Numeral base, numeral classifier, and noun
Word order harmonization

Marc Allassonnière-Tang and One-Soon Her
CNRS/University Lyon 2 | National Chengchi University

Greenberg (1990a: 292) suggests that classifiers (CLF) and numeral bases tend to harmonize in word order, i.e. a numeral (Num) with a base-final [*n base*] order appears in a CLF-final [Num CLF] order, e.g. in Mandarin Chinese, *san1-bai3* (three hundred) '300' and *san1 zhi1 gou3* (three CLF $_{\text{animal}}$ dog) 'three dogs', and a base-initial [*base n*] Num appears in a CLF-initial [CLF Num] order, e.g. in Kilivila (Eastern Malayo-Polynesian, Oceanic), *akatu-tolu* (hundred three) '300' and *na-tolu yena* (CLF $_{\text{animal}}$-three fish) 'three fish'. In non-classifier languages, base and noun (N) tend to harmonize in word order. We propose that harmonization between CLF and N should also obtain. A detailed statistical analysis of a geographically and phylogenetically weighted set of 400 languages shows that the harmonization of word order between numeral bases, classifiers, and nouns is statistically highly significant, as only 8.25% (33/400) of the languages display violations, which are mostly located at the meeting points between head-final and head-initial languages, indicating that language contact is the main factor in the violations to the probabilistic universals.

Keywords: numeral, numeral base, classifier, noun, word order, harmonization

1. Introduction

A complex numeral with an internal multiplicative structure of [$n \times base$], e.g. *three hundred*, where *n*, i.e. *three*, is the multiplier and *base*, i.e. *hundred*, is the multiplicand, has two possible orders: [*n base*] and [*base n*], given that multiplication is commutative. Both orders are indeed attested in languages of the world (Her et al. 2019). For instance, in Mandarin Chinese (Sinitic), like English, the decimal base follows the multiplier, e.g. *san1 bai3* (three hundred) 'three hundred'. On the other hand, in Kilivila (Eastern Malayo-Polynesian, Oceanic), the decimal base precedes the multiplier, cf. *akatu-tolu* (hundred three) 'three hundred'. Greenberg (1990a)

https://doi.org/10.1075/lali.00069.all
Language and Linguistics 21:4 (2020), pp. 511–556. issn 1606-822x | e-issn 2309-5067 © ILAS

proposes fifty-four generalizations about numeral systems in languages. This present study is particularly concerned about Generalization 28:

> If there are any numerals in which the expression of the multiplier follows that of the multiplicand, the language is one in which the numeral follows the noun.
> (Greenberg 1990a: 292)

Greenberg's claim is thus that a [*base n*] numeral (Num) should come after the noun (N) it quantifies, i.e. [N Num], while [*n base*] and [Num N] are the default word orders. Such a hypothesis can be seen as a case of the head-parameter, i.e. head-initial numerals (i.e. [*base* n] numerals) appear in head-initial nominal constituents (i.e. [N Num]), with base and N as syntactic heads in their respective constituent, and the same principle applies to head-final numerals, thus [*n base*] in [Num N]. This is in essence also what Greenberg suggests. Note that M and U in the following quote from Greenberg refer to "multiplier" and "unit" respectively; thus, MU is [*n base*], and UM is [*base n*].

> Since, as we have seen, the most common syntactic treatment of multiplication is to equate it with the QN (quantifier-noun) construction, i.e. *three tens* like *three houses* and *tens three* like *knives three* in most languages, the two orders harmonize, MU with QN and UM with NQ. (Greenberg 1990a: 292)

However, some languages require an additional element known as (numeral) classifier[1] (CLF) in the QN (quantifier-noun) construction mentioned by Greenberg. In Mandarin Chinese, for example, the expression for *three houses* is *san1 ge0 fang2zi0* (three CLF $_{\text{general}}$ house). For such classifier languages, the following provision to Generalization 28 is proposed:

> Where there are numeral classifiers, it is the order numeral + classifier that is fundamental and conforms to this generalization and not classifier phrase + noun.
> (Greenberg 1990a: 292)

Thus, a classifier language will have the [CLF Num] order in harmony with the order of [*base n*] in complex numerals; otherwise, [Num CLF] and [*n* base] obtain. If this turns out to be valid, it indicates that CLF is the head, and Num, the modifier. In light of the head-parameter, assuming that N is the head of a nominal construction, we further propose that the order between N and Num is harmonized with the order between CLF and Num. Hence, we can restate Generalization 28 as (1a–b) and also expand it to include (1c).

1. A more detailed characterization of (numeral) classifiers is provided in § 2 and § 3.2.

(1) Greenberg's Generalization 28 reformulated and expanded
   a. **Base-N harmonization**:
   If Num is base-initial, i.e. [*base n*], then an N-initial order obtains between N and Num; otherwise, an N-final order obtains.
   b. **Base-CLF harmonization**:
   If Num is base-initial, i.e. [*base n*], then a CLF-initial order obtains between CLF and Num; otherwise, a CLF-final order obtains.
   c. **CLF-N harmonization**:
   If a CLF-initial order obtains between CLF and Num, then an N-initial order obtains between N and CLF; otherwise, an N-final order obtains.

The aim of this paper is to consider these three word order generalizations as probabilistic universals (Dryer 1998; Velupillai 2012: 31). In other words, we expect to observe the statistically significant tendency of harmonization between numeral base, CLF, and N within the sample of languages we extracted from the languages of the world. We refer to such harmonization as a tendency rather than a strict universal, since empirical evidence on the fundamental diversity of languages does not support the notion of absolute universals (Evans & Levinson 2009). The paper is organized as follows. § 2 first considers a functional motivation behind the proposed universals involving numeral classifiers within a multiplicative theory. § 3 presents the data source of the 400 languages used in this study. § 4 then applies statistical methods to test the three word order generalizations in (1). § 5 discusses the results of the statistical analyses and examines some of the cases of violation. § 6 concludes the paper with a summary of the study, its limitations, and future prospects.[2]

2. Literature review

In classifier languages, a numeral classifier may be required when a numeral is employed in the quantification of a noun, i.e. "numeral classifiers occur within "pseudopartitive" constructions, which consist of a specifier (numeral, quantifier or determiner), classifier and noun" (Kilarski 2014: 33–34). Such classifiers come in two varieties: sortal classifiers apply to count nouns (2a) and mensural classifiers may apply to count nouns (2b) and/or mass nouns (2c). Sortal classifiers and mensural classifiers are thus two subcategories of a single syntactic category, referred to as (numeral) classifier, CLF in short, and have the same syntactic structure. Yet, note that mensural classifiers, *aka* measure words, are not the same

2. The detailed data and code can be found at the GitHub repository of the authors' https://github.com/marctang/word-order-harmonization.

structure as what may be commonly observed in English with expressions such as *three cups of tea*, which can be referred to as measure terms. Measure terms are considered as lexical means of categorization since "their choice is neither paradigmatic nor obligatory" (Kilarski 2013: 9). Mensural classifiers are part of the grammaticalized categorization of nouns (Aikhenvald 2000: 116–120; Grinevald 2000: 58–59). More importantly, their syntactic behavior is very different: the English measure terms are nouns since they take plural marking and require the preposition "of", e.g. *three bottles of wine*. Mensural classifiers such as (2a) and (2b) in Chinese are not nouns as they do not take plural marking and directly precede N (Her & Hsieh 2010; Her 2012).

(2) Examples of sortal and mensural classifiers in Mandarin Chinese
   a. *wu3 ben3 shu1*
      five CLF $_{volume}$ book
      'five books'
   b. *wu3 xiang1 shu1*
      five MENS $_{box}$ book
      'five boxes of books'
   c. *wu3 xiang1 tu3*
      five MENS $_{box}$ soil
      'five boxes of soil'

In previous studies, order harmonization between numeral bases and classifiers has been interpreted as a manifestation of a cognitive functional connection between two formally distinct elements: numeral bases and classifiers are both multiplicand; thus, they behave in a similar way with regard to their respective position with the multiplier (Her 2017a: 280; 2017b: 42–43). For instance, in (3a), if the numeral base (e.g. hundred) is located after the multiplicand (e.g. three), it is base-final. In such a situation, the classifier (which is also the multiplicand) tends to be positioned after the numeral, i.e. *san1-bai3 tiao2* (three-hundred CLF $_{long}$). The opposite order occurs if the numeral base is base-initial (3b).

(3) Example of the harmonization between numeral bases and classifiers
   a. *san1-bai3 tiao2 yu2* (base-CLF-final, Mandarin Chinese)
      three-hundred CLF $_{long}$ fish
      '300 fish'
   b. *na-akatu-tolu yena* (base-CLF-initial, Kilivila)
      CLF $_{animal}$-hundred-three fish
      '300 fish'

Greenberg (1990b: 172) is again among the first to view the relation between numerals and sortal classifiers as multiplication. More accurately, Greenberg considers all sortal classifiers to be a multiplicand with the precise value of *one*. By way

of illustration, in Mandarin Chinese, *san1 zhi1 gou3* (three CLF $_{\text{animal}}$ dog) 'three dogs' may be interpreted as (3×1 dog), in which the sortal classifier carries the mathematical value of one along with the semantic feature of animals. This view has been further developed in recent studies (Au Yeung 2005, 2007; Her 2012)[3] where the distinction between sortal and mensural classifiers is characterized in terms of their mathematical values as a multiplicand, i.e. the value of a sortal classifier is necessarily *one*, but the value of a mensural classifier is not. As an example, in Mandarin Chinese, *wu3 da3 shu1* (five MENS $_{\text{dozen}}$ book) 'five dozen of books' equals to (5×12 book). Under this multiplicative view, a classifier and a numeral base both function as multiplicands. Thus, given Num formed by an *n* and a base, the base must be aligned in word order with CLF and the two must not be interrupted, thus creating the effect of base-classifier harmonization (Her 2017a: 292). This view also nicely accounts for the following word order typology: Out of the six possible orders among Num, CLF, and N in (5), only four orders are attested in languages in the world (Greenberg 1990b: 185; Aikhenvald 2000: 104–105).

(4) Six mathematically possible word orders of [Num, CLF, N] in classifier languages
   a. √ [Num CLF N] (many languages, e.g. Mandarin Chinese)
   b. √ [N Num CLF] (many languages, e.g. Thai)
   c. √ [CLF Num N] (few languages, e.g. Ibibio [Niger-Congo])
   d. √ [N CLF Num] (few languages, e.g. Jingpho [Tibeto-Burman])
   e. * [CLF N Num] (no language)
   f. * [Num N CLF] (no language)

Why do languages allow only these four orders? Given a multiplicative numeral formed by *n*, and *base*, along with a classifier, Generalization 28 allows only two possibilities: [CLF [*base n*]$_{\text{Num}}$] and [[*n base*]$_{\text{Num}}$ CLF]. As shown in (5), the four attested orders in (4) are precisely the ones that observe this harmonization, with N appearing on either edge of the constituent. Generalization 28 thus predicts that CLF and the base of a numeral formed by *n* and *base* must be adjacent. Thus, (5a'), for example, is ill-formed, because the base and the classifier are separated by the multiplier (*n*) of the numeral.

(5) Twelve possible word orders of [[*n, base*], CLF, N]
   a. √ [[*n* ***base***] **CLF** N] (base-CLF harmonization)
   a.' * [[*base n*] CLF N]
   b. √ [N [*n* ***base***] **CLF**] (base-CLF harmonization)
   b.' * [N [*base n*] CLF]

3. Note that while Her (2012) explicitly gives Greenberg (1990b) the due credit for this multiplicative view of classifiers, Au Yeung (2005; 2007) seems to be unaware of this fact.

c. √ [**CLF** [***base*** *n*] N] (base-CLF harmonization)
c.' * [CLF [*n base*] N]
d. √ [N **CLF** [***base*** *n*]] (base-CLF harmonization)
d.' * [N CLF [*n base*]]
e. * [CLF N [***base*** *n*]]
e.' * [CLF N [*n* ***base***]]
f. * [[*n* ***base***] N CLF]
f.' * [[***base*** *n*] N CLF]

(Her 2017a: 292)

Her et al. (2019) investigated six specific groups of classifier languages, Sinitic, Miao-Yao, Austro-Asiatic, Tai-Kadai,[4] Tibeto-Burman, Indo-Aryan, and demonstrated a statistically significant harmonization in terms of word order between the numeral bases and classifiers. However, their approach was restricted to the use of a sample of languages from six specific language groups, which is thus not phylogenetically weighted. In this paper, we aim to fill this methodological gap with a phylogenetically and geographically weighted language sample that includes both classifier languages and non-classifier languages.

Furthermore, given our knowledge of genuine exceptions, this study hypothesizes that the two word-order statements in Generalization 28 are probabilistic universals, not absolute universals. Probabilistic universals are based on statistical analysis of an observation which "hold[s] for most, but not all, languages", as opposed to absolute universals, which allow no exceptions (Dryer 1998; Velupillai 2012: 31). There are other more fundamental reasons too. First, using Bayesian and frequentist statistical methods to justify inviolable patterns cross-linguistically would need an unrealistic amount of data (Evans & Levinson 2009; Piantadosi & Gibson 2014: 736), and even all current languages of the world would not be sufficient, since "we can never know that there is not another language that fails to confirm to the universal, either one that was once spoken or a hypothetical language that is possible but never actually spoken due to historical accident" (Dryer 1998). Furthermore, language change is constant and generally results in an intermediary stage of language structure. A putative absolute universal which requires a binary or an atomic judgment can thus be expected to encounter exceptions. Therefore, we opt to apply "the same theory-hypothesis-statistics triangle that characterizes most sciences" (Bickel 2014: 119) to falsify the null hypothesis of no association through statistical methods, as opposed to the alternative hypothesis which represents the association between the word orders of numeral bases, classifiers, and noun phrases.

