Yet, we base our current decision on the published data available.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

(Morev 2000). With regard to Miao-Yao, few numeral classifier languages are found, 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: 15).[7] We thus did not include them in our dataset (the same logic applies for other language groups). As for the Austro-Asiatic group, Bauer (1992: 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 represents a high level of diversity probably due to different language contact situations (Adams 1986: 256–257). Such a 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: 45–46). Finally, Indo-Aryan languages show a very small ratio of classifier languages, which is expected since Indo-European languages generally rely on other systems of nominal classification such as grammatical gender (Corbett 2013; Luraghi 2011).

We are aware that such a general picture is still subject to controversy, as some studies attest that Miao-Yao, Austro-Asiatic, Tai-Kadai, and Tibeto-Burman widely use classifier devices (Xu 2013: 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 a restricted number of speakers. We only included in our dataset classifier languages that 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.

Furthermore, our current aim is to verify the harmonization of multiplicative numerals and numeral classifiers. We thus 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-

7 Noun classifiers also occur next to the noun, but they are independent of other constituents such as numerals and are thus distinguished from numeral classifiers. Noun classifiers are generally found in Australian languages and in Mesoamerican languages (Seifart 2010: 722).

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

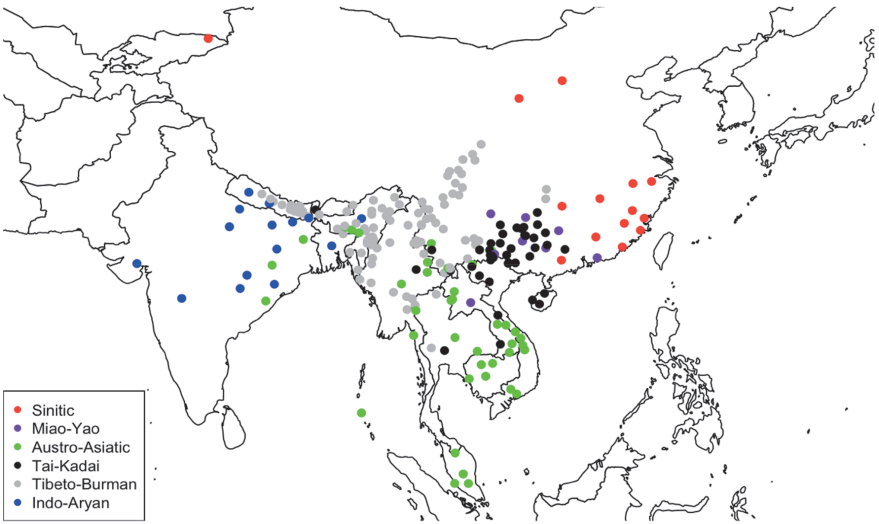

Figure 3: Spatial overview of the 219 languages surveyed.

branches of a language group rather than an across-the-board property of 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 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: 31). We do not argue 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 and Gibson 2014: 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 did have data on all current languages, an observation in the data does not theoretically justify that it applies to all languages. There is no way of knowing if languages no longer spoken or hypothetically possible human languages that have not emerged due to historical accident do not contradict an absolute universal (Dryer 1998).

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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. 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 a probabilistic approach is hence "explored in the same theory-hypothesis-statistics triangle that characterizes most sciences" (Bickel 2014: 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 odds 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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

The data required to testify the probabilistic universal only involves a binomial variable, i.e., with/without multiplication. We thus 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 tails is $10/2 = 5$ times. Nevertheless, the presence of multiplicative numerals in languages of the world does not follow such a 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 follows: 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 a 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 refers to the probability of success.

(8) Formula of the Exact Binomial Test

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

An exact binomial two-tailed test with a 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 ($p < 0.001$) and allows 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 the first probabilistic universal.

We are aware that a more extensive survey of languages of the world is required to fully 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.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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 six exceptions. As shown in Table 5, most of the observed languages are base-final and C/M-final.[8] 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

The information encoded in Table 5 is equivalently shown via a bar plot of a 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.

