Research Article

First Author, Second Author, and Third Author*

Defining numeral classifiers and identifying classifier languages of the world

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: This paper presents a precise definition of numeral classifiers, steps to identify a numeral classifier language, and a database of 3338 languages, where 723 languages have been identified as having a numeral classifier system. The database, named World Atlas of Classifier Languages (WACL), has been systematically constructed in the last ten years via a manual survey of relevant literature and also an automatic scan of digitized grammars followed by manual checking. The open-access release of WACL is thus a significant contribution to linguistic research in providing (i) a precise definition and examples of how to identify numeral classifiers in language data and (ii) the largest dataset of numeral classifier languages in the world. As such it offers researchers with a rich and stable data source for conducting typological, quantitative, and phylogenetic analyses on numeral classifiers. The database will also be expanded with additional features relating to numeral classifiers in the future in order to allow more fine-grained analyses.

Keywords: Classifiers, Database, Numeral classifiers, Sortal classifiers, Nominal classification

1 Why numeral classifiers?

Categorization is one of the most frequent and essential tasks realized by humans, as elements and experience encountered may be more efficiently stored and retrieved in the brain if they are categorized and organized (Clahsen, 2016, p.599; Lakoff and Johnson, 2003, p.162-163). This need is reflected in language via various mechanisms, one the most common being nominal classification systems (Fedden and Corbett, 2018; Kemmerer, 2014, 2017), among which the two most frequent types are grammatical gender and numeral classifiers (Aikhenvald, 2003; Audring, 2016; Corbett, 1991; Grinevald, 2015; Seifart, 2010). Examples of grammatical gender are the common/neuter distinction in Swedish (Indo-European, Europe), the masculine/feminine/neuter1/neuter2 distinction in Mian (Trans-New Guinea, Papunesia; Fedden, 2011), and the noun classes found in languages such as Swahili (Niger-Congo, Africa). Examples of numeral classifiers are the mostly shape-based classification of referents in languages like Mandarin Chinese (Sino-Tibetan, Asia), Nepali (Indo-European, Asia), and Tariana (Arawakan, South America). As shown in (1), classifiers can highlight various inherent features of a referent, including humanness (1a), shape (1b), and animacy (1c). The surveys in the World Atlas of Language Structures Online (WALS, Dryer and Haspelmath, 2013) on gender/noun class systems (Corbett, 2013, 43.6%, 112 on 257 languages having gender/noun class) and classifier systems (Gil, 2013, 35%, 140 on 400 languages having a classifier system) give some indication of the worldwide prevalence of these systems. These systems are studied across various fields such as linguistics, neuroscience, cognition, anthropology, and psychology, as they provide a window of analysis as to how the human mind works.

First Author, Institution, Department, City, Country of first author, e-mail: author_one@xx.yz
First Author, Institution2, Department, City, Country of first author, e-mail: author_one@xx.yz
Second Author, Institution, Department, City, Country of second author, e-mail: author_two@xx.yz
***Corresponding author: Third Author,** Institution, Department, City, Country of third author, e-mail: author_three@xx.yz

(1) Examples of numeral classifiers

a. *tin* *jana* *manche*
three CLF.HUMAN man

'three men' (Nepali, Allassonnière-Tang and Kilarski, 2020, p.127)

b. *yi4* *tiao2* *yu2*
one CLF.LONG fish

'one fish' (Mandarin Chinese)

c. *pa-ita* *tfinu*
one-CLF.ANIMAL dog

'one dog' (Tariana, Aikhenvald, 1994, p.423)

Nominal classification systems are neither redundant nor arbitrary, as they fulfil various lexical and discourse functions (Allassonnière-Tang and Kilarski, 2020; Eliasson and Tang, 2018; Her and Lai, 2012; Vittrant and Allassonnière-Tang, 2021). Taking grammatical gender as an example, the association between meaning and gender is far from being arbitrary (Allassonnière-Tang et al., 2021a; Basirat and Tang, 2018; Veeman et al., 2020). By way of illustration, the information of form and semantics can be used by machine learning and deep learning methods to predict the gender of nouns with an accuracy of around 90% in languages such as French, German, and Russian (Basirat et al., 2021). Likewise, the presence/absence of nominal classification systems is not arbitrary, and is subject to the influence of linguistic as well as non-linguistic factors (Allassonnière-Tang and Her, 2020; Her and Tang, 2020; Her et al., 2019; Tang and Her, 2019). For instance, shared tendencies of nominal classification systems often correlate with human cognitive biases. Within classifier languages, the most common classifiers relate to humanness, animacy, long-shape, and round-shape (Croft, 1994). This is hypothesized to relate to the cognitive saliency of these features: the first two features differentiate between humans and other entities (animals or objects), while the latter two features are salient shapes in our perception (Kemmerer, 2017, p.408).

While grammatical gender has been involved in linguistic studies from an early time, classifiers were of only minor interest in linguistic theories up to the end of the 20th century (Kilarski, 2013, 2014) when scholars converged on the pre-existing perspective that "the sex principle, which underlies the classification of nouns in European languages, is merely one of a great many possible classifications of this kind" (Boas, 1911, p.37). Linguistic works during that period mostly worked on establishing typologies of classifier systems and nominal classification in general (Adams and Conklin, 1973; Aikhenvald, 2000; Allan, 1977a; Craig, 1986; Denny, 1976; Grinevald, 2000; Lichtenberk, 1983; Seiler, 1986; Senft, 2000; Wils, 1935). More recent work on classifiers and nominal classification have focused on identifying the functions of such systems (Allassonnière-Tang and Kilarski, 2020; Contini-Morava and Kilarski, 2013), establishing their canonical morphosyntactic properties (Corbett and Fedden, 2016), identifying properties of concurrent systems (Fedden and Corbett, 2017) as well as the organization of categories of object concepts in the brain (Kemmerer, 2017, 2019).

During this development, the relevance of numeral classifiers to linguistics and other fields such as cognitive science has also been put forward. For instance, one of the most important functions of numeral classifiers relates to the count/mass distinction (Contini-Morava and Kilarski, 2013; Jackendoff, 1991; Wu and Her, 2021). See Supplementary Material A for an extended discussion on the subject.

Investigating such hypotheses quantitatively requires a large database on numeral classifier languages. Especially since findings about one classifier language might not be generalized to other classifier languages. For example, a large number of experimental studies of classifier focused on Mandarin, while a higher diversity would be ideal (Saalbach and Imai, 2012). However, large-scale structured data on numeral classifiers are scarce. As an example, the WALS (Gil, 2013) provides information on the presence/absence (and obligatoriness) of numeral classifiers, with 140 languages having numeral classifiers in a sample of 400 languages. As another example, the AUTOTYP database (Bickel and Nichols, 2002) has data for 272 classifier languages. While such samples are highly valuable, they are a rather small representation of the more than 7000 existing languages (Hammarström et al., 2019). Beside databases, individual research papers/books also provide data on classifiers in languages of the world (e.g., Allassonnière-Tang et al., 2021b; Greenberg, 1972, 1990a; Nichols, 1992), however, these contributions generally consider different types of classifiers with varying definitions. A more substantial and precisely defined database of numeral classifier languages worldwide with geographic information is essential to the research on the distribution of classifiers in language families and subgroups, the probable origin of numeral classifiers and the subsequent areal diffusion of this grammatical feature (Her and Li, in press), the interaction of classifiers with other classification systems, e.g., genders and noun classes, and also with other grammatical features, e.g., numeral bases and plural markers. The current study aims at providing such a data source by clarifying the definition of numeral classifiers and conducting a global search on more than 3000 languages worldwide.

2 What are numeral classifiers?

Even though the term 'numeral classifier' is quite frequently found in the literature on nominal classification (Aikhenvald, 2000, p.30; Bisang, 1999, p.113; Dixon, 1986, p.105; Grinevald, 2000, p.61), different sources tend to use different terms and some variety of names are found in the literature of nominal classification typologies and language descriptions (Blust, 2009, p.292; Wu and Her, 2021, p.42). Examples are classifiers, quantifiers (Adams, 1989), measure or quantitative words (Li, 1924), company words (Liu, 1965), specifiers (Huffman, 1970), projectives (Hurd, 1977), numeratives, numerical determinative (Chao, 1968), among others. Nevertheless, this is not as alarming as it first appears to be, as a detailed reading of the sources show that similar definitions are frequently used in spite of the naming.