4. Tai-Kadai has also been referred to as Kra-Dai (Ostapirat 2000; 2005). We use "Tai-Kadai" in our paper to facilitate cross-checking with existing databases.

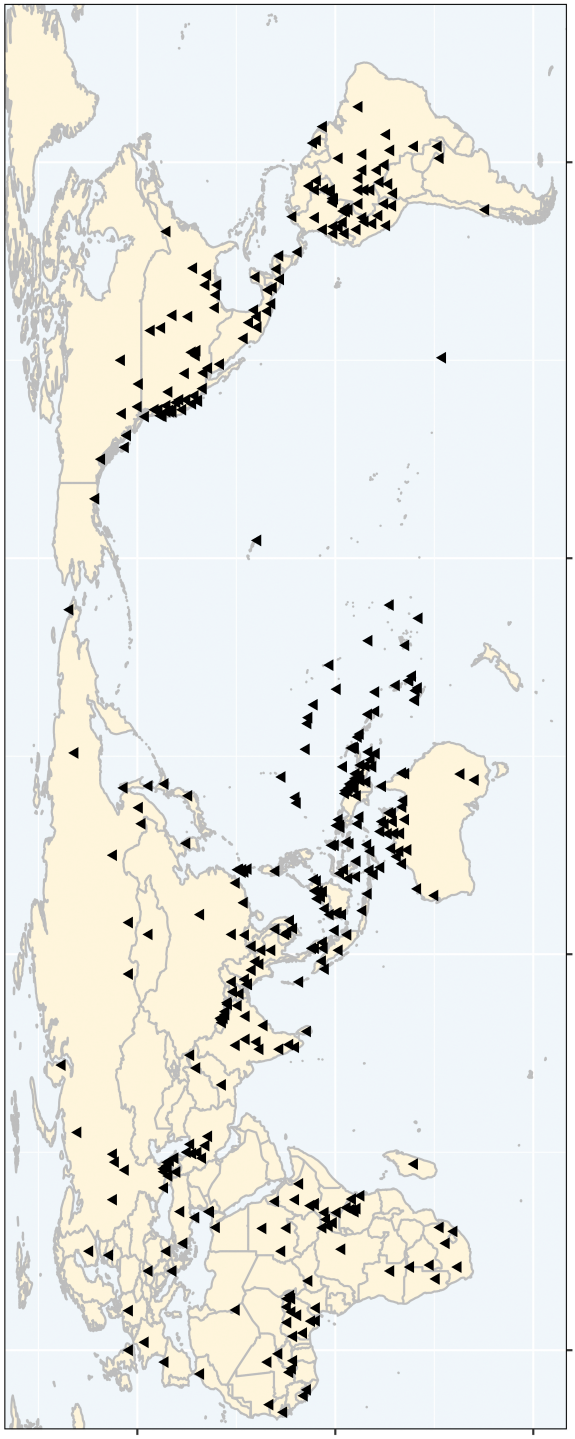
3. Methodology

This sample contains the 400 languages in Gil (2013)'s study on numeral classifier languages. It is defined as weighted in the sense that most language families widely accepted by specialists are represented. Moreover, linguistic diversity is also taken into consideration, e.g. since the Austronesian family accounts for nearly 17.14% (1,262/7,363) languages of the world (Lewis 2009), a similar ratio is applied in the dataset (19.00%, 76/400). The same logic is enforced with regard to spatial distribution. For instance, since the Pacific region contains 18.74% (1,380/7,363) languages of the world (Lewis 2009), an equal scale is illustrated in the database (18.50%, 74/400). Such diversification is also represented within the subgroups of each language family and smaller regions in terms of geography. For further details, please refer to *The world atlas of language structures*. The languages included in our study are displayed in Map 1.

We acknowledge that such data are not an absolute representative of languages of the world and probably suffers from Galton's problem, e.g. additional phylogenetic and geographical testing could ensure a better-balanced list of languages. Nevertheless, we estimate that it is appropriate for the purpose at hand, which is a preliminary analysis on the probabilistic universals proposed. Our data sources can be summarized as follow. With regard to numeral systems, we primarily rely on Chan (2018).[5] Information on classifiers is mainly from Gil (2013), while word orders are largely obtained from Dryer (2013). Note that even though initially we annotated languages according to these databases, we conducted a cross-check of every language in the dataset by verification through language grammars. Thus, a number of points in the data have been modified after a judgment based on a comparison of different studies on the same language. The main reason behind such verification is that we may use different definitions from past researchers of the analyzed categories. Hence, our interpretation of one source of data can diverge from the annotations of other researchers.

Furthermore, we acknowledge a binary classification of languages in our dataset as either classifier languages or non-classifier languages, though some recent studies on nominal classification weigh languages according to their level of canonicity (i.e. prototypicality) with regard to classifier systems (Corbett 2003a; Grinevald 2015; Corbett & Fedden 2016; Fedden & Corbett 2017). Given the purpose of a preliminary analysis, the binary classification is sufficient, as

5. Detailed page numbers from Chan (2018) 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.



Map 1. The 400 languages included in the current analysis

another project applying the canonical approach is in progress to provide additional information related to the research question of this study.

3.1 Numeral base

As mentioned in § 2, we define numeral bases according to the following formula of numeral composition: $(n \times base)+m$, where $m < base$ (Comrie 2013). We gathered data on the numeral systems of the languages in the dataset to cross-check their respective order between the multiplier, *n*, and the multiplicand, *base*. In Table 1, numeral bases in Makassar (Western Malayo-Polynesian) consistently follow the multiplier, e.g. within the numeral 20, the decimal *ten* [sampulo] is positioned after the multiplier *two* [rua], and merge to form [ruampulo]. Makassar is thus annotated as a base-final language.

Table 1. The numeral system of Makassar (Chan 2018)

1. seʔre	10. sampulo	100. sibilaŋŋaŋ
2. rua	20. ruampulo	200. ruambilaŋŋaŋ
3. tallu	30. tallumpulo	1,000. sisaʔbu
4. appaʔ	40. patampulo	2,000. ruassaʔbu
5. lima	50. limampulo	
6. annaŋ	60. annampulo	
7. tuɟu	70. tuɟupulo	
8. sagantuɟu (7+1)	80. sagantuɟupulo	
9. salapaŋ (10-1)	90. salapaŋpulo	

Yet, not all languages employ such a transparent system. In Ngada (Central Malayo-Polynesian), for instance, the decimals are consistently base-initial, e.g. as shown in Table 2, 30 [bulu təlu] is literally composed of 10 [səbulu] and 3 [təlu]. Even though some phonological changes occur (i.e. [səbulu] changes to [bulu] in decimals), the general base-initial order is rather straightforward, except the thousands, however, which are base-final. While 200 [ŋasu zua] is base-initial as 100 [ŋasu]×2 [zua], 2000 [zua ribu] is base-final, i.e. 2 [zua] × 1,000 [ribu]. Such cases are annotated as base-initial since the majority of the multiplicative numerals are base-initial.

Another type of problematic data includes loan words and the use of foreign numeral systems. Belhare (Sino-Tibetan), for example, only retains three traditional numerals (*i* 'one', *sik* 'two', and *sum* 'three), while higher numbers are actually loans from Nepali (Chan 2018). In such cases, we consider the loan numerals

Table 2. The numeral system of Ngada (Chan 2018)

1. ʔəsa	10. səbulu	100. sə ŋasu
2. zua	20. bulu zua	200. ŋasu zua
3. təlu	30. bulu təlu	1,000. sə ribu
4. vutu	40. bulu vutu	2,000. zua ribu
5. lima	50. bulu lima	
6. lima əsa	60. bulu lima əsa	
7. lima zua	70. bulu lima zua	
8. zua butu	80. bulu zua butu	
9. tarəsa	90. bulu taraəsa	

as part of the language, and thus do not annotate Belhare as a language without numeral bases, i.e. deprived of a multiplicative system. There are, however, several languages in our dataset that are genuinely without a multiplicative system. For instance, most, if not all, numeral systems in Australian languages do not involve multiplication (Epps et al. 2012: 51). As shown in Table 3, the Australian language Mawng has only two basic number words for one and two and employs only addition, e.g. *ngarrkarrk la y-arakap* (two and one) 'three' is a combination of 2+1, and *ngarrkarrk la ngarrkarrk* (two and two) 'four' is formed by 2+2. *wurrkamaj yurnu* 'five' literally means 'one side of the hand', while the highest numeral *wurrkamaj yurnu la ngarrkarrk la ngarrkarrk* 'nine' reflects the same principle of addition with its structure as 5+2+2.

Table 3. The numeral system in Mawng (Chan 2018)

1. *-arakap*
2. *ngarrkarrk*
3. *ngarrkarrk la y-arakap*
4. *ngarrkarrk la ngarrkarrk*
5. *wurrkamaj yurnu*
9. *wurrkamaj yurnu la ngarrkarrk la ngarrkarrk*

An overview of the spatial distribution of numeral bases within the 400 languages of our dataset is shown in Map 2. Base-final languages such as Makassar are shown in circles, whereas base-initial languages (e.g. Ngada) are represented by squares. Those without a multiplicative system (e.g. Mawng) are drawn as triangles. Our findings generally correlate with previous studies, as the majority of languages possess a multiplicative numeral system, and those without a multi-

plicative system are mostly located in Australia and Amazonia, with a few in Africa (Comrie 2013). The detailed numbers and their distribution are listed in § 4.

To summarize, information on numeral bases was generally retrieved with relatively few cases of disagreement within the data. Minor divergences in the counting systems have also been noted and taken into account. A similar methodology is applied for numeral classifiers.

3.2 Numeral classifier

With comparison to numeral bases, classifiers are described in a much more opaque manner in the literature, as they "go by an exasperating variety of names" (Blust 2009: 292), e.g. individual classifiers, numeral classifiers, words of measure, quantifiers, unit words, numeratives, projectives, among others. Thus, while applying the definition given in § 2, we need to verify whether a language does or does not have genuine classifiers, as considerable confusion exists in the literature and elements labelled as classifiers in different languages could be incomparable. For example, Japanese classifiers are commonly attested in the literature and classified in Gil (2013) as obligatory. As shown in (6b) and (6d), the absence of classifiers within the context of enumeration results in ungrammaticality. Moreover, the dominant word order in Japanese is [Num CLF N] (Yamamoto & Keil 2000), as in (1a) and (1c).

(6) Classifiers in Japanese

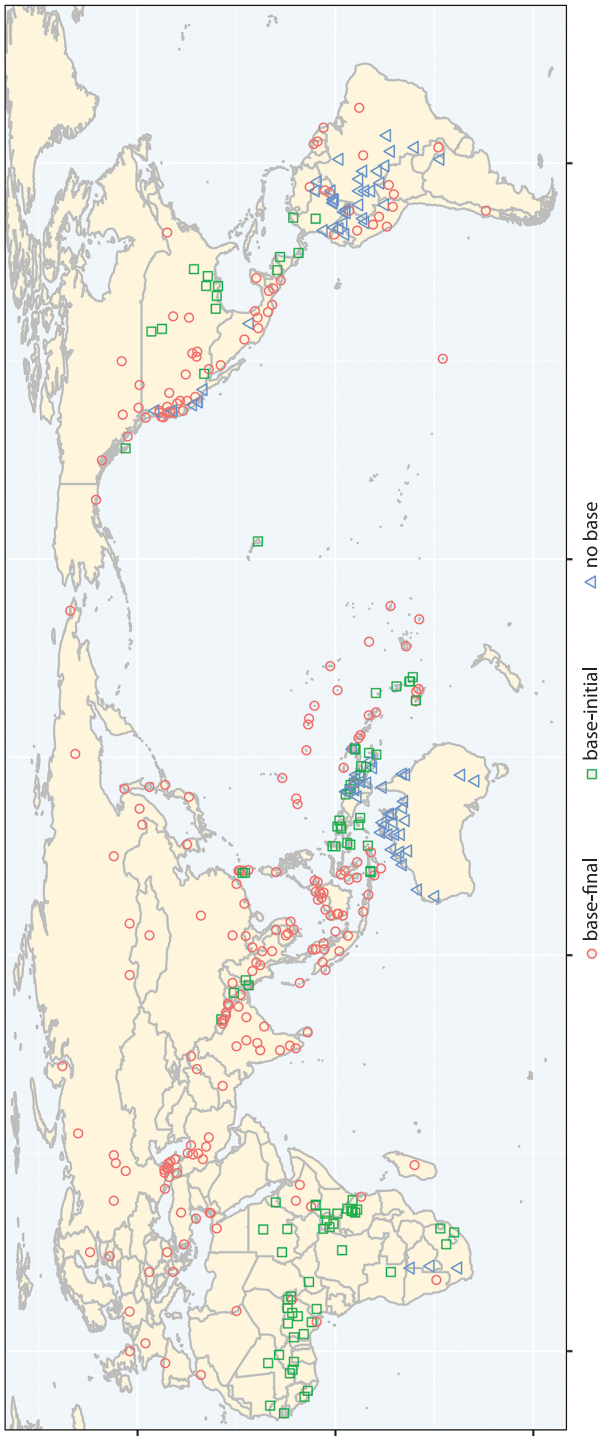
a. *ni-hiki-no* *inu*
two-CLF $_{\text{animal}}$-GEN dog
'two dogs'

b. * *ni-inu*
two-dog
'two dogs'

c. *ni-dai-no* *kuruma*
two-CLF $_{\text{mech}}$-GEN car
'two cars'

d. * *ni-kuruma*
two-car
'two cars' (Mano 2012: 620)

There are cases where different sources disagree. For example, several South-American languages from the Huitotoan and Tucanoan families are annotated as classifier languages by Gil (2013), e.g. Waura (Arawakan), Tuyuca (Tucanoan), Tucano (Tucanoan), Ocaina (Huitotoan), Siona (Tucanoan), Orejon (Tucanoan), and Miraña (Huitotoan); yet, a closer look of their respective language grammars provided evidence that these languages have a concordial nominal classification



Map 2. Spatial distribution of numeral bases in the 400 languages of our dataset

system which is more likely to be defined as grammatical gender rather than classifier (Derbyshire & Payne 1990: 256). Hence, they are reassigned the value "absent" with regard to classifiers. For example, in (7), the Miraña general class marker (GCM) is present on the noun, numeral, and verb. Treating these morphemes as classifiers is inappropriate; as Seifart (2005: 13) points out: "since the use of class markers in expressions such as numerals, demonstratives, and verbs follows a uniform pattern that shares many characteristics with "canonical agreement" (Corbett 2003a; 2003b; 2003c). Thus, the use of class markers in these expressions has little in common with numeral classifiers".