The proportions of the two base-parameter tokens are thus 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

8 Two languages require some explanation. Sunwar (Tibeto-Burman) 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 (Tibeto-Burman), all four attested C/M word orders in (2) 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 discussion in Section 5).

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

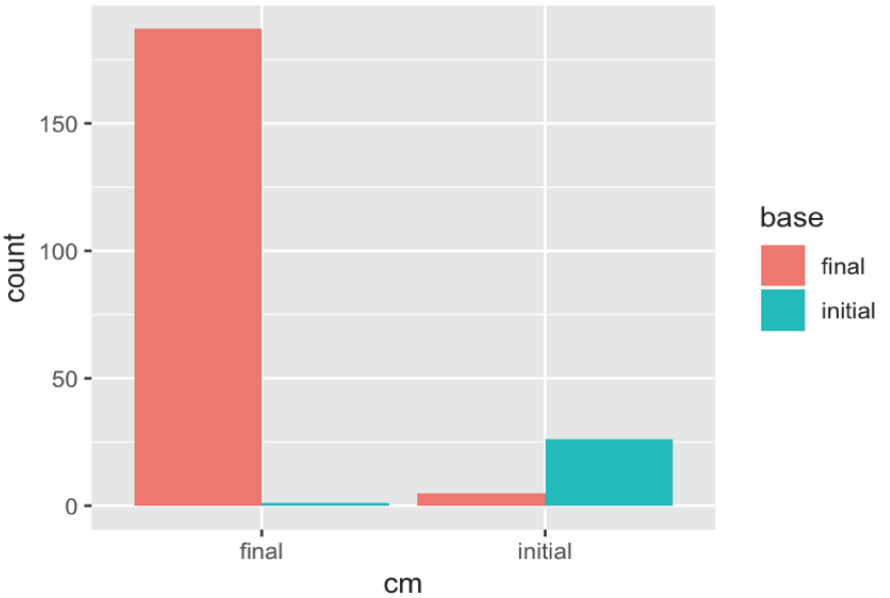

Figure 4: Bar plot of the two-dimensional table.

across two groups is due to chance (Sullivan and Feinn 2012: 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. It is therefore preferable 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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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%)

the C/M-parameter, the distribution of languages in our dataset should be as shown in Table 6.

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 χ^2 (1) = 163.38 and $p < 0.001$ below the level of high significance permit 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 five (C/M-initial and base-initial) and may have influenced the result of the Chi-square test. We thus 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 means 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 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.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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. 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 is equal to 0.884. The results of the correlation between the base-parameter and the C/M-parameter in our data thus not only show a statistically significant association but also a strong effect size.

5 Discussion

In this section, we provide additional details with regard to the distribution of word order patterns within the languages of our dataset. Then we discuss the effect of contact on languages of the area through a case study of Rabha (Tibeto-Burman).

5.1 An overview of languages in the dataset

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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

(12), 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 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 Mandarin

a. 三-百 二-十 一
san-bai *er-shi* *yi*
three-hundred two-ten one
'three hundred twenty-one'

b. 五-百 本/箱 書
wu-bai *ben/xiang* *shu*
five-hundred C_{volume}/M_{box} book
'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 'ten' precedes the multiplier 'four' within the numeral structure. In (13b), the classifier *sak* 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 a hypothesis is in accordance with the two probabilistic universals in terms of numeral base and C/M, but we leave this for future studies.

(13) Base-initial and C/M-initial word order in Garo, adapted from Burling (2004: 243; 245)

a. *sot-bri*
ten-four
'forty'

b. *me?chik* *sak-bri*
woman C-four
'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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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	

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 (Bodo, Deori, Idu-Mishmi, Kok Borok, Konyak Naga, Tiwa) are also exclusively Tibeto-Burman. This suggests the influence of language phylogeny and language contact.