To start with, it is necessary to distinguish between several types of classifiers, which can be identified based on the classifier locus (Aikhenvald, 2000; Grinevald, 1999, 2000; Kilarski and Allassonnière-Tang, 2021; Vittrant and Allassonnière-Tang, 2021): numeral classifiers, noun classifiers, genitive classifiers, deictic classifiers, verbal classifiers, and locative classifiers (Grinevald, 2000, p.62-68; Seifart, 2010, p.721). As indicated by their names, these constructional types of classifiers are differentiated based on the grammatical construction in which they occur, i.e. their distribution in the clause. In this study, we focus on numeral classifiers, which occur in numeral constructions, as shown in (1) and (5). Numeral classifiers systems are divided in two main subtypes based on different semantic (and sometimes syntactic) behaviours (Her et al., 2017; Peyraube and Wiebusch, 1993, p.52-53). First, sortal classifiers highlight or single out some inherent features of the referent denoted by a count noun (Her and Hsieh, 2010). They may also make explicit information about a given referent that the noun itself leaves unspecified, and they fulfill several semantic and discourse functions. For example, in Mandarin, the classifier for humans can be used to highlight that a teacher being referred to is respectable, which is not an information inherently specified for the noun 'teacher', c.f., *yi2 wei4 lao3shi1* (one CLF.HUMAN teacher) 'a teacher'. Second, mensural classifiers[1] are used for measuring both mass nouns and count nouns according to their physical properties (Bisang, 1999, p.121; Aikhenvald, 2000, p.115); however, unlike sortal classifiers, which do not alter the quantity of the nominal, mensural classifiers specify the quantity. For instance, in example (2a) from Mandarin Chinese, the noun 'fish' is used with a sortal classifier, *zhi1*, which highlights animacy. In (2b), the mensural classifier *xiang1* 'box' contributes new information about the quantity measured. Removing the sortal classifier *zhi1* in (2a) would result in **san1 yu2* (three fish), which is ungrammatical in ordinary speech but the meaning of 'three fish' is fully recoverable. Consequently, removing the mensural classifier *xiang1* in (2b) would result in a meaning of 'three fish'; the originally intended meaning of 'three boxes of fish' is no longer available. Finally, mass nouns such as 'water' can only be used with mensural classifiers, as shown in (2c).

[1] What Beckwith (1998) considers the third type, i.e., the so-called 'groupal' classifiers, are in fact a subtype of mensural classifiers under our definition of sortal and mensural classifiers; see Table 1 below.

(2) Sortal and mensural classifiers in Mandarin Chinese

a. *san1* *zhi1* *yu2*
three CLF.ANIMAL fish

'three fish'

b. *san1* *xiang1* *yu2*
three MENS.BOX fish

'three boxes of fish'

c. *san1* *ping2* *shui3*
three MENS.BOTTLE water

'three bottles of water'

A revealing insight into the potential cognitive function of numeral classifiers and the difference between sortal and mensural classifiers is based on the underlying multiplicative relation between the numeral, as a multiplier, and the classifier, as a multiplicand (Her, 2012; Her et al., 2017), inspired by Greenberg's (Greenberg, 1990b, p.172) original observation that "all the classifiers are...merely so many ways of saying 'one' or, more accurately, 'times one'." Sortal and mensural classifiers thus converge as the multiplicand of the numeral, but diverge in the mathematical values they encode, i.e., sortal classifiers encode the precise value of 'one' and mensural classifiers can represent any value, numerical or non-numerical, that is not necessarily 'one'. In (2b) above, *san1 xiang1 yu2* 'three boxes of fish' does not denote 'three fish' specifically, though the total number of fish can accidentally be three if each box contains exactly one fish. That is to say, while the mensural classifier 'box' also involves a multiplication, it is not necessarily 'times one', as opposed to sortal classifiers. In (3a), however, *san1 zhi1 yu2*, where *zhi1* is a sortal classifier, necessarily denotes 'three fish'. This can be further demonstrated by (3), where both *tiao2* and *wei3* are sortal classifiers like *zhi1*.

(3) The same noun with different classifiers in Mandarin Chinese.

a. *san1* *zhi1* *yu2*
three CLF.ANIMAL fish

'three fish'

b. *san1* *tiao2* *yu2*
three CLF.LONG fish

'three fish'

c. *san1* *wei3* *yu2*
three CLF.TAIL fish

'three fish'

All three sortal classifiers thus form the same multiplicative relation with the numeral *san1* 'three', i.e., $[3 \times 1]$, and the total number of fish denoted by all three expressions is thus 'three'. The formal definition of sortal classifiers as multiplicands with the value of 'one', shown in Table 1, further affords the advantage of a mathematically precise taxonomy of numeral classifiers (Wu and Her, 2021), which also departs from those offered in the literature.

All classifiers thus function as a multiplicand of the numeral, the multiplier, in the quantifying phrase and constitute a coherent syntactic category. Sortal classifiers are unique in that their inherent mathematical value must be numerical and fixed at 'one'; all other elements in the same syntactic position are thus mensural classifiers, whose values are anything but 'one'. The ones with a fixed numerical value other than 'one' are mensural classifiers like *shuang1* 'pair', which have the exact value of 'two'. Some other mensural

Tab. 1: Taxonomy of classifiers in Mandarin Chinese based on mathematical values (Her et al., 2017, p.3).

Numerical or not	Fixed or not		Examples	Classifier type
Numerical	Fixed	1	*zhi1* ANIMAL, *tiao2* LONG	Sortal classifier
Numerical	Fixed	$\neg 1$	2: *shuang1* 'pair', 12: *da3* 'dozen'	Mensural classifier
Numerical	Variable	$> 1(\neg 1)$	*qun2* 'group', *bang1* 'gang'	Mensural classifier
Non-numerical	Fixed	$\neg n(\neg 1)$	*jin1* 'catty', *ma3* 'yard'	Mensural classifier
Non-numerical	Variable	$\neg n(\neg 1)$	*di1* 'drop', *wan3* 'bowl'	Mensural classifier

classifiers have a variable numerical value, e.g., *qun2* 'group' may be any number larger than 'two'. Mensural classifiers can also have a fixed, or standard, non-numerical value, which can be weight, height, volume, time, money, etc., e.g., *ma3* 'yard' must be the exact length prescribed. Finally, mensural classifiers may also have a variable non-numerical value, e.g., *wan3* 'bowl' may be big or small in terms of volume.

This multiplicative relation between the numeral and the classifier also has crucial consequences in the constituent structure of the classifier construction and the typology of classifier word orders. In essence, given the multiplicative unit $[multiplier \times multiplicand]$ formed by the numeral and the classifier, the two must form a syntactic constituent which forms a larger constituent with the noun. This premise explains why among the theoretically possible word orders between numeral, classifier, and noun, the noun does not intervene between the numeral and the classifier (Her, 2017).

Mensural classifiers in numeral classifier languages are often compared to nominal terms of measure in non-classifier languages such as English due to the information of quantity they both provide. These two are often confused due to their similar semantic functions but should be differentiated with regard to their different syntactic behaviour (Croft, 1994, p.152; Her, 2012, p.1682). For instance, terms of measure in English are nouns (i.e., strictly lexical items) since they can take plural morphology and require the preposition 'of', cf. 'three cups of tea', when quantifying a noun. In a numeral classifier language, the classifiers do not take plural marking (if present in the language), and syntactically they behave as sortal classifiers in quantifying the noun directly without the mediation of an adposition. Sortal classifiers and mensural classifiers thus constitute the two subcategories of the distinct lexical category of numeral classifiers from nouns in most of classifier languages. Following this definition, sortal classifiers are typically a closed class while almost every noun of the lexicon can be used as a mensural classifier providing an appropriate context. In this study, we only consider sortal classifiers, i.e., mensural classifiers or terms of measure are not sufficient for a language to be marked as a numeral classifier language. Therefore, in the following text, we use the term 'classifier' to refer to sortal numeral classifiers.

Lastly, it makes no difference as to whether the classifier morphemes are bound (as seen in 1c) or free morphemes (as in 1a and 1b). The compulsory nature of the sortal classifier also varies according to languages. For instance, classifiers are considered obligatory with the numerals in Burmese but optional in Malay (Goddard, 2005, p.96; Nomoto and Soh, 2019). This variance of obligatoriness is language-specific and extremely context-specific (Nomoto, 2013) and is not extensively discussed cross-linguistically. In this study, we mark a language as having numeral classifiers whether their use is obligatory or optional.

As a practical guide to the previously defined criteria, to identify numeral classifiers in a given language, the following steps can be conducted:

(i) *consider all grammatical quantifying phrases.* By definition a quantifying phrase must have a quantifier and a nominal, but may also include other obligatory and optional morphemes. For example, we can consider the quantifying phrases in Mandarin as shown in (3).

(ii) *divide these morphemes into classes on distributional grounds.* Taking again the example of numerals in (3), we identify numerals (e.g., *san1* 'three'), classificatory morphemes (e.g., *zhi1* and *tiao2*), and nouns (e.g., *yu2* 'fish).

(iii) *if there is a class which is closed.* Following the previous example with Mandarin, we identify that the classificatory morphemes represent a closed class, as opposed to numerals and nouns.

(iv) *and if the members of that class can/must occur with an* open *class of nominals.* Following the previous example again, the classificatory morphemes can occur with nouns.