(7) Classifiers in Miraña
a. *tsa-:pi* *gwa-hpi*
one-GCM.MASC.SG human-GCM.MASC.SG
'one man'
b. *kátú:βɛ-bɛ* *gwa-hpi*
fall-GCM.MASC.SG human-GCM.MASC.SG
'he fell, the man' (Seifart 2005: 158)

Similar cases are found in Austronesian and neighbouring languages. For instance, classifiers are attested in Malayo-Polynesian languages (Klamer 2014: 111) and Tuvaluan is marked as classifier language in Gil (2013). However, (Besnier 2002 [2000]: 367) points out that even though certain Tuvaluan inflectional and derivational morphemes, e.g. collective morphemes, numeral modifiers, some quantifiers, among others, may resemble classifiers, "their use is restricted to certain parts of the lexicon and is not intrinsically linked to enumeration, and hence they should not be considered as classifiers in the usual sense of the term". Tuvaluan is considered a non-classifier language in our database.

On the other hand, there are several languages that are labelled as non-classifier languages in Gil (2013), correctly in our view, in spite of existing works claiming otherwise. Bulgarian (Indo-European), for example, is considered to have a marginal classifier system by (Cinque & Krapova 2007). However, as shown in (8), the putative classifier takes grammatical number and engages in number agreement with the verb. Given the mutual exclusiveness between classifiers and plural markers (Borer 2005: 93), languages with such structures are marked as non-classifier languages in our dataset.

(8) Classifier-like structure in Bulgarian
samo dyama dúši *novi* *studenti* *doidoxa*
only two person.PL new.PL student.PL come.PST.3PL
'only two new students came' (Cinque & Krapova 2007: 47)

Another example involves Hindi (Indo-European), which is regarded in some literature as a classifier language, e.g. Toyota (2009: 125) states, "Hindi as well as other Indo-Aryan languages also uses classifiers, such as 'a cup of', 'two cups of', etc., but the nouns involved in the phrase are all mass nouns." Yet, even though the pattern of grammatical number behaves similarly to mensural classifiers, only sporadic examples are found. For instance in (9a), the singular form of 'rupee' *rupayā* is used in presence of a numeral instead of the plural form *rupaye* 'rupees'. Nonetheless, CLF is not present between Num and N. However, in (9b), *pyālā* 'cup' is referred to via the singular form *pyālā* rather than the plural *pyāle* 'cups'. The individual case of *pyālā* fits the definition of mensural classifiers in this paper. Nonetheless, their use is restricted as in Tuvuluan, we therefore do not label Hindi as a classifier language in the dataset.

(9) Classifier-like structures in Hindi
   a. *tīn rupayā*
      three rupee
      'three rupees'
   b. *tīn pyālā cāy*
      three cup tea
      'three cups of tea' (Toyota 2009: 125)

Similar data was provided with regard to German and Russian (Sussex & Cubberley 2009: 314–315), e.g. in German, structures such as *fünf Stück Brötchen* (five piece bread) 'five bread rolls' were occasionally considered as classifiers since *Stück* would not take plural marker. Nevertheless, examples of this type are restricted to specific nouns and contexts. Thus, we did not consider languages with such observation as classifier languages.

The opposite divergence also occurred, i.e. several languages are considered classifier languages in Gil (2013), correctly in our view, in spite of existing works claiming otherwise. For example, Kham (Sino-Tibetan), according to Watters (2002: 180) does not have "true classifiers in the classical sense". We still annotate Kham as a classifier language since the language examples fit our definition of classifiers. As shown in (10), the general classifier *-bu* can be used with both animate and inanimate nouns.

(10) Classifiers in Kham
   a. *tu-bu mi*
      one-CLF $_{general}$ person
      'one person'
   b. *tu-bu zihm*
      one-CLF $_{general}$ house
      'one house' (adapted from Watters 2002: 180)

An overview of the spatial distribution of classifier languages within the 400 languages of our dataset is shown in Map 3. CLF-final languages such as Mandarin Chinese are marked with circles, whereas CLF-initial languages (e.g. Kilivila) are indicated by squares. Finally, languages without classifiers (e.g. Mawng) are shown by triangles. Our findings concord with previous studies, i.e. the majority of classifier languages are in South and East Asia, whereas sporadic cases are observed in parts of Europe, Africa and the Americas (Gil 2013). The detailed numbers of their distribution are listed in § 4.

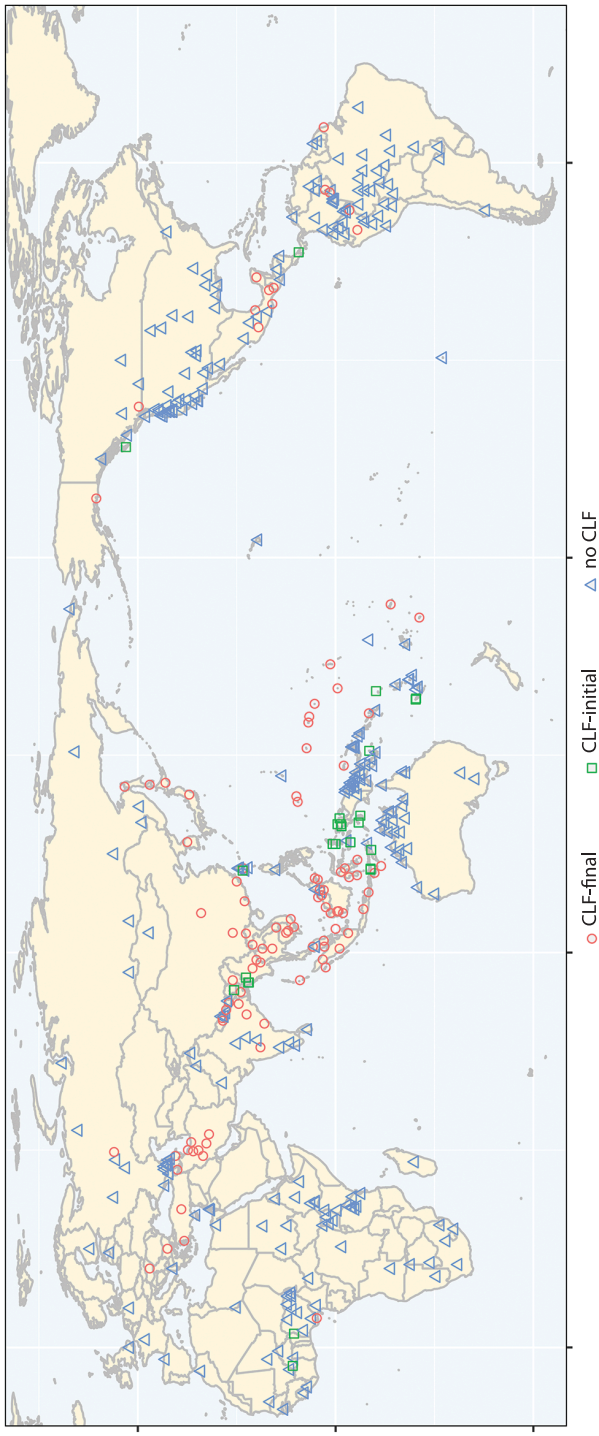
To summarize, classifiers displayed much more divergence in the literature compared to that of numeral bases. Careful examination of actual examples is required to verify the existence or absence of classifiers in each language of the dataset. Different types of situations were encountered, e.g. several languages were recorded as classifier languages while our investigation has shown the contrary, and vice-versa, based on the definition of classifiers applied in this paper. When necessary, linguists working on the language in question were consulted. Our findings thus also highlight the importance of theoretical definition within the process of database-building.

3.3 Order of the numeral and the noun

With regard to the word order of numerals and nouns, we apply Dryer's (2013) methodology and consider "the order of cardinal numerals with respect to a noun they modify". French (Indo-European), for example, has Num consistently positioned before N, as in (11), and is hence annotated as a N-final language.

(11) Numeral-Noun word order in French
  a. *deux livres*
     two book.PL
     'two books'
  b. *trois tables*
     three table.PL
     'three tables'

Divergence among previous studies also exists. For instance, Chimariko (Hokan) is classified as N-initial by Dryer (2013); however, our investigation revealed that both N-initial and N-final orders are possible, as "numerals occur together with nouns in noun phrases, either preceding or following the noun", as shown in (12a) and (12b), respectively (Jany 2009: 58). Yet, since the N-initial order is described as more common, we maintain the annotation from Dryer (2013).



Map 3. Spatial distribution of classifier languages in the 400 languages of our dataset

(12) Word order of numeral and noun in Chimariko
a. *č'imar xotai h-eṭahe-sku-t uwa-tku-t*
man three 3-run.away-DIR.ASP go-DIR-ASP
'three men came as fugitives'
b. *ya-x-amam-na-n p'un ʔiṭi-lla ʔuleeda h-imam-da*
1PL.A-NEG-see-NEG-ASP one man-DIM sibling 3-see-ASP
'we did not see it, a boy saw it' (Jany 2009: 58)

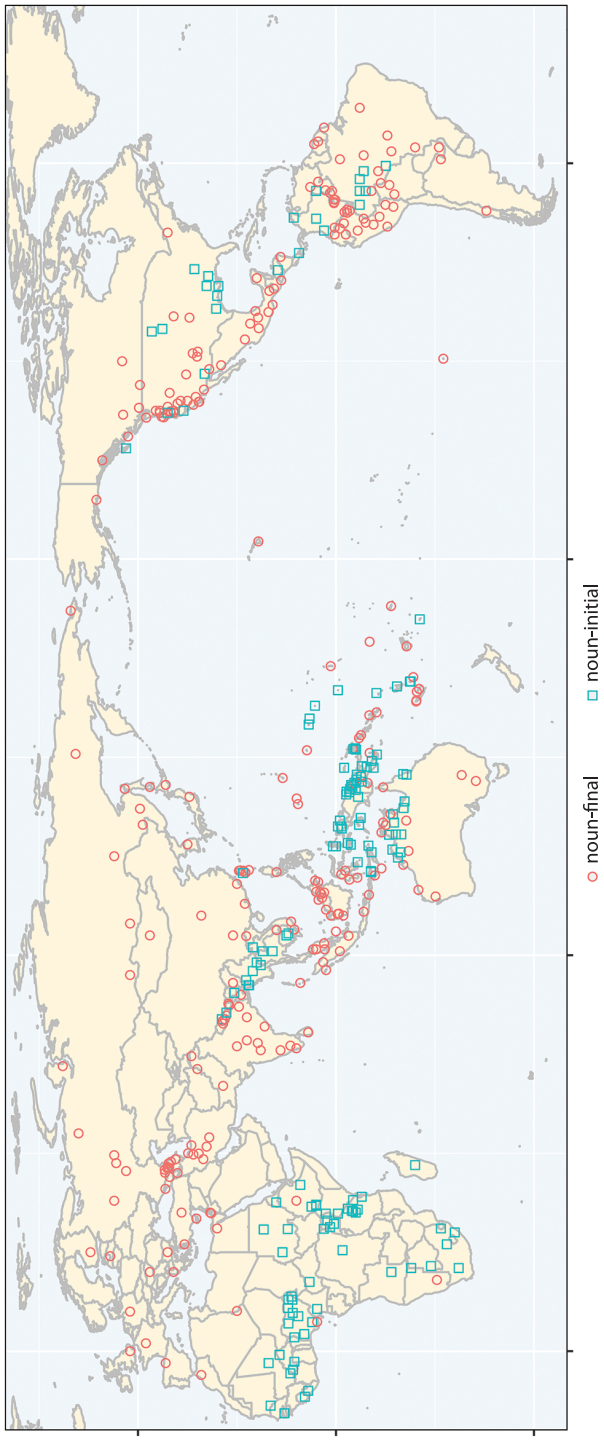
Several other languages had their annotations changed. Arhuaco (Chibchan), for example, is attested as not having a dominant order between N and Num (Dryer 2013), since the order of Num with regard to N serves to mark definiteness (Frank 1985: 41). Note, however, even if both word orders are possible, "quantifiers and numerals usually follow the head noun" (Frank 1985: 5). Hence, we view Arhuaco as a language with an N-initial order, as shown in (13).