Therefore, we display the distribution of the base-parameter via spatial representation in Figure 5. The gray dots represent base-final languages, and

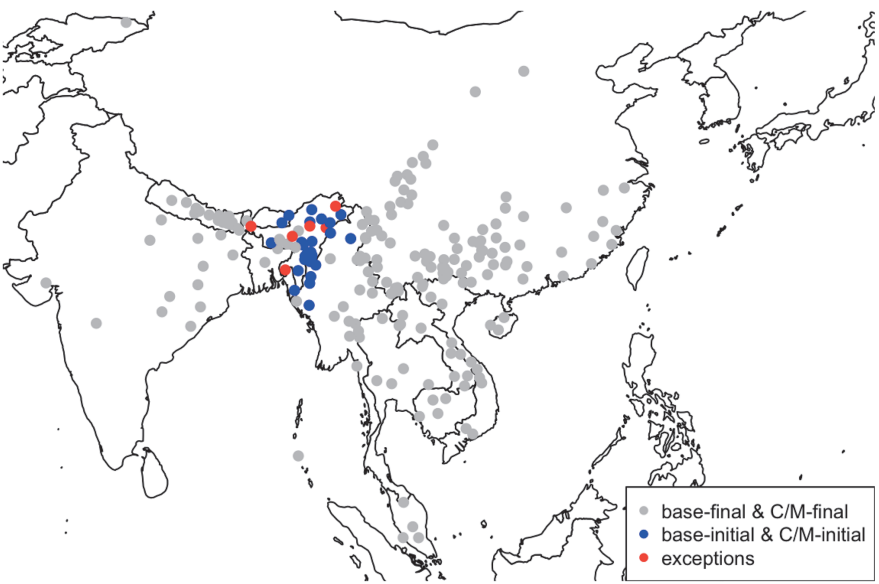

Figure 5: Spatial overview of word order harmonization in SMATTI.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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.

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; Gvozdanović 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.

5.2 A case study of Rabha

Rabha is spoken in the Indian state of Assam, with around 50,000 speakers according to *Ethnologue* (Simons and Fenning 2018). Unlike most other languages in our dataset, which have either an initial or a final base and C/M word order, Rabha uses base-initial, base-final, C/M-initial, and C/M-final word orders, as illustrated in (14) with data from Joseph (2007) and Chan (2017). The first observed word order is C/M-final (14b) and base-final (14d). However, it may also be C/M-initial (14a) and base-initial (14c).

(14) Base-parameter and C/M-parameter in Rabha, adapted from Joseph (2007: 439; 435; 672; 844)

a. *hat* *pak-ŋatham*
week C-three
'three weeks'

b. *pas-jon* *kai*
five-C person
'five people'

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

d. *tin-so*
three-hundred
'three hundred'

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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 (Simons and Fennig 2018). In the south, on the other hand, is Bengali (Indo-Aryan), which is also a prestigious language enjoying the status of the official language of Bangladesh, with more than 82,500,000 speakers solely in India. The two languages therefore exert great influence on the other regional languages such as Rabha e.g., words borrowed from both languages can be found in Rabha. Furthermore, Rabha and Assamese are also relatively similar (e.g., in terms of phonetics), which allows most Rabha texts to be written in Assamese scripts. Due to such similarities crossed with imbalance 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 (Kondakov 2013: 7). Some communities of Rabha speakers have even given up Rabha and shifted entirely to Assamese (Joseph 2007). In the communities where Rabha still survives, it is the numeral system that has almost been replaced by the Assamese one. 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'

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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. It is not surprising to find a base-initial system in Rabha, as it belongs phylogenetically to the Bodo-Garo group of Tibeto-Burman languages, which is spoken by the Rabha people living around the Brahmaputra valley and the Arakan Moutains. Most numeral systems within Bodo-Garo, excluding Rabha, are all base-initial, c.f., Atong, Bodo, Deori, Dimasa, Garo, Kok Borok, Tiwa, and Usoi (Chan 2017).

Table 10: The original numeral system of Rabha (Joseph 2007: 844).

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. p*hala*	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*

There is one obvious borrowed element in the system of Table 10, namely 'one thousand'. The word for 'thousand' *hajar* is most likely borrowed from the Assamese numeral *hezar*. In this indigenous system, 'one thousand' is formed by the base *hajar* followed by the numeral *sa* 'one'. In other words, albeit the same borrowed element for the base of 'thousand', the old system uses it base-initially while the borrowed system uses it base-finally.