(v) *and if the members of that class single out a property particular to the meaning of the quantified nominal.* Following the example in Mandarin, the classifier *zhi1* singles out the feature ‘animal’ while the classifier *tiao2* singles out the feature ‘long’.

(vi) and if the members of that class preserve cardinality of countable nominals, the language has a classifier system.

First, we assume in point (i) that one can identify quantifying phrases, judge their grammaticality, do morpheme division on them and translate them. We also assume that issues relating to morpheme class division (point ii) and the distinction open vs closed class (point iii) can be resolved (Evans, 2000). Without points (ii-iii) many languages with compounds would qualify as languages with optional classifiers. Point (iv) ensures that the morphemes we are after do not relate only to a restricted set of nominals, but serve to, in principle, classify any nominal.

Point (v) is perhaps the most important characteristic of classifiers. Classifiers have meaning (cf. the discussion in Allan 1977b, p.290-294), which is a precision of the meaning of the classified nominal. As we state the requirement, the compatibility of a given classifier and nominal is determined by the classifier and the meaning (not its specific form) of the nominal. Since we are working with an open class of nominals, this implies that there are nominals where more than one classifier is compatible. Save for exceptions, it further implies that classifier compatibility need not be stored separately in the mental or descriptive lexicon of nouns of a classifier language. Classifiers here form a continuum towards gender, which we understand to be languages where only one gender is compatible with each noun, often with incomplete predictability, and therefore needs to be stored in the lexicon. Languages where most nouns have a fixed gender often have a closed set or class of nouns whose gender can alternate based on meaning (cf. Singer 2016, p.7), similar to classifiers as we define them here. The dividing line is what is the exception and what is the open-ended system, so that classifier languages have an average > 1 classes per noun against gender languages with ≈ 1. For example, on the one hand, a noun is either masculine or feminine in French. On the other hand, nouns in Mandarin Chinese can be used with different classifiers. As shown in (3) earlier, the noun ‘fish’ can be used interchangeably with classifiers for animals, long objects, or tails.

Finally, point (vi) relates to the counting functionality of classifiers, and thus the fact that they require the noun to be quantified to be a count noun (Allan, 1977a; Her, 2017, p.288).

We make no direct reference to the matter of classifier or gender (concordial) agreement. As is well-known, some languages are attested with nominal classification systems that are repeatedly marked on different elements of a clause (Derbyshire and Payne, 1990, p.256). As shown in (4) with Miraña, the general class marker (GCM) is present on the noun, numeral, and verb.

(4) Concordial markers in Miraña (Seifart, 2005, p.158)

a. *tsa-:pi* *gwa-hpi*
one-GCM.M.SG human-GCM.M.SG

‘one man’

b. *kát*ɯ́:βɛ*-bɛ* *gwa-hpi*
fall-GCM.M.SG human-GCM.M.SG

‘he fell, the man’

While agreement is deemed a necessary (but not sufficient) requirement for gender/noun class status (Corbett, 1991, p.146), this does not detract from our characterization of numeral classifiers. Hence, to judge whether a language has numeral classifiers (as defined here) one does not need to know the grammar of the language beyond the quantifying phrase. The potential issue of multifunctionality is addressed in a similar way. A classifier in a given language can be described as representing several classifier types simultaneously. For example, in languages such as Mandarin and Cantonese, some numeral classifiers can also be referred to as 'bare noun classifiers' (Simpson et al., 2011), indicating that those classifiers may occur with a noun but without a numeral to infer a definite interpretation. We do not quantify this multifunctionality in our identification process. That is to say, if a language has sortal classifiers within our definition, it counts as being a classifier language, regardless if those classifiers can have different functions outside of the quantifying phrase and be referred to as different classifier types in the literature.

The methodology used in this paper, which finds a lineage in Greenberg's (Greenberg, 1990a, p.172) insight that sortal classifiers express 'times one', significantly departs from the often informal and vague definitions found in previous studies. Gil (2013), for example, relies heavily on the concept of 'countability' in identifying sortal numeral classifiers. However, as shown in Table 1, sortal classifiers and mensural classifiers all require nouns of low countability, but only the ones with the precise numerical value of 'one' are sortal classifiers. Furthermore, the reliance on countability might also induce the serious misconception that non-classifier languages such as English have mensural numeral classifiers. Our methodology shows that while English, and other non-classifier languages, have terms of measurement such as pair, group, yard, and bowl that function exactly like Mandarin Chinese mensural classifiers semantically, they are syntactically nouns, not sortal classifiers at all. This methodology has helped clarifying that the Archaic Chinese in oracle bone inscriptions has mensural classifiers but not sortal classifiers and that Proto-Tibeto-Burman is a non-classifier language (Her and Li, in press). We are now in the process of using this methodology to re-examine putative classifier languages that seem to be borderline cases, especially the ones in Africa, Europe, and Taiwan.

3 Manual survey of literature and automatic scan of grammars

Based on the definitions provided in Section 2, we conducted two parallel surveys to identify languages that have numeral classifiers. During these surveys, we gathered as many language grammars that could be found as an attempt to cover as many languages as possible. First, a manual survey of language grammars was conducted to identify which languages were described as having numeral classifiers. The language examples available in each grammar were then used for applying the definition provided in Section 2. This method is, as far as we know, the most commonly used to construct databases such as WALS (Dryer and Haspelmath, 2013) and Autotyp (Bickel and Nichols, 2002; Nichols et al., 2013). In parallel, we also conducted an automatic survey in the collection of digitized grammatical descriptions from the DReaM Corpus (Virk et al., 2020). For the purposes of the present study, we selected the subset of descriptions that were (i) written in English as the meta-language, (ii) a grammar or grammar sketch[2] and (iii) a description of only one language — so that its contents could arguably be attributed to exactly that language. The resulting collection consisted of 7126 source documents describing 3240 languages spanning all areas of the world. The manual survey and the automatic survey (See Supplementary Material B) resulted in a sample of 3338 languages, which includes 723 numeral classifier languages.[3] Further details are provided in Section 4.

[2] These are the description types of Hammarström and Nordhoff (2011) that cover a spectrum the grammar of a language and would therefore be expected to bring up classifiers if and only if such are present in the language being described.

[3] By considering the result of the manual checking as the gold data, the overall accuracy of the automatic survey is around 0.89, which is comparable with human inter-coder agreement of similar tasks (Donohue, 2006, p.67-68). The precision is near 0.95 and much higer than the recall (0.62). Most errors are classifier languages that are not identified as such by the automatic survey. For more details on this quantitative comparison, please refer to Hammarström et al. (2021, p.32).

4 Results

A geographic visualization of the numeral classifier languages found in our surveys is shown in Figure 1. The data includes 723 (22%, 723/3338) numeral classifier languages and 2615 (78%, 2615/3338) languages without numeral classifiers. The data matches with the existing literature in two ways. First, numeral classifiers are rare, as only 22% of the languages have such a system. Our database also allows us to refine the attested distribution in existing online databases. As an example, Gil (2013) lists 140 numeral classifier languages in a database of 400 languages, which results in a proportion of 35%. Our data provide a more detailed idea of the scarcity of numeral classifier languages in languages of the world. This divergence of distribution can mostly be explained by a different of coverage and definition. First, it is possible that the WALS sample of 400 languages might have been coincidentally biased towards having numeral classifiers. Second, while Gil (2013) also considers sortal classifiers, the definition varies with ours. For example, Eyak (Athabaskan-Eyak-Tlingit) is annotated as a classifier language in WALS. However, considering the available references, we observe that "classifiers are strictly verb prefixes" in Eyak (Krauss, 2015, p.122). By comparing the two databases, there is a mismatch of annotation for 42 languages. 15 languages are annotated as having classifiers in WALS but not in our database, while 27 languages are annotated as not having classifiers in WALS but annotated as having classifiers in our database. If we were to replace these mismatching points with our data, the proportion of classifier languages would further increase in the WALS sample, which hints toward the first possibility that the option was accidentally biased towards classifiers. While we acknowledge that it would be interesting to compare the checklists of the two sources, there is no available checklist of criteria available for Gil (2013), we thus do not conduct such a comparison here.