(13) Word order of numeral and noun in Arhuaco
a. *peri mouga*
dog two
'two dogs'
b. *tsinu in'gui zei*
pig one GEN
'someone's pig' (Frank 1985: 5–6)

Some languages not included in Dryer (2013) were also difficult to interpret in terms of word order. In Ngarinman (Australian), for example, "numerals are found preceding and following the head equally" (Meakins & Nordlinger 2014: 104), as shown in (14). Yet, given the observation that (14a) "may reflect a 'younger' variety" (Meakins & Nordlinger 2014: 106) compared to the traditional (14b), we annotate the language with N-initial as the dominant pattern.

(14) Word order of numeral and noun in Ngarinman
a. *jindagu girri-nggu yuwa-ni junggard-ngarna tebel-da*
one woman-ERG put-PST smoke-ASSOC table-LOC
'one woman put the packet of cigarettes on the chair'
b. *nyila=wula=nyunu baya-la warlagu-lu gujarra-lu*
that=3UA.S=RR bite-PRS dog-ERG two-ERG
'those two dogs are fighting each other' (Meakins & Nordlinger 2014: 106)

An overview of the spatial distribution of the word order between numerals and nouns within the 400 languages of our dataset is shown in Map 4. N-final languages are marked as circles, whereas N-initial languages are shown in squares. Our findings largely match the observations of previous studies, as the two orders show geographical patterns (Dryer 2013). N-initial languages are outnumbered



Map 4. Spatial distribution of numeral & noun order in the 400 languages of our dataset

by N-final languages and are mostly found in sub-Saharan African, North-East of India along with adjacent areas, and the eastern islands of New Guinea plus northern parts of Australia. N-final languages cover the other parts of the world, including (but not limited to) Europe, South Asia, and most of the Americas.

To summarize, divergence with data from past research was encountered during our survey. Grammars were consulted and examples were scrutinized to judge whether the annotation in previous studies should be modified or maintained to concord with the definition applied in this paper. As a result, we were able to retrieve information on numeral bases, classifiers, and word order between Num and N for the 400 languages targeted. The next section demonstrates the macro-analysis based on our dataset.

4. Statistical analysis

In this section, we first scrutinize the three statistical universals proposed in (1), i.e. the base-N harmonization (§ 4.1), base-CLF harmonization (§ 4.2), and CLF-N harmonization (§ 4.3). In general, we follow the methodology suggested by Levshina (2015) and carry out the calculations using R (R Core Team 2020). The overall distribution of the features is displayed via bar plots to show the general tendencies. A calculation of probability and effect size are then performed by the Chi-square test of independence and Cramer's V.

In § 4.4, we combine the three features, i.e. order of numeral bases, CLF, and nouns, to verify the effect of language family with regard to the probabilistic universal. In other words, we check if a specific language group behaves differently from the others with regard to the proposed probabilistic universal. Likewise, we also investigate the probability of language contact through multidimensional scaling and the application of the Gower distance combined with a Mantel test.

4.1 Numeral base and order between the numeral and the noun

We first examine the proposed statistical universal (1a), i.e. base-N harmonization: If Num is base-initial, i.e. [*base n*], then an N-initial order obtains between N and Num; otherwise, an N-final order obtains. The null hypothesis is that there is no correlation between the base-order and N-order.

Among the 400 targeted languages, the majority of the languages' numeral systems employ multiplication (81.25%, 321/400) and most are base-final (56.50%, 226/400). Such tendencies confirm the observations by previous studies, e.g. Comrie (2013); Her (2017a).

The raw numbers are displayed in Table 4. Besides the base-final tendency, we observe that the N-final order is preferred over the N-initial order: 63.00% (252/400) and 37.00% (148/400), respectively. Incidentally, the label "no dominant order" is not included in Table 4 since no such languages were attested in our dataset.

Table 4. Distribution of base-order and CLF-order in the 400 languages

	N-final	N-initial
base-final	201 (50.25%)	25 (6.25%)
base-initial	7 (1.75%)	88 (22.00%)
base-absent	44 (11.00%)	35 (8.75%)

We visualize the numbers of Table 4 with a bar plot in Figure 1. The x-axis indicates the order of N with respect to Num, while the y-axis represents the frequency of the languages with different base systems. Such distribution shows that base-final structures tend to co-occur with the N-final word order, and vice-versa. Nevertheless, there are a fair number of languages with N-initial and base-final orders. Further analysis is thus required to decide whether (1a) is statistically significant or not.

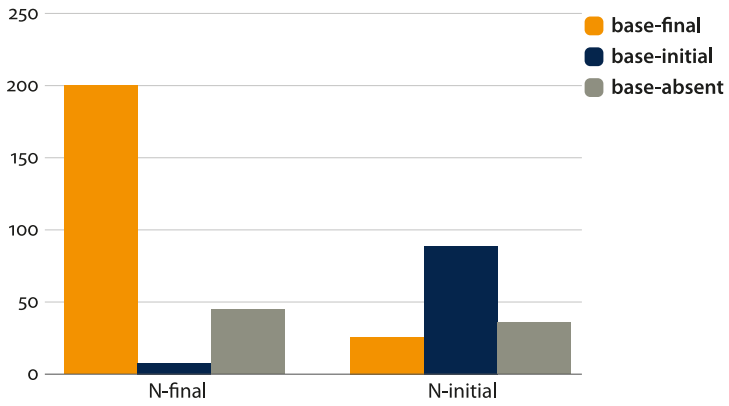


Figure 1. Bar plot of numeral base and noun order

The probability of our observation is obtained via the Pearson's Chi-square (χ^2) test of independence. The null hypothesis suggests no association between the variables, while the alternative hypothesis of (1a) states that the variables (i.e. base-order and N-order) are correlated. In statistical terms, we cannot prove that the alternative hypothesis is right, but we reject the null hypothesis based on the output of the Chi-square test. Such a test compares the observed and expected fre-

quencies of a dataset. The frequency of observations assumed by the null hypothesis is shown in Table 5.

Table 5. Expected random distribution of numeral base and noun

	N-final	N-initial
base-final	142 (35.50%)	84 (21.00%)
base-initial	60 (15.00%)	35 (8.75%)
base-absent	50 (12.50%)	29 (7.25%)

Observed frequencies indicate the actual observation in the data, while expected frequencies refer to the frequencies anticipated based on the assumption that the variables are independent and that the null hypothesis is true. The expected frequencies are calculated by dividing the product of the marginal frequency of a row and the marginal frequency of a column by the total number of observations.

The Chi-square test was used to scrutinize if the difference between our observations (Table 4) and the expected random distribution (Table 5) is statistically significant. The formula of the Chi-square test is shown in Figure 2. 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.

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

Figure 2. Formula of the Chi-square test

By applying this test, the generated *p*-value is numerically indistinguishable from zero (2.2e-16), i.e. $\chi^2(2) = 193.17, p < 0.001$. This allows us to reject the null hypothesis of no association. Since we do not have cell values smaller than five in this analysis, the Fisher's exact test is not required as an additional verification.

Nevertheless, we still measured the effect size, which indicates the magnitude of the cross-group divergence, whereas the statistical significance shows the probability that the observed cross-group divergence is due to chance (Sullivan & Feinn 2012: 279). By way of illustration, a small *p*-value tells us that the variation across experiment groups is less likely to result from chance, whereas the effect size explains how strong is the association between the variables. The effect size was measured via Cramer's V, which is an extension of *the φ* ('phi') *coefficient* (i.e. mean square contingency coefficient). Such measure is analogous to the *Pearson correlation coefficient* and measures the level of association between two binary variables. As shown in Figure 3, Cramer's V is calculated by the square of the Chi-

squared statistic of our contingency table divided by the product of the total number of subjects and the degree of freedom of our observations.

$$\varphi = \sqrt{\frac{\chi^2}{n \cdot df}}$$

Figure 3. Formula of Cramer's V

Our 3×2 table has a degree of freedom which equals to (3−1)(2−1) = 2. A Cramer's V smaller than 0.21 thus represents a small effect size; between 0.21 and 0.35 indicates a moderate effect; bigger than 0.35 displays a strong effect. Based on our data, the generated Cramer's V equals to 0.695. Hence, the correlation between base-order and N-order in our dataset indicates a statistically significant association assimilated with a strong effect size.

4.2 Numeral base and numeral classifier

We now examine the proposed statistical universal (1b), i.e. base-CLF harmonization: If Num is base-initial, i.e. [*base n*], then a CLF-initial order obtains between CLF and Num; otherwise, a CLF-final order obtains. The null hypothesis is that there is no correlation between the base-order and CLF-order.

Classifier languages only account for a third of the total in our dataset (33.00%, 132/400). Such a ratio is consistent with previous studies, as classifiers are mostly found in Asia and languages of other regions more commonly employ other types of nominal classification systems, e.g. grammatical gender (Aikhenvald 2000; Corbett 2013; Gil 2013). Among classifier languages, the CLF-final order (26.50%, 106/400), where CLF follows Num, is much more prevalent than the CLF-initial order. Significantly, there are no attested classifier languages lacking a multiplicative numeral system, a fact predicted by the multiplicative theory of classifiers (Her 2012; 2017a; 2017b). Table 6 lists the detailed numbers of our pair analysis.

Table 6. Distribution of numeral base and CLF order in the 400 languages

	CLF-final	**CLF-initial**	**CLF-absent**
base-final	106 (26.50%)	3 (0.75%)	117 (29.25%)
base-initial	0 (0.00%)	23 (5.75%)	72 (18.00%)
base-absent	0 (0.00%)	0 (0.00%)	79 (19.75%)

The data encoded in Table 6 is displayed in Figure 4 via a bar plot. The x-axis symbolizes the word order of CLF, while the y-axis represents the frequency of

base-final (orange), base-initial (dark blue), and base-absent (olive green) languages, respectively. The plot demonstrates that there is an apparent correlation between the order of numeral base and CLF, as most of the CLF-final languages are base-final, and vice versa. On the other hand, the three possibilities of base-order are fairly evenly distributed within non-classifier languages, though the base-final order again enjoys a slight advantage.

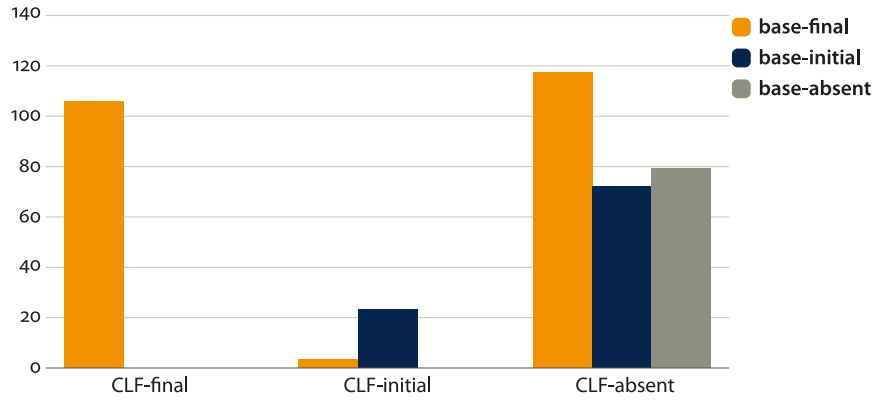


Figure 4. Bar plot of numeral base and CLF

Hence, the visual representation shows an apparent harmonization between base-order and CLF-order, though further statistical analysis is necessary to determine the effect size and statistical significance. The probability is generated via the Pearson's Chi-square (χ^2) test of independence. The null hypothesis is formulated as the absence of association between the variables, i.e. base-order and CLF-order are not correlated. On the other hand, the alternative hypothesis of (1b) states the opposite, i.e. the base-order and CLF-order are correlated. Table 7 shows the expected frequencies under the null hypothesis.

Table 7. Expected random distribution of order between numeral bases and CLF

	CLF-final	CLF-initial	CLF-absent
base-final	60 (15.00%)	15 (3.75%)	151 (37.75%)
base-initial	25 (6.25%)	6 (1.50%)	64 (16.00%)
base-absent	21 (5.25%)	5 (1.25%)	53 (13.25%)

The Chi-square test was again used to scrutinize if the difference between our observations (Table 6) and the expected random distribution (Table 7) is statistically significant. The obtained *p*-value is numerically indistinguishable from zero

(2.2e-16), i.e. χ^2 (4) = 163.65, $p < 0.001$. Thus, it is below the level of high statistical significance (p value < 0.01) and permits us to reject the null hypothesis.