Intriguingly, C/Ms in Rabha can also be divided into two groups, depending on their word order, which by and large corresponds to their respective origin. The C/M-final classifiers are mostly borrowed from Assamese and can only be used with base-final numerals borrowed from Assamese. The C/M-initial classifiers, mostly indigenous, can only appear with the three surviving indigenous numbers, which are part of a base-initial system no longer in use. This leads to a divergence in the paradigms of numeral phrases. As shown in (15), the C/M-initial word order is used for the three remaining indigenous numerals, i.e., one,

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

two, and three; however, the C/M-final word order is found with other borrowed numerals such as four and five.

(15) The numeral phrase paradigm in Rabha

a.	*kai sak-sa*	(person C-one)	'one person'
b.	*kai kam-iŋ*	(person C-two)	'two persons'
c.	*kai me-tham*	(person C-three)	'three persons'
d.	*sari-jon kai*	(four-C person)	'four persons'
e.	*pas-jon kai*	(five-C person)	'five persons'

Rabha also allows, not without constraints, classifiers from both systems to be used with numerals from the other system. Borrowed classifiers have to be 'nativized' into base-initial classifiers before they are used with indigenous base-initial numerals. As shown in (16), a speaker of Rabha may use the borrowed classifier *dal* in a C/M-final order when counting with the numeral system borrowed from Assamese for four and above (16a). However, if the speaker counts with the indigenous system for three and beyond, the C/M-initial order has to be used, even if the classifier is not indigenous. Finally, the C/M-initial order is also used if the speaker uses indigenous classifiers with indigenous numerals (16c).

(16) The borrowed classifier *dal* with borrowed and indigenous numerals
[Joseph 2007: 440]

a.	*cari-dal*	*bá*
	four-C	bamboo
	'four bamboos'	
b.	*bá*	*dal-sa*
	bamboo	C-one
	'one bamboo'	
c.	*bá*	*tiŋ-sa*
	bamboo	C-one
	'one bamboo (counting with indigenous classifier)'	

The opposite is observed if indigenous classifiers are combined with numerals borrowed from Assamese. As an example in (17), indigenous classifiers are generally in the C/M-initial order if used with the indigenous numerals between one and three (17a). However, indigenous classifiers are in the C/M-final order if they are connected with base-final numerals borrowed from Assamese (17b); unless they are used as nouns, e.g., in (17c), *doŋ* is considered as a noun, whereas the classifier *-ta* is used instead. Rabha thus provides strong

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

evidence for the base-C/M harmonization, which is not only confirmed cross-linguistically, but also observed language-internally.

(17) Indigenous classifier *doŋ* with indigenous and borrowed numerals [Joseph 2007: 443]

a. *mai doŋ-aniŋ*
paddy C-two
'two ears of paddy'
b. *cari-doŋ mai*
four-C paddy
'four ears of paddy'
c. *cari-ta doŋ mai*
four-C ear paddy
'four ears of paddy'

The observed violations to the probabilistic universal in Tibeto-Burman can therefore 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, for example, is C/M-initial but base-final (Emeneau 1956). It is thus not surprising that, except for 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. *nui̯*	20. *tʃinui̯*	200. *nui̯ ra*
3. *tʰam*	30. *tʰamtʃi*	1000. *hadʒar ʂa*
4. *burui̯*	40. *buruitʃi*	2000. *nui̯ hadʒar*
5. *ba*	50. *batʃi*	
6. *doʊk*	60. *doʊktʃi*	
7. *ʂıni*	70. *ʂınitʃi*	
8. *tʃar*	80. *tʃatʃi*	
9. *ʃıku*	90. *ʂikutʃi*	

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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 5. 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 pressure from the neighboring base-final classifier languages, which are politically, socially, and economically more powerful. Tibeto-Burman languages could have gradually adopted the numeral systems of their neighbors and also acquired their classifier feature, while the more isolated languages in the mountainous areas would have been more protected by the geographic barriers and retained more of their original systems.