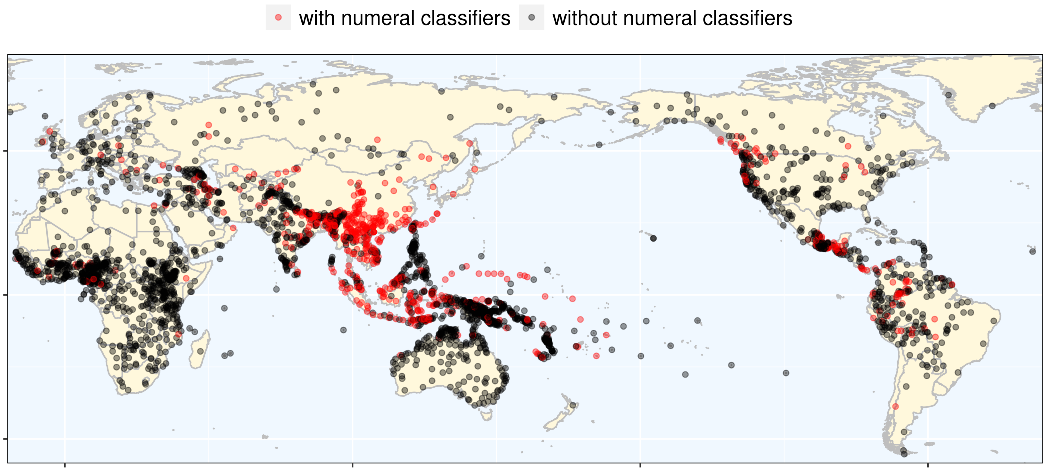

Fig. 1: The spatial distribution of numeral classifier languages. Each point represents a language.

Second, in terms of geographic distribution, the existing literature suggests that numeral classifiers are mostly found in Asia, while outside of Asia, they "are rare overall, but cluster along the Pacific rim in a pattern that, though clearly subcontinental in size, happens to span three macroareas: North Asian Coast (Old World), Oceania (Pacific), and the western coastline of North America, Mesoamerica, and South America (New World)" (Nichols, 1992, p.200). Our data matches with this overview (Table 2), in which we consider continents instead of glottoareas. The latter is not considered since it merges Europe and Asia into Eurasia and introduces noise into the visualization of the geographical distribution, as Europe has few classifier languages, while Asia has a lot of classifier languages. First, classifiers are mostly found in Asia. Second, classifiers are the least attested in Europe and Africa, while they are present but less frequent in the Americas and the Pacific when compared with Asia. More precisely, within the Pacific, numeral classifier languages are mostly found in Papunesia and are extremely rare in Australia. The scarcity of the classifiers in the Americas is likely due to the fact that only numeral classifiers (more specifically sortal classifiers) are included in our data, which excludes other types of classifiers that are generally found in languages spoken in South America.

Tab. 2: The proportion of numeral classifier languages across continents. The 'proportion on total' refers to the percentage of numeral classifier languages distributed across continents. For example, 70.1% of all the classifier languages are found in Asia. The 'proportion per continent' indicates the percentage of numeral classifier languages within each continent. For example, 45.1% of the languages in Asia are numeral classifier languages. The number of classifier languages and total languages differs from the numbers mentioned in the text (723/3338), because only languages with identified coordinates are mentioned in this table.

	Proportion on total		Proportion per continent	
Continent	**Count**	**Percentage**	**Count**	**Percentage**
Africa	29/680	4.3%	29/756	3.8%
Americas	111/680	16.3%	111/579	19.2%
Asia	477/680	70.1%	477/1058	45.1%
Europe	10/680	1.4%	10/112	8.9%
Pacific	53/680	7.8%	53/596	8.8%

The geographic distribution of numeral classifier languages can also be visualized in terms of proportion within each continent. As an example, while 70% of the numeral classifier languages are found in Asia, it is also necessary to understand how frequent are numeral classifier languages amongst languages of Asia. For instance, it is possible that the high proportion of numeral classifier languages in Asia is solely due to the fact that much more languages are found in Asia. To avoid such biases, it is necessary to visualize the proportion of numeral classifier languages in each continent. The results show that the ranking based on proportion across areas gives a similar proportion as the ranking calculated based on each individual area: Asia has the highest proportion of numeral classifier languages, followed by the Americas, while the proportion of numeral classifier languages is generally low in the Pacific, Africa, and Europe[4]. While the geographical distribution of classifiers generally matches with the literature, there are also divergence with observations from previous studies. As an example, some studies (Nichols, 1992; Sinnemäki, 2019) observe that numeral classifiers are commonly found in the Pacific than elsewhere. We do not engage in this issue within this paper, nevertheless we suggest that our database enables further testing of these observations from different perspectives.

Finally, we also visualize the distribution of numeral classifier languages across language families. Numeral classifier languages are found in 56 of the 203 language families included in the data. The proportion of numeral classifier languages of each of these families is listed in Figure 2. We observe that few families are only consisting of numeral classifier languages. Interestingly, these families are located either in Asia (Japonic and Hmong-Mien) or the Americas (Jodi-Saliban, Huavean, and Haida), which once again matches with the existing literature on the geographic distribution of numeral classifier languages. Furthermore, only 22 out of the 56 families have half or more than half of their languages as numeral classifier languages. The majority of the families have a small proportion of numeral classifier languages.

[4] As it is frequently questioned, the classifier languages found in Europe include languages such as Hungarian, which has optional classifiers (Csirmaz and Dekany, 2014; Dekany and Csirmaz, 2017).

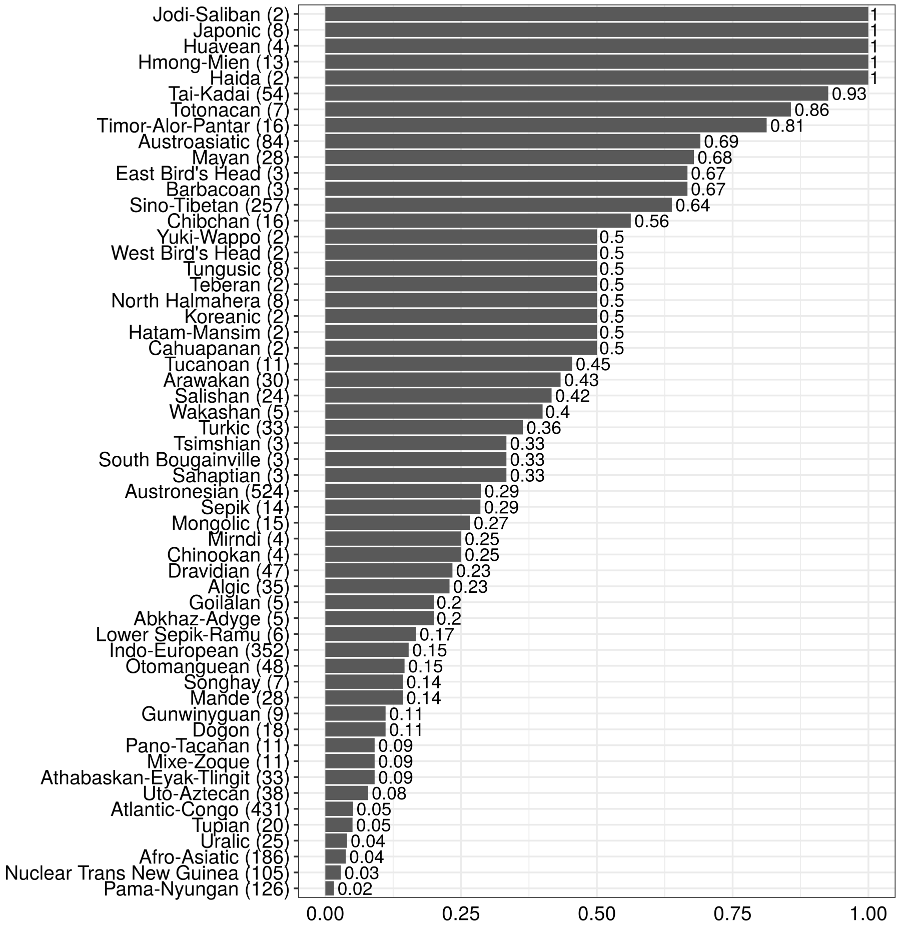

Fig. 2: The proportion of numeral classifier languages per family. The numbers in parenthesis refers to the number of languages included in the data for each family. Families without numeral classifier languages or families for which there is only one data point are not listed here.

As a summary, on the one hand, the data match with the existing literature by showing that Asia is a hotbed for numeral classifier languages. On the other hand, the data provide additional details with regard to the geographic and phylogenetic distribution of numeral classifier languages, which is helpful for the development of future studies. For example, the proportion of classifier languages per family shown in Figure 2 gives hints as to which language families could be suggested for studies on the evolution of numeral classifier with phylogenetic methods.

5 Summary and future development

The product of our clarified definition of numeral classifiers and our surveys is a database of numeral classifier languages. While its content match with the existing literature and provide additional details about the distribution of numeral classifier languages worldwide, we acknowledge that additional details and feedback from the linguistic community is needed to further enlarge and deepen our survey. Therefore, following the FAIR principles, we also aim at releasing the data obtained through our surveys as an online open-access database, which is named *The World Atlas of Classifier Languages* and abbreviated as WACL.