However, even though the total quantity of our observations reaches the threshold in terms of statistical significance, three of the values are lower than 1 (i.e. CLF-final and base-initial, CLF-final and base-absent, CLF-initial and base-absent) and may have affected the output of the Chi-square test (Sheskin 2011: 646; Levshina 2015: 214). Hence, we also ran a two-tailed Fisher's exact test to verify the results obtained via the Chi-square test.

The Fisher's exact test calculates the probability of obtaining the values via the hypergeometric sampling distribution of the hypergeometric-likelihood measure. Thus, the product of the factorial of the sum of each row and column is divided by the product of the factorial of the value in every cell along with the factorial of the total amount of observations. The formula of the Fisher's exact test is shown in Figure 5. Assuming an $m \times n$ contingency table with m columns and n rows, C and R represents the sum of each row and column, whereas V indicates the individual value of every cell in the contingency table. Finally, n equals the sum of all the observations in the data.

$$\mathrm{P} = \frac{\prod_{i=1}^{m} C_i! \prod_{j=1}^{n} R_j!}{n! \prod_{i=1}^{m} \prod_{j=1}^{n} V_{ij}!}$$

Figure 5. Formula of the Fisher's exact test

The p-value of the Fisher's exact test is equally numerically indistinguishable from zero (2.2e-16), i.e. $p < 0.001$. Thus, both the Chi-square test of independence and the Fisher's exact test allows us to reject with high statistical significance the null hypothesis of no association between the variables of numeral bases and CLF.

We still needed to calculate the effect size via Cramer's V to obtain the strength of association between the variables. Since we are dealing with a 3×3 table, the degree of freedom equals (3−1)(3−1) = 4. A Cramer's V smaller than 0.15 represents a small effect size; between 0.15 and 0.25 indicates a moderate effect; bigger than 0.25 displays a strong effect. Based on our data, the generated Cramer's V equals to 0.452. Hence, the results of the correlation analysis between base-order and CLF-order in our dataset shows a statistically significant association combined with a strong effect size.

4.3 Classifier and order between the numeral and the noun

Next, we examine the proposed statistical universal (1c), i.e. CLF-N harmonization: If a CLF-initial order obtains between CLF and Num, then an N-initial order obtains

between N and Num; otherwise, an N-final order obtains. The null hypothesis is that there is no correlation between the CLF-order and N-order.

The raw numbers are displayed in Table 8. As observed in § 4.1 and § 4.2, the CLF-final (26.50%, 106/400) and N-final (63.00%, 252/400) word orders represent the majority of the data in comparison to the CLF-initial (6.50%, 26/400) and N-initial (37.00%, 148/400) parameters. The "no dominant order" column is again excluded from this table since no languages were annotated as such in our data.

Table 8. Distribution of CLF-order and N-order in the 400 languages

	N-final	N-initial
CLF-final	90 (22.50%)	16 (4.00%)
CLF-initial	3 (0.75%)	23 (5.75%)
CLF-absent	159 (39.75%)	109 (27.25%)

A visualization of the numbers of Table 8 is shown in Figure 6. While the CLF-final languages tend to apply N-final word order, the distribution of N-order is apparently more even within non-classifier languages. Thus, the plot suggests an association between the CLF-order and N-order when the first feature is present in the language. On the other hand, the N-order seems to be randomly distributed within non-classifier languages. Once again, further statistical analysis is required.

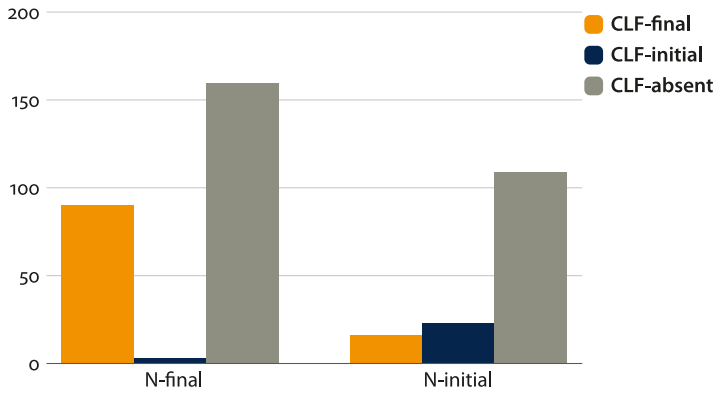


Figure 6. Bar plot of CLF-order and N-order

The Pearson's Chi-square (χ^2) test of independence was applied to measure the probability of the null hypothesis as the lack of association between the variables, whereas the alternative hypothesis indicates a correlation between the variables (i.e. CLF-order and N-order). The frequencies of observations expected by the null hypothesis are shown in Table 9.

Table 9. Expected random distribution of CLF-order and N-order

	N-final	N-initial
CLF-final	67 (16.75%)	39 (9.75%)
CLF-initial	16 (4.00%)	10 (2.50%)
CLF-absent	169 (42.25%)	99 (24.75%)

The Chi-square test outputs a p-value which is numerically close from zero (3.241e-12), i.e. $\chi^2(2) = 52.91, p < 0.001$. Hence, it is below the threshold of high significance (p value < 0.01) and allows us to reject the null hypothesis of no association between the variables. Nonetheless, as in § 4.2, one cell value is smaller than five in this analysis (i.e. CLF-initial and N-final = 3). Hence, the two-tailed Fisher's exact test is required as additional evidence. The obtained results are $p < 0.001$ (5.937e-13). Therefore, both tests support us in rejecting the null hypothesis. However, it is interesting to note that the p-value observed in this pair of comparison is relatively higher than the tests in § 4.1 and § 4.2. It thus implies that the association between the word order of classifiers and nouns is slightly less strong than the correlation among numeral bases and classifiers. This subject is further developed in § 5.

Furthermore, we also calculated the effect size by Cramer's V. Our 3×2 table has a degree of freedom of (3−1)(2−1) = 2. Hence, a Cramer's V smaller than 0.21 indicates a small effect size; between 0.21 and 0.35 points toward a moderate effect; bigger than 0.35 represents a strong effect. The Cramer's V of classifier and noun order equals to 0.364. Thus, the association of CLF-order and N-order in our dataset displays a statistically significant association combined with a strong effect size. Nevertheless, as found in terms of probability and effect size, such association is less strong compared to the relation among numeral bases and classifiers.

4.4 Overview and preliminary analysis

We now combine the three features which were analyzed by pairs and scrutinize their interaction across language families. Table 10 is a summary of the numbers mentioned in § 4.1, § 4.2, and § 4.3. While non-classifier languages have a relatively balanced distribution in terms of base-order and N-order, classifier languages show an apparent correlation between base-order, CLF-order, and N-order. With regard to the proposed probabilistic universal, only 8.25% (33/400) of the languages in our dataset represent exceptions (highlighted in grey), while the majority of the violations are related to N-order. Thus, the null hypothesis of no association can be rejected; while we find that the correlation between CLF-order and base-order is stronger than the association with N-order, as demonstrated by the statistical tests operated in the previous sections.

Table 10. Distribution of base-order, CLF-order, and N-order in 400 languages; shaded cells indicate violations of the probabilistic universal

		Base-final	Base-initial	Base-absent
CLF-final	N-final	90 (22.50%)	0 (0.00%)	0 (0.00%)
	N-initial	16 (4.00%)	0 (0.00%)	0 (0.00%)
CLF-initial	N-final	1 (0.25%)	2 (0.50%)	0 (0.00%)
	N-initial	2 (0.50%)	21 (5.25%)	0 (0.00%)
CLF-absent	N-final	110 (27.50%)	5 (1.25%)	44 (11.00%)
	N-initial	7 (1.75%)	67 (16.75%)	35 (8.75%)

To further explain this phenomenon, it would be relevant to investigate the origin of the exceptions, i.e. are certain language groups more likely to have non-aligned word orders due to their internal mutation? Or is divergence more likely to be generated by influence of contact? Intuitively, language families with more speakers and/or languages may show more diversity due to a higher possibility of inner-variation. Hence, we first analyzed via *simple linear regression* the 106 language families included in our dataset with regard to their speaker population and language diversity (i.e. amount of languages). In Figure 7, the y-axis represents the entropy of word order alignment, i.e. the measure of disorder in terms of word order within a language family. For instance, the higher the entropy, the stronger the lack of alignment between numeral bases, classifiers, and nouns. The x-axis indicates the logarithm of the speaker population (left) and the quantity of languages (right). The language families with high entropy are highlighted with their respective names.

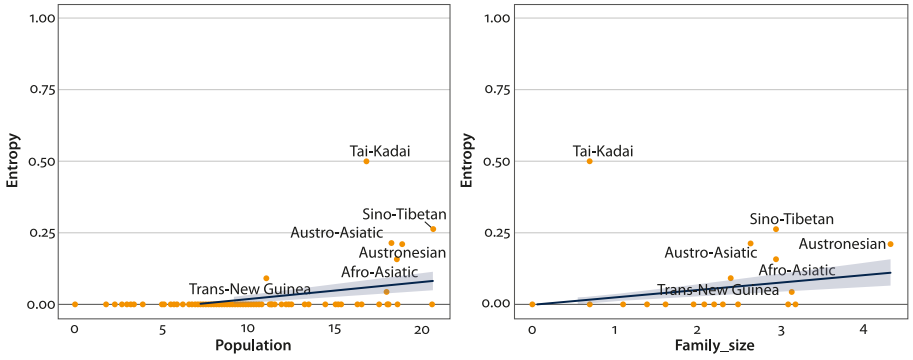


Figure 7. The entropy of word order alignment compared with speaker population (left) and quantity of languages within language families (right)

Simple regression is used to test if the speaker population of different language groups can significantly predict the entropy of alignment between numeral bases, CLF, and N. The results of the regression indicates that the predictor cannot explain the variance ($R^2 = 0.01795$, $F(1, 95) = 1.737$, $p > 0.05$). Likewise in terms of linguistic diversity, the amount of languages per family is also not a significant predictor for the entropy of alignment among the three linguistic structures we investigated ($R^2 = 0.00644$, $F(1, 95) = 0.6158$, $p > 0.05$). Thus, we cannot reject the null hypothesis of no association between the size of language families and our observations on word order alignment.

However, we did find that several language groups show higher entropy in comparison to others, i.e. word alignment is less consistent among language families such as Afro-Asiatic (15.78%), Austro-Asiatic (21.42%), Austronesian (21.05%), Niger-Congo (4.34%), Sino-Tibetan (26.31%), Tai-Kadai (50.00%), and Trans-New Guinea (9.09%). Therefore, we also considered the hypothesis that such a phenomenon is due to language contact. We computed the geographical distances and linguistic distances of languages through *Multidimensional Scaling* (MDS) in two dimensions and tested the correlation between the two distance matrices via a *Mantel test*. First, we plotted by dimensionality reduction the geographical distance between the 400 languages of our dataset on a two-dimensional space. As shown in Figure 8, the location of the 400 languages on the globe is reduced to a two-dimensional representation to ease display and later computations. For example, the relative isolation of the data point referring to *Hawaiian* reflects the geographical location of the language and its speakers in the North Pacific Ocean. The squares represent the languages aligned with order-initial features, while the circles indicate order-final languages. Violations are displayed with triangles. Such a spatial distribution is expected to fairly reflect the geographical distribution of the 400 languages in the world map since the *Kruskal Stress* measure of our plot (0.12) falls between 0.1 and 0.2 (Levshina 2015: 341). It does not reach the level of excellence (Stress < 0.05), but we estimated it sufficient for this preliminary analysis. Interestingly, most of the exceptions are indeed found at the meeting point of the clouds of head-initial and head-final languages, which visually support our language contact hypothesis.

We then computed the linguistic distances by comparing the variation between each language with regard to numeral bases, CLF, and N. For instance, a language annotated as [base-final, CLF-final] is expected to be more distant linguistically to a [base-initial, CLF-initial] language than a [base-final, CLF-initial] language. The *Gower general coefficient of similarity* is applied to generate such linguistic distance based on the three features examined in our dataset. As demonstrated in Figure 9, v indicates the quantity of variables, while i and j represents the observations, and s_{ijk} incarnates the similarity between i and j for the variable k. The Gower coefficient is thus obtained by dividing the sum of the product of

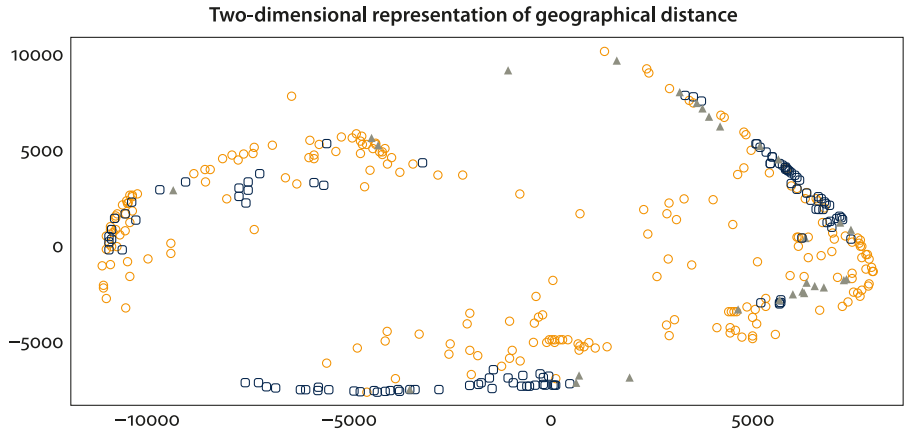


Figure 8. Two-dimensional representation of geographical distance

the similarity between i and j for the k variables and their weight by the sum of the weight of the k variables. The resulting distance matrix shows a Kruskal Stress measure of 0.15. We therefore estimate that this two-dimensional distribution fairly represents the linguistic distance of the variables within the 400 languages of our dataset without losing significant information during the process.

$$s_{ij} = \frac{\sum_{k=1}^{v} s_{ijk} w_k}{\sum_{k=1}^{v} w_k}$$

Figure 9. Formula of the Gower coefficient of similarity

Finally, to find out whether there is a correlation between the geographical distances and the linguistic distances within our dataset, we applied a *simple Mantel test with Kendall correlation coefficient*. The computation process of the Mantel test is shown in Figure 10. Assuming that l_{ij} indicates the linguistic distance and g_{ij} represents the geographical distance between the tokens i and j among the total of n observations, the null distribution z_m is derived by the sum of products of distances. Such distribution was then tested by the standard normal deviate through a standardized amount of permutations, i.e. 999 (Diniz-Filho et al. 2013: 476).

$$z_m = \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij} g_{ij}$$

Figure 10. Main formula of the Mantel test

The output of the Mantel test shows a statistically highly significant correlation with small effect size ($r_m = 0.09522$; $p < 0.001$). In other words, the linguistic distance based on the three parameters of word alignment (i.e. numeral bases, classifiers, and nouns) is correlated with the geographical distance between languages. Figure 11 displays an overview of this correlation. The x-axis indicates the geographical distance between languages in kilometres. The y-axis refers to the Mantel correlation coefficient given different geographical distances. Orange dots represent classes with statistically significant correlation coefficient ($p < 0.05$). The line represents the expectation of the Mantel statistic under the assumption of no correlation between linguistic and geographical distances. Values above the black line indicate positive correlation, whilst points below indicate negative correlation.