6 Conclusion

In this paper, we reviewed the two probabilistic linguistic universals developed from the multiplicative theory that unites numeral bases and classifiers. The 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. Both were tested in the world's foremost hotbed of classifier languages, namely SMATTI (Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai, Tibeto-Burman, Indo-Aryan). 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 few exceptions encountered 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 only selected languages from six specific language groups. 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 including classifier languages and 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. Additional research is likewise needed in this regard.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

Acknowledgments: We thank the anonymous reviewers and the editors for their constructive comments, which led to significant improvements of the paper. All remaining errors are our own. We gratefully acknowledge the financial support by Taiwan's Ministry of Science and Technology (MOST) via the following grants awarded to O.-S Her: 101-2410-H-004-184-MY3, 104-2633-H-004-001, 104-2410-H-004-164-MY3, and 106-2410-H-004-106-MY3.

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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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

References

Adams, Karen L. 1986. Numeral classifiers in Austroasiatic. In Colette G. Craig (ed.), *Noun classes and categorization*, 241–262. Amsterdam & Philadelphia: John Benjamins.

Aikhenvald, Alexandra Y. 2000. *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.

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

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

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

Bauer, Christian. 1992. Review of: Adams Karen Lee: *Systems of numeral classification in the Mon-Khmer, Nicobarese and Aslian subfamilies of Austroasiatic*. 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.

Beckwith, Christopher I. 2007. Phoronyms: *Classifiers, class nouns, and the pseudopartitive construction*. New York: Peter Lang.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

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

Bhattacharya, Tanmoy. 2001. Numeral/quantifier-classifier as a complex head. In Norbert Corver & Henk C. van Riemsdijk (eds.), *Semi-lexical categories: The function of content words and the content of function words*, 191–221. Berlin & New York: Mouton de Gruyter.

Bickel, Balthasar. 2014. Linguistic diversity and universals. In N. J. Enfield, Paul Kockelman & Jack Sidnell (eds.), *The Cambridge handbook of linguistic anthropology*, 101–124. Cambridge: Cambridge University Press.

Bisang, Walter. 1999. Classifiers in East and Southeast Asian languages: Counting and beyond. In Jadranka Gvozdanović (ed.), *Numeral types and changes worldwide*, 113–186. München: Walter de Gruyter.

Biswas, Priyanka. 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.

Burling, Robbins. 2004. *The language of the Modhupur Mandi (Garo)*. Vol. 1: *Grammar*. New Delhi: Bibliophile South Asia.

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

Comrie, Bernard. 2006. Numbers, language, and culture. Presented at the Jyväskylä Summer School, Jyväskylä.

Comrie, Bernard. 2013. Numeral bases. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/131, Accessed on 2018- 07-13.)

Corbett, Greville G. 2013. Number of genders. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/30, Accessed on 2018- 07-13.)

Dayal, Veneeta. 2014. Bangla plural classifiers. *Language and Linguistics* 15(1). 47–87.

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

Emeneau, Murray B. 1956. India as a linguistic area. *Language* 32(1). 3–16.

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

Gil, David. 2013. Numeral classifiers. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/55, Accessed on 2018-07-13.)

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

Greenberg, Joseph H. 1990 [1978]. Generalizations about numeral systems. In Keith Denning & Suzanne Kemmer (eds.), *On language. Selected writings of Joseph H. Greenberg*, 271–309. Stanford, CA: Stanford University Press. [First published 1978 in Joseph H. Greenberg., Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of the human language, vol. 3: Word structure*, 249–295. Stanford, CA: Stanford University Press.]

Gvozdanović, Jadranka. 1999. *Numeral types and changes worldwide*. Berlin & New York: De Gruyter.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