The content of WACL (Further details in Supplementary Material C) will be published under the CLLD framework (Forkel, 2014, https://clld.org/) under the CLDF format (Forkel et al., 2018) and hosted at the location *REMOVED TO KEEP ANONYMITY OF THE AUTHORS*. It will be updated on a yearly basis with a GitHub repository and a Zenodo frozen version. The version included in this paper is version 1. The building of WACL supports crowd science and will welcome comments and suggestions from the linguistic community to correct and/or expand the content of WACL. For example, even though the content of WACL is the result of automatic and manual scans, the content of WACL may be updated based on feedback from the linguistic community. WACL will also be expanded with additional features such as the obligatoriness/optionality of classifiers, detailed examples for each language in the database, differentiation of sub-categories of numeral classifiers (e.g., sortal vs. mensural classifiers), the inventory of classifiers in each language, among others. Opportunities of collaboration from various parties and/or institutions are also welcomed to suggest changes and/or new data points in the database.

Acknowledgment: To be added

Funding: To be added

Data Availability Statement and supplemental data

The content of the database will be available under the CLDF format and stored in a Github repository. The content of the database will be displayed and searchable through the website *REMOVED TO KEEP ANONYMITY OF THE AUTHORS*. All these files will be freely available under an open-source license. The content of the database will also have updated releases, which will be stored in Zenodo and assigned a DOI.

References

Adams, K. L. (1989). *Systems of numeral classification in the Mon-Khmer, Micobarese and Aslian Subfamilies of Austroasiatic*. Pacific Linguistics, Canberra.

Adams, K. L. and Conklin, N. F. (1973). Toward a theory of natural classification. In Corum, C., Smith-Stark, T. C., and Weiser, A., editors, *Papers from the ninth regional meeting of the Chicago Linguistic Society*, pages 1–10. University of Chicago, Chicago.

Aikhenvald, A. (2003). 4: Numeral Classifiers. In *Classifiers*, pages 98–124. Oxford: Oxford University Press.

Aikhenvald, A. Y. (1994). Classifiers in Tariana. *Anthropological Linguistics*, 36(4):407–465.

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

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

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

Allassonnière-Tang, M., Brown, D., and Fedden, S. (2021a). Testing Semantic Dominance in Mian Gender: Three Machine Learning Models. *Oceanic Linguistics*, 60(2):302–334.

Allassonnière-Tang, M. and Her, O.-S. (2020). Numeral base, numeral classifier, and noun: Word order harmonization. *Language and Linguistics*, 21(4):511–556.

Allassonnière-Tang, M. and Kilarski, M. (2020). Functions of gender and numeral classifiers in Nepali. *Poznan Studies in Contemporary Linguistics*, 56(1):113–168.

Allassonnière-Tang, M., Lundgren, O., Robbers, M., Cronhamn, S., Larsson, F., Her, O.-S., Hammarström, H., and Carling, G. (2021b). Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Humanities and Social Sciences Communications*, 8(1):331.

Audring, J. (2016). Gender. In Aronoff, M., editor, *Oxford research encyclopedia of linguistics*. Oxford University Press, Oxford.

Bale, A. and Coon, J. (2014). Classifiers are for numerals, not for nouns: Consequences for the mass/count distinction. *Linguistic Inquiry*, 45(4):695–707.

Basirat, A., Allassonnière-Tang, M., and Berdicevskis, A. (2021). An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*, 7(1):20200048.

Basirat, A. and Tang, M. (2018). Lexical and morpho-syntactic features in word embeddings: A case study of nouns in Swedish. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2:663–674.

Beckwith, C. I. (1998). Noun specification and classification in Uzbek. *Anthropological Linguistics*, 40(1):124–140.

Bickel, B. and Nichols, J. (2002). Autotypologizing databases and their use in fieldwork. In Austin, P., Dry, H., and Witternburg, P., editors, *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26 - 27 May 2002*. ISLE and DOBES, Nijmegen.

Bisang, W. (1999). Classifiers in East and Southeast Asian languages: Counting and beyond. In Gvozdanović, J., editor, *Numeral Types and Changes Worldwide*, volume 118 of *Trends in Linguistics: Studies and Monographs*, pages 113–186. Mouton de Gruyter.

Blust, R. (2009). *The Austronesian languages*. Pacific Linguistics, Canberra.

Boas, F. (1911). Chinook. In Boas, F., editor, *Handbook of American Indian Languages 1*, volume 40 of *Smithsonian Institution Bureau of American Ethnology Bulletin*, pages 559–678. Government Printing Office, Washington, D. C.

Borer, H. (2005). *Structuring Sense, part I*. Oxford University Press, Oxford.

Borer, H. and Ouwayda, S. (2010). Men and their apples: Dividing plural and agreement plural. In *Handout of a talk presented at GLOW Asia 8*, Beijing.

Bugaeva, A. (2012). Southern Hokkaido Ainu. In Tranter, N., editor, *The Languages of Japan and Korea*, pages 461–509. Routledge, New York.

Campbell, L. (1985). *The Pipil language of El Salvador*. De Gruyter Mouton, Berlin.

Cathcart, C., Hölzl, A., Jäger, G., Widmer, P., and Bickel, B. (2020). Numeral classifiers and number marking in Indo-Iranian: A phylogenetic approach. *Language Dynamics and Change*, pages 1–53.

Chao, Y. (1968). *A grammar of spoken Chinese*. University of California Press, Berkeley.

Chiarelli, V., El Yagoubi, R., Mondini, S., Bisiacchi, P., and Semenza, C. (2011). The syntactic and semantic processing of mass and count nouns: An ERP study. *PLoS ONE*, 6(10):1–15.

Chierchia, G. (1998). Plurality of mass nouns and the notion of semantic parameter. In Rothstein, S., editor, *Events and grammar*, pages 53–104. Kluwer, Dordrecht.

Chierchia, G. (2010). Mass nouns, vagueness and semantic variation. *Synthese*, 174(1):99–149.

Clahsen, H. (2016). Contributions of linguistic typology to psycholinguistics. *Linguistic Typology*, 20(3):599–614.

Contini-Morava, E. and Kilarski, M. (2013). Functions of nominal classification. *Language Sciences*, 40:263–299.

Corbett, G. G. (1991). *Gender*. Cambridge University Press, Cambridge.

Corbett, G. G. (2013). Number of Genders. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Corbett, G. G. and Fedden, S. (2016). Canonical gender. *Journal of Linguistics*, 52(3):495–531.

Cowper, E. and Hall, D. C. (2012). Aspects of individuation. In Massam, D., editor, *Count and mass across languages*, pages 27–53. Oxford University Press, Oxford.

Craig, C. (1986). *Noun classes and categorization*. John Benjamins, Amsterdam.

Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45(2):145–171.

Csirmaz, A. and Dekany, E. (2014). Hungarian is a classifier language. In Simone, R. and Masini, F., editors, *Word Classes: Nature, typology and representations*, pages 141–160. John Benjamins, New York.

Dekany, E. and Csirmaz, A. (2017). Numerals and quantifiers. In Alberti, G. and Laczko, T., editors, *Syntax of Hungarian: Nouns and noun phrases*, pages 1044–1150. Amsterdam University Press, Amsterdam.

Delahunty, G. P. and Garvey, J. J. (2010). *The English language: From sound to sense*. Parlor Press, West Lafayette.

Denny, P. (1976). What are noun classifiers good for? *Papers from the 12th regional meeting of the Chicago Linguistic Society*, pages 122–132.

Derbyshire, D. C. and Payne, D. L. (1990). Noun classification systems of Amazonian languages. In Payne, D. L., editor, *Amazonian linguistics, Studies in Lowland South American languages*, pages 243–271. University of Texas Press, Austin.

Dixon, R. M. W. (1986). Noun class and noun classification. In Craig, C., editor, *Noun classes and categorization*, pages 105–112. John Benjamins, Amsterdam.

Doetjes, J. (2012). Count/mass distinctions across languages. In Maienborn, C., Heusinger, K. v., and Portner, P., editors, *Semantics: An international handbook of natural language meaning, part III*, pages 2559–2580. Mouton de Gruyter, Berlin.

Donohue, M. (2006). Review of the The World Atlas of Language Structures. *LINGUIST LIST*, 17(1055):1–20.

Dryer, M. S. and Haspelmath, M. (2013). WALS Online. Place: Leipzig.

Eliasson, P. and Tang, M. (2018). The lexical and discourse functions of grammatical gender in Marathi. *Journal of South Asian Languages and Linguistics*, 5(2):131–157.

Enfield, N. J. (2004). Nominal classification in Lao: a sketch. *STUF - Language Typology and Universals*, 57(2-3):117–143.

Evans, N. (2000). Word classes in the world's languages. In Booij, G., Lehmann, C., and Mugdan, J., editors, *Morphology: a Handbook on Inflection and Word Formation*, volume 1, pages 708–732. Berlin: Mouton de Gruyter.

Fedden, S. (2011). *A grammar of Mian*. Walter de Gruyter, Berlin.

Fedden, S. and Corbett, G. G. (2017). Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: a journal of general linguistics*, 2(1):1–47.

Fedden, S. and Corbett, G. G. (2018). Extreme classification. *Cognitive Linguistics*, 29(4):633–675.