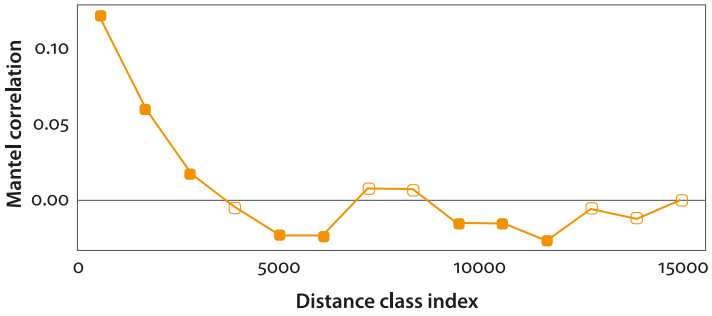


Figure 11. Mantel correlogram of Mantel correlation coefficient and geographic distance

The graph shows that the correlation between linguistic and geographic distances is stronger within a limited geographical range and decreases along with the increase of distance. To be more precise, the closer the geographical distance between languages, the more likely they are to bear similar linguistic features. However, this correlation becomes weaker when the geographical distance between languages increase. Languages within a range of near 500 kilometres tend to be similar in terms of linguistic features. However, the correlation drastically decreases after 500 kilometres and becomes negative or non-existent after near 4,000 kilometres. This is expected as languages that are 10,000 or 15,000 kilometres apart from the target language are equally unrelated to the target language, since they have an equally small (or nonexistent) level of contact with the target language. These results suggest that language contact is one of the significant influencing factors. However, such speculation requires additional analysis, as we barely included in this test 400 languages scattered around the world. Furthermore, additional controlling factors should be added to falsify the interaction between different variables. These limitations are further discussed in the following section.

5. Discussion

A spatial distribution of our data is displayed in Map 5. The circles represent languages with a head-final alignment, e.g. Mandarin Chinese is base-final, CLF-final, and N-final. The squares indicate head-initial languages, whereas violations are shown as triangles. Most of the exceptions are located at the intersection of the two tendencies, i.e. North-East of India, Eastern coasts of Africa, and North-East of the Oceania. We thus suspect that such divergence is due to language contact. We provided preliminary statistical analysis in § 4 from a macro-perspective. In this section, we scrutinize some of the languages of our dataset which represent violations to the probabilistic universal and provide detailed language examples as a micro-analysis.

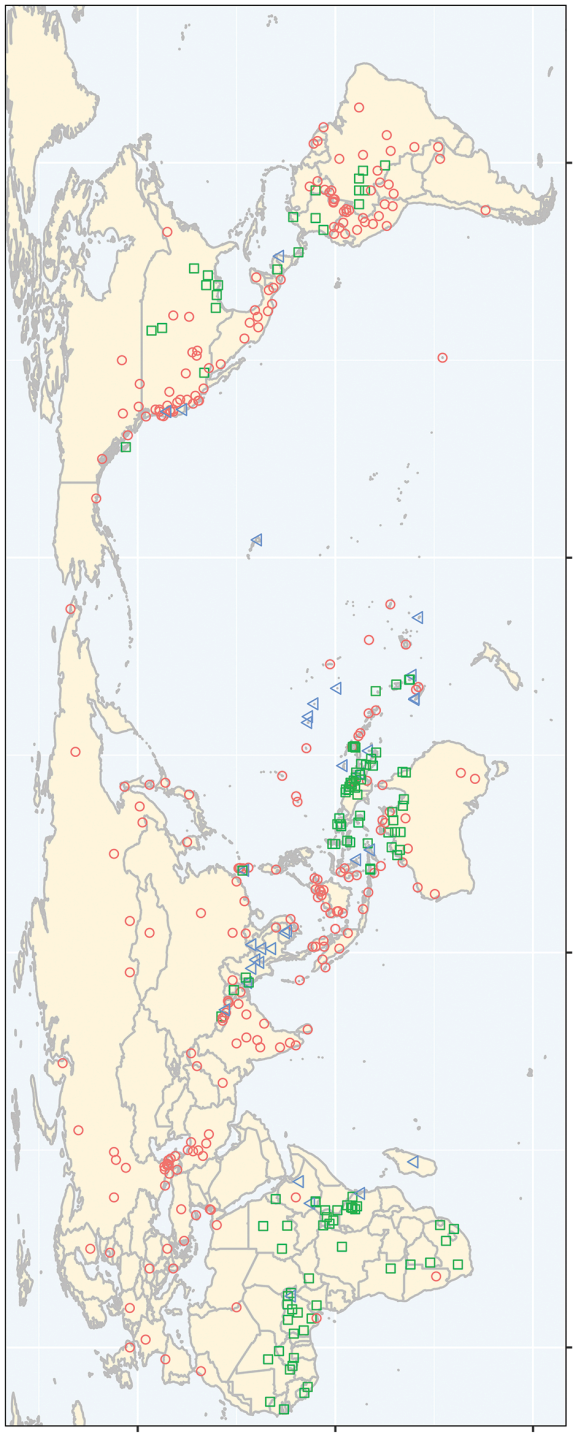
5.1 Examples of word order harmonization

Examples of alignment include three major types of situation. First of all, a language may possess numeral bases along with CLF, and have both of them aligned with the order of Num and N (27.75%, 111/400). By way of illustration in Cantonese (Sino-Tibetan), the numeral system is consistently base-final, as shown in Table 11.

Table 11. The numeral system of Cantonese (Chan 2018)

1. jɐt^{5}	10. sɐp^{2}	100. jɐt^{5} pak^{3}
2. i^{22}	20. i^{22} sɐp^{2}	200. i^{22} pak^{3}
3. sam^{53}	30. sam^{53}sɐp^{2}	1,000. jɐt^{5} ts^{h}in^{55}
4. sei^{33}	40. sei^{33} sɐp^{2}	2,000. i^{22} ts^{h}in^{55}
5. ṃ^{13}	50. ṃ^{13} sɐp^{2}	
6. lok^{2}	60. lok^{2} sɐp^{2}	
7. ts^{h}ɐt^{5}	70. ts^{h}ɐt^{5} sɐp^{2}	
8. pat^{3}	80. pat^{3} sɐp^{2}	
9. kɐu^{35}	90. kau^{53} sɐp^{2}	

With regard to CLF and N in Cantonese, both are aligned with the base-final Num, as CLF is positioned after Num, and N follows Num (Jiang & Hu 2010: 230–231). Thus, Num commonly precedes CLF and N (15a), whereas in a definite construction with the quantity of "one", CLF precedes N (15b).



Map 5. Spatial distribution of the 400 languages after alignment analysis

(15) Classifier in Cantonese
a. *sāam zek gau*
three CLF $_{animal}$ dog
'three dogs'
b. *zek gau zungji sek juk*
CLF $_{animal}$ dog like eat meat
'the dog likes to eat meat' (adapted from Jiang & Hu 2010: 230–231)

The second type of alignment involves non-classifier languages in which base-order concords with the word order of Num and N (44.25%, 177/400). For instance, Lucazi (Niger-Congo) is a non-classifier language with a base-initial numeral system. As shown in Table 12, the numeral base of decimals and higher numbers is consistently located after the multiplier digit, cf. *makúmi avalí* (10×2) 'twenty' and *vihíta vivalí* (100×2) 'two hundred'.

Table 12. The numeral system of Lucazi (Chan 2018)

1. *-mó*	10. *likúmi*	100. *cihítà*
2. *-vàli*	20. *makúmi avalí*	200. *vihíta vivalí*
3. *-tátù*	30. *mákumi atátu*	1,000. *likùlùkázì*
4. *-u̱ánà*	40. *mákumi au̱ána*	2,000. *makùlùkázì mavalí*
5. *-tánù*	50. *mákumi atánu*	
6. *-tánù nà -mó*	60. *mákumi atánu nalimó*	
7. *-tánù nà -vàli*	70. *mákumi atánu navalí*	
8. *-tánù nà -tátù*	80. *mákumi atánu natátu*	
9. *-tánù nà -u̱ánà*	90. *mákumi atánu nau̱ána*	

In terms of N-order in Lucazi, N regularly precedes Num, as in (16). Thus, we may observe that in Lucazi, the base-final setting is mirrored in the word order of Num and N, as the head noun of the phrase is in the final position.

(16) Order of Num and N in Lucazi
a. *vangombe likúmi*
cattle ten
'ten head of cattle'
b. *vangombe makúmi avalí*
cattle ten two
'twenty head of cattle' (Fleisch 2000: 102–103)

The third main type of alignment involves languages which have neither a multiplicative system nor classifiers (19.75%, 79/400). For instance, Djingili (Australian)

has only numerals from one to five. As demonstrated in Table 13, *kujkarrarni-kujkarrarni* 'four' is a reduplication of *kujkarrarni* 'two', whereas *marndamarnda* 'five' literally means 'hand'.

Table 13. The numeral system of Djingili (Chan 2018)

1. *kungkubarnu*
2. *kujkarrarni*
3. *murrkunbala*
4. *kujkarrarni-kujkarrarni*
5. *marndamarnda*

Djingili is not annotated as a classifier language (Gil 2013). The word order between Num and N, as shown in (17), is N-final. Given that there is no possible comparison with numeral bases and CLF, we consider that languages such as Djingili conform to the proposed probabilistic universals. Given the purpose to scrutinize the existence of exceptions to the hypotheses, the lack of violation is viewed as consistency with the probabilistic universals.

(17) Order of Num and N in Djingili
ngaja-nga-ju murrkunbala bayin-bala wijink-urri-ju nyambala
see-1SG-do three.MASC people-PL $_{\text{animate}}$ stand-3PL-do DEM $_{\text{n}}$
iurrju-mbili wijink-urru-ju
sandy.ridge-LOC upright-3PL-do
'I see three men standing on a sandy ridge' (Pensalfini 1997: 264)

5.2 Examples of word order disharmonization in classifier languages

Violations count as 8.25% (33/400) of the dataset and can be divided in two main categories, i.e. languages with classifiers and non-classifier languages. The first type constitutes the majority and accounts for 5.25% (21/400). Among them, 4.25% (17/400) have aligned bases and CLF, along with a disharmonized word order of Num and N. For example, the numeral system of the Tibeto-Burman language Newar (Sino-Tibetan) is base-final, as shown in Table 14.

Numerals in Newar commonly occur in bound form prefixed to CLF, as shown in (18). Newar thus has a CLF-final word order. However, the order between Num and N is N-initial, as in both examples the nouns *ki:* 'bug' and *santrAsi* 'orange' are positioned before Num-CLF. Similar observation are made by Dryer (2013) in languages such as Thai and Burmese.

Table 14. The numeral system of Newar (Chan 2018)

1. tɕʰə- / -tɕʰi	10. dʑi- / dʑĩː-	100. sətɕʰi-
2. ni- / nəsi	20. niː-	200. nisəː-
3. swə- / swə̃ː-	30. swiː-	1,000. dwəːtɕʰi-
4. pʲe- / pi-/ pʲẽː-	40. piː-	2,000. nidwəː-
5. nja-	50. neː-	
6. kʰu-	60. kʰwiː-	
7. n̥ɛː-	70. n̥ɛː-	
8. tɕja-	80. tɕɛː-	
9. gu- / gũː-	90. gwiː-	

(18) Classifiers in Newar
a. *ki: khu-mha*
bug six-CLF $_{\text{animate}}$
'six bugs'
b. *santrAsi cha-mA*
orange one-CLF $_{\text{plant}}$
'an orange tree' (Shakya 1997: 3–9)

A detailed analysis reveals that the picture is actually more complex, as diachronic change in Newar due to language contact could be the cause of such internal divergence in word order. Within the old texts, two different CLF-orders are found: Num-CLF and CLF-Num. However, according to (Kiryu 2009: 59–61), "all the modern Newar dialects allow only the Num-CLF order. In them it is not possible to place the classifier before the numeral". Furthermore, in modern Newar, "the numeral precedes the classifier, the classifier phrase can be case-marked, both orders (NP-[Num-CL] and [Num-CL]-NP) are allowed." Examples are given in (19).