Her, One-Soon. 2012a. Distinguishing classifiers and measure words: A mathematical perspective and implications. *Lingua* 122(14). 1668–1691.
Her, One-Soon. 2012b. Structure of classifiers and measure words: A lexical functional account. *Language and Linguistics* 13. 1211–1251.
Her, One-Soon. 2017a. Deriving classifier word order typology, or Greenberg's Universal 20A and Universal 20. *Linguistics* 55(2). 265–303.
Her, One-Soon. 2017b. Structure of numerals and classifiers in Chinese: Historical and typological perspectives and cross-linguistic implications. *Language and Linguistics* 18(1). 26–71.
Her, One-Soon, Ying-Chun Chen & Nia-Shin Yen. 2017. Mathematical values in the processing of Chinese numeral classifiers and measure words. *PLOS ONE* 12(9). 1–9.
Her, One-Soon, & Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11(3). 527–551.
Hu, Houxuan. 1983. *Jiaguwen heji* [The great collection of the oracle inscriptions]. Beijing: China Social Sciences Publishing House.
Jacquesson, François. 2007. Kok-Borok. A short analysis. *Hukumu, 10th anniversary volume*, 109–122. Agartala: Kokborok Tei Hukumu Mission.
Joseph, Thatil Umbavu V. 2007. *Rabha*. Leiden: Brill.
Kondakov, Alexander. 2013. *A sociolinguistic survey of the Rabha dialects of Meghalaya and Assam*. Dallas: SIL international.
Luraghi, Silvia. 2011. The origin of the Proto-Indo-European gender system: Typological considerations. *Folia Linguistica* 45(2). 435–464.
Matisoff, James A. 1995. Sino-Tibetan numerals and the play of prefixes. *Bulleting of the National Museum of Ethnology (Osaka)* 20(1). 105–251.
Mazaudon, Martine. 2009. Number-building in Tibeto-Burman languages. In Stephen Morey & Mark W. Post (eds.), *North East Indian linguistics*, 117–148. Cambridge: Foundation Books.
Morev, Lev N. 2000. Some afterthoughts on classifiers in the Tai languages. *Mon-Khmer Studies* 30. 75–82.
Mortensen, David R. 2017. Hmong-Mien languages. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*, 1–25. Oxford: Oxford University Press.
Nomoto, Hiroki. 2013. *Number in classifier languages* (PhD dissertation). University of Minnesota, Minneapolis.
Peyraube, Alain. 1991. Some remarks on the history of Chinese classifiers. *Santa Barbara Papers in Linguistics* 3. 106–126.
Peyraube, Alain. 1998. On the history of classifiers in Archaic and Medieval Chinese. In Benjamin K. T'sou (ed.), *Studia Linguistica Serica*, 131–145. Hong Kong: City University of Hong Kong.
Piantadosi, Steven T. & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756.
Seifart, Frank. 2010. Nominal classification. *Language and Linguistics Compass* 4(8). 719–736.
Senft, Gunter. 1986. *Kilivila: The language of the Trobriand Islanders*. Berlin & New York: De Gruyter.
Senft, Gunter. 2000. What do we really know about nominal classification systems. In Gunter Senft (ed.), *Systems of nominal classification*, 11–49. Cambridge: Cambridge University Press.
Simons, Gary F. & Charles D. Fennig (eds.). 2018. *Ethnologue: Languages of the world (21st edition)*. Dallas, Tex.: SIL International. (http://www.ethnologue.com).

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM

Sinnemäki, Kaius T. K. to appear. On the distribution and complexity of gender and numeral classifiers. In Francesca Di Garbo & Bernhard Wälchli (eds.), *Grammatical gender and linguistic complexity*. Berlin: Language Science Press.

Sullivan, Gail M. & Richard Feinn. 2012. Using effect size – or why the p value is not enough. *Journal of Graduate Medical Education* 4(3). 279–282.

Tai, James & Lianqing Wang. 1990. A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association* 25(1). 35–56.

Velupillai, Viveka. 2012. *An introduction to linguistic typology*. Amsterdam & Philadelphia: John Benjamins.

Wu, Fuxiang, Feng Shengli & Huang Zhengde. 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, Dan 2013. *Plurality and classifiers across languages in China*. München: Walter de Gruyter.

Yi, Byeong-Uk. 2009. Chinese classifiers and count nouns. *Journal of Cognitive Science* 10. 209–225.

Yi, Byeong-Uk. 2011. What is a numeral classifier? *Philosophical Analysis* 23. 195–258.

Zhang, Niina Ning. 2013. *Classifier structures in Mandarin Chinese*. Berlin & Boston: De Gruyter Mouton.

Brought to you by | Universite Lumiere Lyon 2
Authenticated
Download Date | 10/3/19 11:32 AM