Forkel, R. (2014). The cross-linguistic linked data project. In Chiarcos, C., McCrae, J. P., Osenova, P., and Vertan, C., editors, *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 60–66. Reykjavik, Iceland: European Language Resources Association (ELRA).

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Nature Scientific Data*, 5(180205):1–10.

Ghomeshi, J. and Massam, D. (2012). The mass count distinction: Issues and perspectives. In Massam, D., editor, *Count and mass across languages*, pages 1–8. Oxford University Press, Oxford.

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

Gillon, B. S. (1999). The lexical semantics of English count and mass nouns. In Viegas, E., editor, *Breadth and depth of semantic lexicons*, pages 19–37. Springer, Dordrecht.

Goddard, C. (2005). *The languages of East and Southeast Asia: an introduction*. Oxford University Press, Oxford, N.Y.

Greenberg, J. H. (1972). Numeral Classifiers and Substantival Number: Problems in the Genesis of a Linguistic Type. *Working Papers on Language Universals*, 9:1–39.

Greenberg, J. H. (1990a). Generalizations about numeral systems. In Denning, K. and Kemmer, S., editors, *On language: Selected writings of Joseph H. Greenberg*, pages 271–309. Stanford University Press, Stanford. [Originally published 1978 in Universals of Human Language, ed by Joseph H. Greenberg, Charles A. Fergson, & Edith A. Moravcsik, Vol 3, 249-295. Stanford; Stanford University Press.].

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

Grinevald, C. (1999). Typologie des systèmes de classification nominale. *Faits de langues*, 7(14):101–122.

Grinevald, C. (2000). A morphosyntactic typology of classifiers. In Senft, G., editor, *Systems of nominal classification*, pages 50–92. Cambridge University Press, Cambridge.

Grinevald, C. (2015). Linguistics of classifiers. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences*, pages 811–818. Elsevier, Oxford.

Guerrero, J. (2015). *El dialecto árabe hablado en la ciudad marroquí de Larache*. Prensas de l'Universidad de Zaragoza, Zaragoza.

Hammarström, H. (2021). Measuring prefixation and suffixation in the languages of the world. In *Proceedings of The 3rd Workshop on Research in Computational Typology and Multilingual NLP*, pages 81–89. Stroudsburg, PA: Association for Computational Linguistics (ACL).

Hammarström, H., Her, O.-S., and Tang, M. (2021). Term-spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In Dobnik, S., Johansson, R., and Ljunglöf, P., editors, *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*, pages 27–34. Linköping: Linköping Electronic Press.

Hammarström, H., Castermans, T., Forkel, R., Verbeek, K., Westenberg, M. A., and Speckmann, B. (2018). Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). *Glottolog 4.1*. Max Planck Institute for the Science of Human History, Jena.

Hammarström, H. and Nordhoff, S. (2011). LangDoc: Bibliographic Infrastructure for Linguistic Typology. *Oslo Studies in Language*, 3(2):31–43.

Hammarström, H., Virk, S. M., and Forsberg, M. (2017). Poor man's ocr post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. *Proceedings of the Digital Access to Textual Cultural Heritage (DATeCH) conference*, pages 71–75.

Hansen, C. (1983). *Language and logic in ancient China*. University of Michigan Press, Ann Arbor.

Her, O.-S. (2012). Distinguishing classifiers and measure words: A mathematical perspective and implications. *Lingua*, 122(14):1668–1691.

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

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

Her, O.-S. and Hsieh, C.-T. (2010). On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics*, 11(3):527–550.

Her, O.-S. and Lai, W.-J. (2012). Classifiers: The many ways to profile one, a case study of Taiwan Mandarin. *International Journal of Computer Processing of Oriental Languages*, 24(1):79–94.

Her, O.-S. and Li, B.-T. (in press). A single origin of numeral classifiers in asia and the pacific: A hypothesis. In *Nominal classification in Asia and Oceania: Functional and diachronic perspectives*. John Benjamins, Amsterdam.

Her, O.-S. and Tang, M. (2020). A statistical explanation of the distribution of sortal classifiers in languages of the world via computational classifiers. *Journal of Quantitative Linguistics*, 27(2):93–113.

Her, O.-S., Tang, M., and Li, B.-T. (2019). Word order of numeral classifiers and numeral bases. *STUF - Language Typology and Universals*, 72(3):421–452.

Huffman, F. (1970). *Modern spoken Cambodian*. Yale University Press, New Haven.

Hurd, C. (1977). Nasioi Projectives. *Oceanic Linguistics*, 16(2):111.

Jackendoff, R. (1991). Parts and boundaries. *Cognition*, 41(1-3):9–45.

Janhunen, J., Peltomaa, M., Sandman, E., and Xiawu, D. (2008). *Wutun*. Number 466 in Languages of the world. Lincom Europa, Germany.

Jolly, L. (1989). Aghu Tharrnggala, a language of the Princess Charlotte Bay region of Cape York Peninsula. Master's thesis, University of Queensland, Brisbane.

Kemmerer, D. (2014). Word classes in the brain: Implications of linguistic typology for cognitive neuroscience. *Cortex*, 58:27–51.

Kemmerer, D. (2017). Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, 32(4):401–424.

Kemmerer, D. (2019). *Concepts in the Brain: The View From Cross-linguistic Diversity*. Oxford University Press, Oxford.

Kilarski, M. (2013). *Nominal classification: A history of its study from the classical period to the present*. John Benjamins, Amsterdam.

Kilarski, M. (2014). The Place of Classifiers in the History of Linguistics. *Historiographia Linguistica*, 41(1):33–79.

Kilarski, M. and Allassonnière-Tang, M. (2021). Classifiers in Morphology. In Aronoff, M., editor, *Oxford Research Encyclopedia of Linguistics*, pages 1–28. Oxford University Press, Oxford.

Krauss, M. (2015). Eyak grammar. Ms., University of Alaska.

Krifka, M. (1995). Common nouns: A contrastive analysis of Chinese and English. In Carlson, G. N. and Pelletier, F. J., editors, *The generic book*, pages 398–411. University of Chicago Press, Chicago.

Lakoff, G. and Johnson, M. (2003). *Metaphors we live by*. University of Chicago Press, London.

Langacker, R. W. (1977). *An overview of Uto-Aztecan grammar: Studies in Uto-Aztecan grammar*. Summer Institute of Linguistics and the University of Texas at Arlington, Dallas.

Lee-Smith, M. W. and Wurm, S. A. (1996). The Wutun language. In Wurm, S. A., Mühlhäusler, P., and Tryon, D. T., editors, *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, volume II.2, pages 883–897. Mouton de Gruyter, Berlin.

Li, J. (1924). *The grammar of Mandarin Chinese*. Shangwu Chubanshe, Beijing.

Lichtenberk, F. (1983). *A Grammar of Manam*. University of Hawaii Press, Honolulu.

Link, G. (1998). *Algebraic semantics in language and philosophy*. CSLI, Stanford.

Liu, S. (1965). *Wei-Jin Nanbeichao liangci yanjiu [A study on classifiers in the Wei-Kin and in the Nanbeichao periods]*. Zhonghua shuju chuban, Beijing.

Lorenzino, G. A. (1998). *The Angolar Creole Portuguese of São Tomé: Its Grammar and Sociolinguistic History*, volume 1 of *Lincom Studies in Pidgin and Creole Linguistics*. Lincom Europa, München.

Mathieu, E. (2012). On the mass-count distinction in Ojibwe. In Massam, D., editor, *Count and mass across languages*, pages 172–198. Oxford University Press, Oxford.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

Nichols, J., Witzlack-Makarevich, A., and Bickel, B. (2013). *The AUTOTYP genealogy and geography database: 2013 release*. Published: Electronic database available https://github.com/autotyp/autotyp-data accessed 2019-02-20.

Nomoto, H. (2013). *Number in classifier languages*. PhD dissertation, University of Minnesota, Minneapolis.

Nomoto, H. and Soh, H. L. (2019). Malay. In Vittrant, A. and Watkins, J., editors, *The Mainland Southeast Asia Linguistic Area*, pages 475–522. De Gruyter Mouton, Berlin.

Ojah, D. (1995). *A critical study of Barpeta dialect*. PhD Dissertation, Gauhati University, Assam.

Peyraube, A. and Wiebusch, T. (1993). Le rôle des classificateurs nominaux en chinois et leur évolution historique : un cas de changement cyclique. *Faits de langues*, 1(2):51–61.

Quine, W. v. O. (1960). *Word and object*. MIT Press, Cambridge.

Saalbach, H. and Imai, M. (2012). The relation between linguistic categories and cognition: The case of numeral classifiers. *Language and Cognitive Processes*, 27(3):381–428.

Sanches, M. and Slobin, L. (1973). Numeral classifiers and plural marking: An implicational universal. *Working Papers in Language Universals*, 11:1–22.

Sandman, E. (2016). *A grammar of Wutun*. PhD Dissertation, University of Helsinki, Helsinki.