(19) Word order in Newar
a. *thva na-hmam mi-m*
this five-CLF $_{\text{animate}}$ people-ERG
'these five people'
b. *gvatha ne-ma bhvana mesa smasta lisyam haya*
cowherd two-CLF $_{\text{animate}}$? buffalo all back bring
'the two herdsmen who brought back all the buffaloes'
(adapted from Kiryu 2009: 61)

Given the head-initial word order in Proto-Tibeto-Burman (Matisoff 1995), the current tendency of Newar favouring the head-final (i.e. base-final, CLF-final, N-final) order has thus been attributed to the influence of the head-final Nepali

language as the prestige language, along with the influence of neighbouring head-final Indo-Aryan languages to the West and Sinitic languages to the East (Kiryu 2009: 65).

A similar observation can be made for other exceptions of this type within Tibeto-Burman. Burmese (Sino-Tibetan), for instance, is also located at the meeting point of head-initial and head-final languages in our data. To the north, there are Chinese, Vietnamese, and Hmong (Bisang 1999: 118), all head-final ([Num-CLF]-N), and to the south, there are Thai and Khmer, where the dominant order is N-[Num-CLF] (Vittrant 2002: 132).

Analogous areal variation is observed in other regions of the world. The diversity attested in the Austronesian family is expected to be partially induced by contact between head-initial and head-final languages too. In Tetun, for example, the use of CLF is "preferable but not obligatory" (Williams-van Klinken et al. 2002: 39), as shown in (20). The examples also show that Tetun is CLF-initial and N-initial, showing harmonization.

(20) Classifiers in Tetun

a. *uma rua*
house two
'two houses'

b. *feto nain rua*
woman CLF $_{\text{human}}$ two
'two women' (Williams-van Klinken et al. 2002: 38–39)

However, the numeral system of Tetun appears to be base-final, as the decimals are commonly composed of a multiplier followed by a multiplicand, cf. in Table 15, *tolunulu* (3×10) 'thirty' and *limanulu* (5×10) 'fifty'. Thus, Tetun is a potential violation to the proposed probabilistic universal. Yet, a closer examination shows that while decimals are indeed base-final, the higher numbers are in fact base-initial, e.g. *atus rua* (100×2) 'two hundred' and *rihun rua* (1,000×2) 'two thousand'. This split can also be attributed to language contact, as higher numbers are commonly expressed via Portuguese or Indonesian, which possess base-final numeral systems, while indigenous Tetun numerals tend to be used only for smaller numbers (Williams-van Klinken et al. 2002: 38).[6]

This language contact hypothesis is also supported by the spatial distribution of languages shown in Map 5, and by the fact that more conservative Austronesian

6. Other factors should also be investigated to strengthen this speculation. For instance, do these larger numbers occur in greater frequency in both spoken and written corpora? Moreover, it is also not uncommon that classifiers are not needed with larger numbers, which could be relevant to the analysis.

Table 15. The numeral system of Tetun (Chan 2018)

1. *ida*	10. *sanulu*	100. *atus ida*
2. *rua*	20. *ruanulu*	200. *atus rua*
3. *tolu*	30. *tolunulu*	1,000. *rihun*
4. *ha:t*	40. *ha:tnulu*	2,000. *rihun rua*
5. *lima*	50. *limanulu*	
6. *ne:n*	60. *ne:nulu*	
7. *hitu*	70. *hitunulu*	
8. *ualu*	80. *ualunulu*	
9. *sia*	90. *sianulu*	

languages in the North-West preserved their original Numeral-Noun order, whereas languages in other regions may have adopted "the post-nominal order that is typical of the non-Austronesian languages of New Guinea" (Donohue 2007: 370–371).

Another type of violation which we speculate to be also motivated by language contact includes a harmonized order between N and the base, but a divergent word order of CLF. For instance in Nêlêmwa (Eastern Malayo Polynesian, Oceanic), the numeral system is consistently base-final, as the decimals and higher numbers are composed of a multiplier followed by a multiplicand, cf. [a:ru ak] (2×20) 'twenty' and [a:nem ak] (5×20) 'one hundred' in Table 16.

Table 16. The numeral system of Nêlêmwa (Chan 2018)

1. pʷaᵑgi:k	10. tuᶯɟic *	100. a:nem ak
2. pʷaⁿdu	20. a:ɣi ak ('ak' = person)	200. tuᶯɟic ak
3. pʷaᵑgan	30. a:ru ak ɣa ᵐbʷat tuᶯɟic	
4. pʷaᵐba:k	40. a:ru ak (2×20)	
5. pʷanem	50. a:ru ak ɣa ᵐbʷat tuᶯɟic	
6. pʷanemᵑgi:k	60. a:ɣan ak (3×20)	
7. pʷanemⁿdu	70. a:ɣan ɣa ᵐbʷat tuᶯɟic	
8. pʷanemᵑgan	80. a:va:k ak (4×20)	
9. pʷanemᵐba:k	90. a:va:k ak ɣa ᵐbʷat tuᶯɟic	

However, CLF is commonly prefixed to Num, while N appears at the right edge of the nominal phrase, as shown in (21). Nêlêmwa is thus a base-final and N-final language with a CLF-initial word order.

(21) Classifiers in Nêlêmwa
a. *aa-xiik* *shalaga*
CLF $_{animate}$-one crab
'one crab'
b. *aa-ru* *ak*
CLF $_{animate}$-two man
'two men' (Bril 2002: 380–382)

Nonetheless, the consulted sources suggest that the default structure in Nêlêmwa is actually head-initial, and that the order between CLF-Num and N may vary depending on the meaning conveyed. For instance, in (22a) the N-initial N-CLF-Num phrase 'three women' refers to the totality of a group, while in (22b) the N-final phrase refers to part of a bigger group. A similar scenario is found in Cham (Malayo-Sumbawan) (Baumgartner 1998: 11), while in Sanuma (Yanomam), which is in general N-initial, an N-final order occurs in a possessive construction or when the noun has already occurred in discourse (Borgman 1990: 129).

(22) Different classifier order in Nêlêmwa
a. *i* *axe thaamwa aa-xan*
3SG see woman CLF $_{animate}$-three
'he saw three women (in total)'
b. *i* *axe aa-xan* *thaamwa*
3SG see CLF $_{animate}$-three woman
'he saw three women (fraction)' (Bril 2002: 386)

Hence, the disharmonization of word orders we found is generally not complete violations to the probabilistic universal in question, as such phenomena are most likely due to language contact and a residue of the indigenous N-initial order still exists. Such observations in turn support our approach of probabilistic universals rather than absolute universals, as language contact may result in an intermediary stage between two tendencies on a continuum.

5.3 Examples of word order disharmonization in non-classifier languages

With regard to non-classifier languages, isolated cases of disharmonization between base-order and N-order are also attested (3.00%, 12/400). In Dizi (Afro-Asiatic), for example, numbers lower than a hundred are base-final, e.g. in Table 17, [útʃ͡ū tàmū] (5×10) 'fifty'. Yet, numbers above hundreds, e.g. [màtū t'à:gn̪̩] (100×2) 'two hundred' and [ʃí t'à:gn̪̩] (1,000×2) 'two thousand', are base-final. Language contact is again the culprit, as Dizi has borrowed numerals from both Cushitic and Amharic, which are base-initial and base-final, respectively (Allan 1976: 381; Chan

2018). Following the methodology outlined in § 3, Dizi is annotated as base-final since the base-final order is more common.

Table 17. The numeral system of Dizi (Chan 2018)

1. k'o:j	10. támū	100. màtū k'o:j
2. t'à:gn̪	20. t'à:gn̪ tàmū	200. màtū t'à:gn̪
3. kà:dū	30. kà:dū tàmū	1,000. ʃí k'o:j
4. k'ùbm̄	40. k'ùbm̄ tàmū	2,000. ʃí t'à:gn̪
5. útʃ͡ū	50. útʃ͡ū tàmū	
6. jàkū	60. jàkū tàmū	
7. tù:sū	70. tù:sū tàmū	
8. ze:d	80. ze:d tàmū	
9. sāgŋ̀	90. sāgŋ̀ tàmū	

As for the order between N and Num, both N-initial and N-final orders are attested, depending on the emphasis chosen by the speaker, as shown in (23) (Allan 1976: 381).

(23) Word order in Dizi

a. *wete jes ts' aniz t' àgn*
cow.PL fine black two
'two fine black cows'

b. *kùgn wete jèda dùenda k' ankàs*
four cow.PL fine fat PL
'four fine fat cows' (adapted from Allan 1976: 381)

However, Dizi is annotated as N-initial since it is the more common word order in the language. Our current analysis demonstrates that annotating with purely binary features does not reflect the degree of variance within a language, neither does it take account of the effect of language contact. Hence, we expect that in future studies with the above-mentioned two factors more satisfactorily taken care of, the statistics will show even stronger support to the proposed probabilistic universals.

5.4 Summary

Three main observations may be retrieved based on the results in § 4 and the language examples provided in § 5.1–§ 5.3. First, while the harmonization of word orders is statistically significant, further qualitative data on individual languages is required to analyze if the harmonization still holds true in situations of language

contact. Second, another factor influencing the level of harmonization involves the syntactic proximity of the elements. In the numbers of § 4 and language examples of § 5, we observe that under the head-parameter, N is more likely to diverge from base and CLF, as 90.91% (30/33) of the violations to the harmonization are due to a difference of N-order. We suggest that such phenomenon can be explained under the multiplicative theory outlined in § 2: the base and CLF both function as a multiplicand and thus Num and CLF form both a multiplicative unit as well as a syntactic constituent (Her & Tsai 2020), and are thus doubly constrained in word order harmonization, while N is constrained only by the head-parameter.[7] Third, as a suggestion for future studies, several language families (e.g. Austronesian, Tibeto-Burman, among others) have shown a high level of entropy in terms of word order harmonization. Further family-internal analysis is thus required to scrutinize whether this entropy results from geographical or phylogenetic factors.

6. Conclusion

Greenberg's (1990a: 292) Generalization 28 concerns the word order harmonization between numeral base and numeral classifier (CLF) and between base and noun (N). We further propose that harmonization between CLF and N should obtain. Three universal tendencies are thus under proposal. A detailed statistical analysis of a geographically and phylogenetically weighted set of 400 languages shows that the harmonization among the three elements is statistically highly significant, as only 8.25% (33/400) of the languages display violations. A fair number of language samples are provided to scrutinize the typologically peculiar cases encountered in the languages of our dataset. Most of the languages representing violations are located at the meeting point between head-final and head-initial languages; we thus speculate that language contact is the main influencing factor in the violations to the probabilistic universals proposed. The main limitation of our study comes from the method of data annotation. For instance, we only included binary choices such as the presence/absence of classifiers in a language. However, the detailed analysis of language samples indicates that languages with borderline structures are not rare and the information of their diversity is not

7. The proximity between Num and CLF is also found in the lexicon, as some language may use numeral bases that are classifiers at the same time. In Japhug (Sino-Tibetan, Qiangic), the base for hundreds is the classifier *-ri* (Jacques 2017: 140–141), e.g. [χsɯ-ri] 'three hundred'. It is possible that some languages went through a stage when the numeral base used to be a classifier like 'hundred' in Japhug, a pathway that would play a role in the statistical tendency brought to light in the paper.

fully captured with binary choices. Future studies could consider the application of continuous measures based on canonical features of the investigated linguistic structure.

Acknowledgments

We are grateful for the financial support by Taiwan's Ministry of Science and Technology (MOST) via the following grants awarded to the second author: 101-2410-H-004-184-MY3, 104-2633-H-004-001, 104-2410-H-004-164-MY3, and 106-2410-H-004-106-MY3. The first author is also thankful for the support of the IDEXLYON (16-IDEX-0005) Fellowship grant (2018–2021). We also thank the three anonymous reviewers for their valuable remarks and suggestions. All remaining errors are our own.

Abbreviations

1	first person	MASC	masculine
2	second person	MDS	Multidimensional Scaling
3	third person	MENS	mensural classifier
A	agent-like argument of canonical transitive verb	N	Noun
		NEG	negative
ASP	aspect	Num	Numeral
ASSOC	associative	PL	plural
CLF	classifier	PRS	present
DEM	demonstrative	PST	past
DIM	diminutive	QN	Quantifier-Noun
DIR	directional	RR	reflexive/reciprocal
ERG	ergative	S	single argument of canonical intransitive verb
GCM	general class marker		
GEN	genitive	SG	singular
LOC	locative	U	Unit
M	Multiplier	UA	unite-augmented

References

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

Allan, Edward J. 1976. Dizi. In Bender, M. Lionel (ed.), *The non-Semitic languages of Ethiopia* (Occasional Papers Series, Committee on Ethiopian Studies, Monograph 5), 377–392. East Lansing: African Studies Center, Michigan State University.

Au Yeung, Ben Wai Hoo. 2005. *An interface program for parameterization of classifiers in Chinese*. Kowloon: Hong Kong University of Science and Technology. (Doctoral dissertation.) https://doi.org/10.14711/thesis-b922442

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

Baumgartner, Neil I. 1998. A grammar sketch of Western (Cambodian) Cham. In Thomas, David (ed.), *Papers in Southeast Asian linguistics no. 15: Further Chamic studies* (Pacific Linguistics Series A-89), 1–20. Canberra: Pacific Linguistics (The Australian National University).

Besnier, Niko. 2002 [2000]. *Tuvaluan: A Polynesian language of the Central Pacific*. London: Routledge. (First published in 2000.) https://doi.org/10.4324/9780203027127

Bickel, Balthasar. 2014. Linguistic diversity and universals. In Enfield, Nick J. & Kockelman, Paul & Sidnell, Jack (eds.), *The Cambridge handbook of linguistic anthropology*, 102–127. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139342872.006

Bisang, Walter. 1999. Classifiers in East and Southeast Asian languages: Counting and beyond. In Gvozdanovic, Jadranka (ed.), *Numeral types and changes worldwide*, 113–186. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110811193.113

Blust, Robert. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics (The Australian National University).

Borer, Hagit. 2005. *Structuring sense, volume 1: In name only*. Oxford: Oxford University Press.

Borgman, Donald M. 1990. Sanuma. In Derbyshire, Desmond C. & Pullum, Geoffrey K. (eds.), *Handbook of Amazonian languages*, vol. 2, 15–248. Berlin: Mouton de Gruyter.

Bril, Isabelle. 2002. *Le nêlêmwa (Nouvelle-Calédonie): Analyse syntaxique et sémantique*. Paris: Peeters.

Chan, Eugene S. L. 2018. *Numeral systems of the world's languages*. (https://mpi-lingweb.shh.mpg.de/numeral/) (Accessed 2018-07-27.)

Cinque, Guglielmo & Krapova, Iliyana. 2007. A note on Bulgarian numeral classifiers. In Alboiu, Gabriela & Avram, Andrei A. & Avram, Laurentia Georgeta & Isac, Daniela (eds.), *Pitar Moş: A building with a view. Papers in honour of Alexandra Cornilescu*, 45–51. Bucharest: Editura Universităţii din Bucureşti.

Comrie, Bernard. 2013. Numeral bases. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/chapter/131) (Accessed 2018-07-27.)

Corbett, Greville G. 2003a. Agreement: Canonical instances and the extent of the phenomenon. In Booij, Geert E. & DeCesaris, Janet & Ralli, Angela & Scalise, Sergio (eds.), *Topics in morphology: Selected papers form the Third Mediterranean Morphology Meeting*, 109–128. Barcelona: Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra.

Corbett, Greville G. 2003b. Agreement: Terms and boundaries. In Griffin, William Earl (ed.), *The role of agreement in natural language: Proceedings of the 5th Annual Texas Linguistics Society Conference*, 109–122. Austin: Texas Linguistics Society.

Corbett, Greville G. 2003c. Agreement: The range of the phenomenon and the principles of the Surrey Database of Agreement. *Transactions of the Philological Society* 101(2). 155–202. https://doi.org/10.1111/1467-968X.00117

Corbett, Greville G. 2013. Number of genders. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/chapter/30) (Accessed 2018-07-27.)

Corbett, Greville G. & Fedden, Sebastian. 2016. Canonical gender. *Journal of Linguistics* 52(3). 495–531. https://doi.org/10.1017/S0022226715000195

Derbyshire, Desmond C. & Payne, Doris L. 1990. Noun classification systems of Amazonian languages. In Payne, Doris L. (ed.), *Amazonian linguistics: Studies in lowland South American languages*, 243–271. Austin: University of Texas Press.

Diniz-Filho, José Alexandre F. & Soares, Thannya N. & Lima, Jacqueline S. & Dobrovolski, Ricardo & Landeiro, Victor Lemes & de Campos Telles, Mariana Pires & Rangel, Thiago F. & Bini, Luis Mauricio. 2013. Mantel test in population genetics. *Genetics and Molecular Biology* 36(4). 475–485. https://doi.org/10.1590/S1415-47572013000400002

Donohue, Mark. 2007. Word order in Austronesian from north to south and west to east. *Linguistic Typology* 11(2). 349–391. https://doi.org/10.1515/LINGTY.2007.026

Dryer, Matthew S. 1998. Why statistical universals are better than absolute universals. *Papers from the 33rd Annual Meeting of the Chicago Linguistic Society*, 123–145. Chicago: Chicago Linguistic Society.

Dryer, Matthew S. 2013. Order of numeral and noun. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The word atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/chapter/89) (Accessed 2018-07-27.)

Epps, Patience & Bowern, Claire & Hansen, Cynthia A. & Hill, Jane H. & Zentz, Jason. 2012. On numeral complexity in hunter-gatherer languages. *Linguistic Typology* 16(1). 41–109. https://doi.org/10.1515/lity-2012-0002

Evans, Nicholas & Levinson, Stephen C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5). 429–448. https://doi.org/10.1017/S0140525X0999094X

Fedden, Sebastian & Corbett, Greville G. 2017. Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: A Journal of General Linguistics* 2(1). 1–47. (Article 34.) https://doi.org/10.5334/gjgl.177

Fleisch, Axel. 2000. *Lucazi grammar: A morphosemantic analysis* (Grammatical Analyses of African Languages 15). Cologne: Rüdiger Köppe.

Frank, Paul Stephen. 1985. *A grammar of Ika*. Philadelphia: University of Pennsylvania. (Doctoral dissertation.)

Gil, David. 2013. Numeral classifiers. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://wals.info/chapter/55) (Accessed 2018-07-27.)

Greenberg, Joseph H. 1990a. Generalizations about numeral systems. In Denning, Keith M. & Kemmer, Suzanne (eds.), *On language: Selected writings of Joseph H. Greenberg*, 271–309. Stanford: Stanford University Press.

Greenberg, Joseph H. 1990b. Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. In Denning, Keith M. & Kemmer, Suzanne (eds.), *On language: Selected writings of Joseph H. Greenberg*, 166–193. Stanford: Stanford University Press.

Grinevald, Colette. 2000. A morphosyntactic typology of classifiers. In Senft, Gunter (ed.), *Systems of nominal classification*, 50–92. Cambridge: Cambridge University Press.

Grinevald, Colette. 2015. Classifiers, linguistics of. In Wright, James D. (ed.), *International encyclopedia of the social & behavioral sciences*, 2nd edn., 811–818. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.53003-7

Her, One-Soon. 2012. 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, One-Soon. 2017a. Deriving classifier word order typology, or Greenberg's Universal 20A and Universal 20. *Linguistics* 55(2). 265–303. https://doi.org/10.1515/ling-2016-0044

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 & Hsieh, Chen-Tien. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11(3). 527–551.

Her, One-Soon & Tang, Marc & Li, Bing-Tsiong. 2019. Word order of numeral classifiers and numeral bases: Harmonization by multiplication. *Language Typology and Universals* 72(3). 421–452. https://doi.org/10.1515/stuf-2019-0017

Her, One-Soon & Tsai, Hui-Chin. 2020. Left is right, right is not: On the constituency of the classifier phrase in Chinese. *Language and Linguistics* 21(1). 1–32.

Jacques, Guillaume. 2017. The morphology of numerals and classifiers in Japhug. In Ding, Picus Sizhi & Pelkey, Jamin (eds.), *Sociohistorical linguistics in Southeast Asia: New horizons for Tibeto-Burman studies in honor of David Bradley*, 135–148. Leiden: Brill. https://doi.org/10.1163/9789004350519_009

Jany, Carmen. 2009. *Chimariko grammar: Areal and typological perspective*. Berkeley: University of California Press. https://doi.org/10.1525/9780520945197

Jiang, Li Julie & Hu, Suhua. 2010. On bare classifier phrases. In Clemens, Lauren Eby & Liu, Chi-Ming Louis (eds.), *Proceedings of the 22nd North American Conference on Chinese Linguistics (NACCL-22) & the 18th Annual Meeting of the International Association of Chinese Linguistics (IACL-18)*, vol. 2, 230–241. Cambridge: Harvard University.

Kilarski, Marcin. 2013. *Nominal classification: A history of its study from the classical period to the present* (Studies in the History of the Language Sciences 121). Amsterdam: John Benjamins. https://doi.org/10.1075/sihols.121

Kilarski, Marcin. 2014. The place of classifiers in the history of linguistics. *Historiographia Linguistica* 41(1). 33–78. https://doi.org/10.1075/hl.41.1.02kil

Kiryu, Kazuyuki. 2009. On the rise of the classifier system in Newar. In Nagano, Yasuhiko (ed.), *Issues in Tibeto-Burman historical linguistics* (Senri Ethnological Studies 75), 51–69. Osaka: National Museum of Ethnology.

Klamer, Marian. 2014. Numeral classifiers in the Papuan languages of Alor and Pantar: A comparative perspective. In Klamer, Marian & Kratochvíl, František (eds.), *Number and quantity in East Nusantara: Papers from 12-ICAL*, vol. 1, 103–122. Canberra: Asia-Pacific Linguistics (The Australian National University). (https://openresearch-repository.anu.edu.au/handle/1885/11917) (Accessed 2020-03-25.)

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins. (https://benjamins.com/catalog/z.195) (Accessed 2020-04-07.) https://doi.org/10.1075/z.195

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world*. Dallas: SIL International. (https://www.ethnologue.com/16/) (Accessed 2020-05-05.)

Mano, Miho. 2012. Compositional mechanisms of Japanese numeral classifiers. In Manurung, Ruli & Bond, Francis (eds.), *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, 620–625. Depok: Faculty of Computer Science, Universitas Indonesia.

Matisoff, James A. 1995. Sino-Tibetan numerals and the play of prefixes. *Bulletin of the National Museum of Ethnology* 20(1). 105–251.

Meakins, Felicity & Nordlinger, Rachel. 2014. *A grammar of Bilinarra: An Australian aboriginal language of the Northern Territory* (Pacific Linguistics 640). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9781614512745

Ostapirat, Weera. 2000. Proto-Kra. *Linguistics of the Tibeto-Burman Area* 23(1). 1–251.

Ostapirat, Weera. 2005. Kra-Dai and Austronesian: Notes on phonological correspondences and vocabulary distribution. In Sagart, Laurent & Blench, Roger & Sanchez-Mazas, Alicia (eds.), *The peopling of East Asia: Putting together archaeology, linguistics and genetics*, 1st edn., 107–131. London: Routledge Curzon. https://doi.org/10.4324/9780203343685_chapter_7

Pensalfini, Robert J. 1997. *Jingulu grammar, dictionary, and texts*. Cambridge: MIT. (Doctoral dissertation.)

Piantadosi, Steven T. & Gibson, Edward. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756. https://doi.org/10.1111/cogs.12088

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (https://www.R-project.org/) (Accessed 2020-03-25.)

Seifart, Frank. 2005. *The structure and use of shape-based noun classes in Miraña (North West Amazon)*. Nijmegen: Radboud University. (Doctoral dissertation.)

Shakya, Daya R. 1997. Classifiers and their syntactic functions in Nepal Bhasa. *Himalaya* (The Journal of the Association for Nepal and Himalayan Studies) 17(1). 1–24.

Sheskin, David J. 2011. *Handbook of parametric and nonparametric statistical procedures*. 5th edn. Boca Raton: Chapman & Hall/CRC Press.

Sullivan, Gail M. & Feinn, Richard. 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

Sussex, Roland & Cubberley, Paul. 2009. *The Slavic languages*. Cambridge: Cambridge University Press.

Toyota, Junichi. 2009. When the mass was counted: English as classifier and non-classifier language. *SKASE Journal of Theoretical Linguistics* 6(1). 118–130. (http://www.skase.sk/Volumes/JTL13/pdf_doc/07.pdf) (Accessed 2020-03-26.)

Velupillai, Viveka. 2012. *An introduction to linguistic typology*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.176

Vittrant, Alice. 2002. Classifier systems and noun categorization devices in Burmese. In Chew, Patrick (ed.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society: Special session on Tibeto-Burman and Southeast Asian linguistics*, 129–148. Berkeley: Berkeley Linguistics Society.

Watters, David E. 2002. *A grammar of Kham*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486883

Williams-van Klinken, Catharina & Hajek, John & Nordlinger, Rachel. 2002. *Tetun Dili: A grammar of an East Timorese language* (Pacific Linguistics 528). Canberra: Pacific Linguistics (The Australian National University).

Yamamoto, Kasumi & Keil, Frank. 2000. The acquisition of Japanese numeral classifiers: Linkage between grammatical forms and conceptual categories. *Journal of East Asian Linguistics* 9(4). 379–409. https://doi.org/10.1023/A:1008308724059

Authors' addresses

Marc Allassonnière-Tang (corresponding author)
Lab Dynamics of Language
UMR 5596
CNRS/University Lyon 2
DDL – MSH-LSE
14 Avenue Berthelot
69363 Lyon Cedex 07
France
marc.tang@univ-lyon2.fr

Publication history

Date received: 25 April 2018
Date accepted: 21 August 2018