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

Seifart, F. (2010). Nominal Classification. *Language and Linguistics Compass*, 4(8):719–736.

Seiler, H. (1986). *Apprehension: language, object and order*. Gunter Narr, Tübingen.

Senft, G. (2000). *Systems of nominal classification*. Cambridge University Press, Cambridge.

Simpson, A., Soh, H. L., and Nomoto, H. (2011). Bare classifiers and definiteness: A cross-linguistic investigation. *Studies in Language*, 35(1):168–193.

Singer, R. (2016). *The dynamics of nominal classification: productive and lexicalised uses of gender agreement in Mawng*. Number 642 in Pacific Linguistics. De Gruyter Mouton, Boston.

Sinnemäki, K. (2019). On the distribution and complexity of gender and numeral classifiers. In Di Garbo, F., Olsson, B., and Walchli, B., editors, *Grammatical gender and linguistic complexity*, pages 133–200. Language Science Press, Berlin.

Tang, M. and Her, O.-S. (2019). Insights on the Greenberg-Sanches-Slobin generalization: Quantitative typological data on classifiers and plural markers. *Folia Linguistica*, 53(2):297–331.

T'sou, B. K. (1976). The Structure of Nominal Classifier Systems. In Jenner, P. N., Thompson, L. C., and Starosta, S., editors, *Austroasiatic Studies Part II*, Oceanic Linguistics Special Publication, pages 1215–1247. University Press of Hawaii, Honolulu.

Veeman, H., Allassonnière-Tang, M., Berdicevskis, A., and Basirat, A. (2020). Cross-lingual Embeddings Reveal Universal and Lineage-Specific Patterns in Grammatical Gender Assignment. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 265–275, Online. Association for Computational Linguistics.

Virk, S. M., Borin, L., Saxena, A., and Hammarström, H. (2017). Automatic extraction of typological linguistic features from descriptive grammars. In Ekštein, K. and Matoušek, V., editors, *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, volume 10415 of *Lecture Notes in Computer Science*, pages 111–119. Berlin: Springer.

Virk, S. M., Hammarström, H., Forsberg, M., and Wichmann, S. (2020). The dream corpus: A multilingual annotated corpus of grammars for the world's languages. *Proceedings of The 12th Language Resources and Evaluation Conferenc*, pages 871–877.

Virk, S. M., Muhammad, A. S., Borin, L., Aslam, M. I., Iqbal, S., and Khurram, N. (2019). Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, page 1247–1256. Varna, Bulgaria: NCOMA Ltd.

Vittrant, A. and Allassonnière-Tang, M. (2021). Classifiers in Southeast Asian languages. In Sidwell, P. and Jenny, M., editors, *The Languages and Linguistics of Mainland Southeast Asia*, pages 733–772. De Gruyter.

Volker, C. A. (1998). *The Nalik language of New Ireland, Papua New Guinea*. Peter Lang, Bern.

Wils, J. (1935). *De nominale klassificatie in de Afrikaansche Negertalen*. PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen.

Wu, J.-S. and Her, O.-S. (2021). Taxonomy of numeral classifiers. In Lee, C., Kim, Y.-W., and Yi, B.-u., editors, *Numeral Classifiers and Classifier Languages: Chinese, Japanese, and Korean*, pages 40–71. Routledge, 1 edition.

Yi, B. U. (2011). What is a numeral classifier? *Philosophical Analysis*, 23:195–258.

Zhang, N. N. (2012). Countability and numeral classifiers in Mandarin. In Massam, D., editor, *Count and mass across languages*, pages 220–237. Oxford University Press, Oxford.

A Supplementary Material - Classifiers in the literature

Count nouns are perceived as semantically bounded entities that can be individuated and counted, while mass nouns refer to things whose parts are not considered as discrete units (Bisang, 1999, p.120; Delahunty and Garvey, 2010, p.156). This distinction is mirrored through language (Chierchia, 1998, 2010; Doetjes, 2012; Gillon, 1999; Quine, 1960), as our brain “differentiates between count and mass nouns not only at the syntactic level but also at the semantic level” (Chiarelli et al., 2011, p.1). This function is generally referred to as ‘individualization’ (Bisang, 1999, p.120) or ‘unitizing’ (Enfield, 2004, p.132).[5] In numeral classifier languages, count nouns use sortal classifiers in contexts of enumeration/quantification and mensural classifiers in contexts of measure, whereas mass nouns must rely on mensural classifiers.[6] As demonstrated in (5), semantically unbounded mass nouns such as ‘water’ cannot apply sortal classifiers (5a) and can only be quantified with mensural classifiers (5b). See Tang and Her (2019) for a theoretical and quantitative analysis on the subject matter.

(5) Individuation by numeral classifiers in Vietnamese (Austroasiatic, Vietnam)

a.

**ba*	*cái*	*nu’ó’c*
three	CLF.GEN	water

‘three water’

b.

ba	*chai*	*nu’ó’c*
three	MENS.BOTTLE	water

‘three bottles of water’

By further analyzing how languages fulfill the function of individuation, previous typological studies found that numeral classifiers and grammatical plural markers[7] (Tang and Her, 2019). follow a complementary-like distribution cross-linguistically. Thus, different hypotheses have been developed to explain this observation (Ghomeshi and Massam, 2012, p.2). First, a typological approach suggests that numeral classifier languages, unlike plural-marking languages, either do not make the mass-count distinction or only make this distinction semantically, but not syntactically, and therefore do not allow nouns to be quantified by numerals directly without classifiers (Allan, 1977a; Bale and Coon, 2014; Chierchia, 1998; Hansen, 1983; Krifka, 1995; Link, 1998; Zhang, 2012). Thus, nouns in numeral classifier languages are all mass nouns or transnumeral nouns, i.e. nouns are not specified for number in the lexicon. A universalist approach, on the other hand, claims that sortal classifiers and plural markers are unified under one grammatical category (Borer, 2005; Borer and Ouwayda, 2010; Cowper and Hall, 2012; Doetjes, 2012; Greenberg, 1990a; Mathieu, 2012; Nomoto, 2013; Sanches and Slobin, 1973; T’sou, 1976; Wu and Her, 2021; Yi, 2011). Under this hypothesis, the mass-count distinction is recognized in both types of languages, where the use of a sortal classifier is analogous to that of a plural marker.

[5] It is important to point out that even though there are cross-linguistic patterns of individualization, the exact count/mass boundary varies between languages.

[6] Sortal classifiers and mensural classifiers are two subtypes of numeral classifiers. The definition of numeral classifiers and sortal/mensural classifiers will be further developed in Section 2.

[7] Grammatical plural (also called grammatical number) engages in grammatical agreement outside the noun phrase. It is distinguished from semantic plural, which are only marked on the noun and relate to collective and associative marking

B Supplementary Material - Grammar survey

Each document in the collection has been OCRed using ABBYY Finereader 14 with English set as the recognition language. In essence, the OCR correctly recognizes most tokens of the meta-language but is hopelessly inaccurate on most tokens of the vernacular being described. This is completely expected from the typical, dictionary/training-heavy, contemporary techniques for OCR, and cannot easily be improved on the scale relevant for the present collection. Some post-correction of OCR output very relevant for the genre of linguistics is possible (see Hammarström et al., 2017) but made little difference for the present study.

Since information extraction from raw text grammatical description has only recently become practical, very little work has so far been done on this task. The first attempts (Hammarström, 2021; Virk et al., 2017, 2019), naturally, have explored the range and strength of hand-written rules for specific features. For the present case, an even simpler technique called keyword extraction seemed possible, namely to simply look for occurrences of the term 'classifier(s)'. At first blush, keyword extraction might seem trivial: simply look for the existence of the keyword and/or its relative frequency in a document, and infer the feature associated with the keyword. Unfortunately, to simply look for the existence of a keyword is too naive. In many grammars, keywords for grammatical features do occur although the language being described, in fact, does not exhibit the feature. For example, the grammar may make the explicit statement that there are "no X" incurring at least one occurrence.[8] Also, what frequently happens is that comments and comparisons are made with other languages — often related languages or other temporal stages — than the main one being described.[9] Furthermore, there's always the possibility that a term occurs in an example sentence, text or title in one of the references. However, such "spurious" occurrences will not likely be frequent, at least not as frequent as a keyword for a grammatical feature which actually belongs to the language and thus needs to be described properly. But how frequent is frequent enough? In order to avoid the labour and subjectivity of tuning a threshold manually, the following heuristic has been developed (described in more detail in Hammarström et al. 2021).

Suppose that we have several different grammars for the same language. As they are describing the same language we can assess the generality of the terms occurring in each document. The generality of a term in one grammar can be calculated from the proportions by which that term occurs in the other grammars for the same language. The overall generality of a whole grammar can then be obtained as the weighted average of the generality of its terms (there are various ways to do this, see Hammarström et al. 2021, p.29-30). The overall generality of a document i constitutes a proportion α_i which we hypothesize to be akin to the ratio between "signal" and "noise" in this document (Hammarström et al., 2021, p.29-30). For languages where we have only one document, we may simply take average α_i for documents of similar size. We can then recapture the question "how frequent is frequent enough?" as: does the frequency of a term in a grammar exceed its noise level $(1 - \alpha_i)$? Assuming that the fraction $(1 - \alpha_i)$ of least frequent tokens are "noise". Simply subtracting the fraction $(1 - \alpha_i)$ of tokens of the least frequent types effectively generates a threshold t separating the tokens being retained versus those subtracted. For example, consider Table 3 below with grammars and grammar sketches of Wutun [wuh]. The grammar of Sandman (2016) has an α_i of 0.91 and contains a total of 100624 tokens.

If we subtract $(1 - 0.91) \cdot 100624 \approx 9056$ tokens from the least frequent types, this leaves only types with frequency of 4 or more, defining the frequency threshold $t = 4$. Each grammar has a corresponding α purity level as described above, the total number of tokens, and the frequency threshold t induced by α_i. The 'classifier' column contains the frequency of this term. The cells with a frequency that exceeds the threshold t for their corresponding grammar are shown in green, indicating that the keyword in question is

[8] One example is the Pipil grammar of Campbell (1985, p.61): "It should be noted that unlike Proto-Uto-Aztecan (Langacker, 1977, p.92-93) Pipil has no productive postpositions. However, it has reflexes of former postpositions both in the relational nouns (cf. 3.5.2) and in certain of the locative suffixes (cf. 3.1.3).".

[9] For example, Lorenzino (1998)'s description of Angolar Creole Portugues [aoa] contains a number of references to the fate of nouns that were masculine in Portuguese, yet the modern Angolar does not have masculine, or other, gender.

Tab. 3: Automatic detection of classifiers in grammars from the language Wutun. The abbreviations are read as follows: G = grammar, S = grammar sketch.

Sources for Wutun [wuh]	bibtype	α_i	t	# tokens	Classifier
Sandman 2016	G	0.91	4	100624	38
Janhunen, Peltomaa, Sandman and Dongzhou 2008	S	0.79	4	41509	31
Lee-Smith and Wurm 1996	S	0.51	9	6025	1
Majority					True

probably genuinely describing the language. In this case, by majority consensus, the machine infers that the language Wutun [wuh] does have classifiers.

Further manual checking was still performed for languages that a) were not included in the manual survey but have been detected by the automatic survey b) were included in the manual survey and had an assessment different from that of the automatic one. The manual checking was also conducted to ensure that only languages with sortal classifiers were included. To facilitate the manual checking, the sentences containing search hits are presented by the machine next to the assessments along with direct links to the underlying documents. Manual correction was necessary to ensure high quality data in the database. In particular, for classifiers, noise could potentially have been introduced due to the variation of definition and terms for classifiers in the literature. An example is Nalik (Austronesian, New Ireland), which was initially identified by the automatic assessment as a classifier language due to the multiple occurrences of the exact term 'classifier' and the specific examples provided in the reference grammar by Volker (1998). Manual checking and subsequent examination of details, however, have determined that the language does not have sortal classifiers following our definition. As an example, one of the putative classifiers attributed to Nalik is the classifier *vi* 'crowd', as shown in (6a). Such a meaning refers to two objects or more. We thus know immediately that it cannot be a sortal classifier, which by definition must be a multiplicand with the numerical value 'one'.

(6) Classifier-like structures in Nalik (Volker, 1998, p.100,120)

a. *a* *vi* *fu-nalik*
ART crowd NSG-boy

'a crowd of boys'

b. *a* *yen* *orolavaat*
ART fish four

'the four fish'

c. *a* *vi* *yen* *orolavaat*
ART crowd fish four

'the four fish'

We then tried to determine whether *vi* is a mensural classifier, in which case its value could be anything except 'one'. Note that the putative classifiers in Nalik such as *vi* are optional; (6b) thus have exactly the same number of fish, i.e., four. If *vi* was a mensural classifier indicating a crowd, the total number of four crowds, i.e., $[4 \times n, n > 2]$, could not possibly be four. Note also that the word order in (6c) is [*vi* Noun Numeral], where the noun intervenes between *vi* and the numeral, thus ruling out the possibility of the two forming the multiplicative relation that is expected between numeral classifiers and numerals. Based on these observations, we conclude that the putative classifiers in Nalik are not numeral classifiers and should instead be treated as either semantic plural and dual markers or object-specific nouns indicating groups, like English 'group' in 'a group of four people' or 'pride' in 'a pride of six lions'.

C Supplementary Material - Data format

Three major types of variables are currently included in the data: metadata, socio-geographic annotations, and information on numeral classifier systems. First, variables related to metadata are listed as follows: Glottocode, ISO 639-3 code, and language name in Glottolog. The Glottocodes and ISO 639-3 codes are two of the most common unique identifiers found in typological studies. These two types of identifiers are thus both included. The language name as found in Glottolog is also included.[10]

Socio-geographic variables included in WACL are: Longitude, Latitude, Glottoarea, Continent, Status, and Family. Geographic information such as longitude, latitude, and Glottoarea (Africa, Australia, Eurasia, North America, Papunesia, South America) are included due to their increasing use in large-scale typological studies to control for geographic factors during statistical analyses. This information is directly extracted from Glottolog. The information of continent (Africa, Americas, Asia, Australia, Europe, Pacific) is also added, to facilitate analyses that would require geographic boundaries different from Glottoareas. An example of how languages are encoded for each variable is shown in Table 4.

Tab. 4: A sample of the data included in WACL. The variables about the subgroups of families, the continents, and the status of languages are not included due to space limitation in the text.

Glottocode	ISO	Name	CLF	Longitude	Latitude	Area	Family	Source
aghu1254	ggr	Aghu Tharnggalu	FALSE	142.426	-13.735	Australia	Pama-Nyungan	Jolly1989
ainu1240	ain	Hokkaido Ainu	TRUE	142.462	43.634	Eurasia	Ainu	Bugaeva2012
alge1239	arq	Algerian Arabic	FALSE	33.230	35.421	Africa	Afro-Asiatic	Guerrero2015
assa1263	asm	Assamese	TRUE	91.293	26.088	Eurasia	Indo-European	Ojah1995

The genealogical affiliation of each language is extracted from Glottolog. In the current version of WACL, we only include the first three levels of each language family. The status of a language is also encoded based on its Agglomerated Endangerment Status as defined by Glottolog. This status reflects how endangered a language is according to an agglomeration of the databases of The Catalogue of Endangered Languages (ELCat), UNESCO Atlas of the World's Languages in Danger, and Ethnologue. This variable includes six values: not endangered, threatened, shifting, moribund, nearly extinct, and extinct (see Hammarström et al. 2018 for details). Finally, the current version of WACL includes information on the presence/absence of numeral classifiers (more specifically sortal classifiers) in each language. The feature is currently binary, with `TRUE` marking classifier languages and `FALSE` referring to non-classifier languages. If a language has one numeral classifier, and it is a sortal classifier, the language is marked as having numeral classifiers, regardless of the obligatoriness of this classifier. The reference that was used to identify the presence/absence of numeral classifiers is included in the 'Source' column. The current display only shows one reference for each language. Additional information about groups of relevant references that have been checked and page numbers are available in the raw data and will be added in the future releases of WACL.

At the current stage, WACL includes metadata and information on the presence/absence of numeral classifier systems for 3338 languages, among which 723 are numeral classifier languages. WACL differs from existing data sources in several ways. First, it provides a precise definition and a series of morphosyntactic tests to further facilitate the identification of numeral classifiers. The content of the data has been automatically and manually checked based on the specified definition and tests, which allows readers to have a more precise and robust view of the distribution of numeral classifier languages in the world. Second, it provides a much larger dataset of numeral classifier languages, as the currently largest available data on numeral classifier

[10] Most of the metadata and socio-geographic annotations are imported from Glottolog. If an item of information is missing in Glottolog, it is also marked as NA in the current version of WACL. These missing values are generally rare. For example, 73 of the 3338 data points do not have information on their geographical location.

languages has 400 languages with 140 numeral classifier languages. The content of WACL thus provides a solid foundation for linguistic analyses. For instance, it is an adequate source of data to investigate the origin of classifiers with quantitative and/or phylogenetic methods. As an example, the presence/absence of classifiers can be tested for correlation with the presence/absence of plural marking (Cathcart et al., 2020) and the presence/absence of grammatical gender (noun class) systems (Sinnemäki, 2019) in different language families, as those systems are hypothesized to be in complementary-like distribution with classifier systems. Furthermore, the presence/absence of classifiers in specific language families can be used to reconstruct the ancestral state of classifier systems in those family and assess existing hypotheses about the origin of classifier systems in languages of the world (Her and Li, in press).