One-Soon Her, Harald Hammarström and Marc Allassonnière-Tang*

Defining numeral classifiers and identifying classifier languages of the world

Abstract: This paper presents a precise definition of numeral classifiers, steps to identify a numeral classifier language, and a database of 3,338 languages, of which 723 languages have been identified as having a numeral classifier system. The database, named World Atlas of Classifier Languages (WACL), has been systematically constructed over the last 10 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 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; nominal classification; numeral classifiers; sortal classifiers; wacl

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: 599; Lakoff and Johnson 2003:162–163). 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 of 257 languages having gender/noun class) and classifier systems (Gil 2013; 35%, 140 of 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 into how the human mind works.

*Corresponding author: Marc Allassonnière-Tang**, CNRS/MNHN/University Paris City, Lab Ecological Anthropology, Paris, France, E-mail: marc.allassoniere-tang@mnhn.fr. https://orcid.org/0000-0002-9057-642X
One-Soon Her, Department of Foreign Languages and Literature, Tunghai University, Taichung, Taiwan; and Graduate Institute of Linguistics, National Chengchi University, Taipei, Taiwan, E-mail: hero@thu.edu.tw
Harald Hammarström, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, E-mail: harald.hammarstrom@lingfil.uu.se

(1)   Examples of numeral classifiers

a. *tin*   *jana*   *manche*
    three   CLF.HUMAN   man
    'three men' (Nepali, Allassonnière-Tang and Kilarski 2020: 127)

b. *yi4*   *tiao2*   *yu2*
    one   CLF.LONG   fish
    'one fish' (Mandarin Chinese)

c. *pa-ita*   *tʃinu*
    one-CLF.ANIMAL   dog
    'one dog' (Tariana, Aikhenvald 1994: 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. 2021; 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: 408).

While grammatical gender has long been involved in linguistic studies, classifiers have been of only minor interest in linguistic theories up to the end of the twentieth 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: 37). Linguistic works during that period mostly aimed to establish 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 noted. 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 of 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 classifiers 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 7,000 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. 2021; Greenberg 1972; Greenberg et al. 1990; 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 3,000 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: 30; Bisang 1999: 113; Dixon 1986: 105; Grinevald 2000: 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: 292; Wu and Her 2021: 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, as a detailed reading of the sources shows that similar definitions are frequently used despite 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: 62–68; Seifart 2010: 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 classifier systems are divided into two main subtypes based on different semantic (and sometimes syntactic) behaviour (Her et al. 2017; Peyraube and Wiebusch 1993: 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 certain 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 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 (Aikhenvald 2000: 115; Bisang 1999: 121); 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.

(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 (1990: 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 (2a), 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*. While the three examples have the same truth value, *zhi1* in (3a) highlights the animacy of fish, *tiao2* in (3b) highlights the elongated shape, and *wei3* in (3c) highlights the tail part.

(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 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,

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

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

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 × 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: 152; Her 2012: 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 given an appropriate context. In this study, we only consider sortal classifiers, that is, 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: 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, perform morpheme division on them and translate them. We also assume that issues relating to morpheme class division (point ii) and the distinction open versus 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: 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: 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: 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: 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: 158)

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

    b. *kátʉ́: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: 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 of whether 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 (1990: 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 can both occur with nouns of low countability (i.e., nouns that typically do not occur in direct construction with numerals), 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. Our methodology has helped clarify 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 those 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 7,126 source documents describing 3,240 languages spanning all areas of the world. The manual survey and the automatic survey (see Supplementary Material B) resulted in a sample of 3,338 languages, which includes 723 numeral classifier languages.[3] Further details are provided in Section 4.

4 Results

A geographic visualization of the numeral classifier languages found in our surveys is shown in Figure 1. The data includes 723 (22%, 723/3,338) numeral classifier languages and 2,615 (78%, 2,615/3,338) languages without numeral classifiers. The data matches with the existing literature in two ways. First, numeral classifiers

---

2 These are the description types of Hammarström and Nordhoff (2011) that cover a spectrum of linguistic description types. They 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: 67–68). The precision is near 0.95 and much higher 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. (2021:32).

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 difference 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: 122). By comparing the two databases, there is a mismatch of annotation for 42 languages. Fifteen 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.

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: 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 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.

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
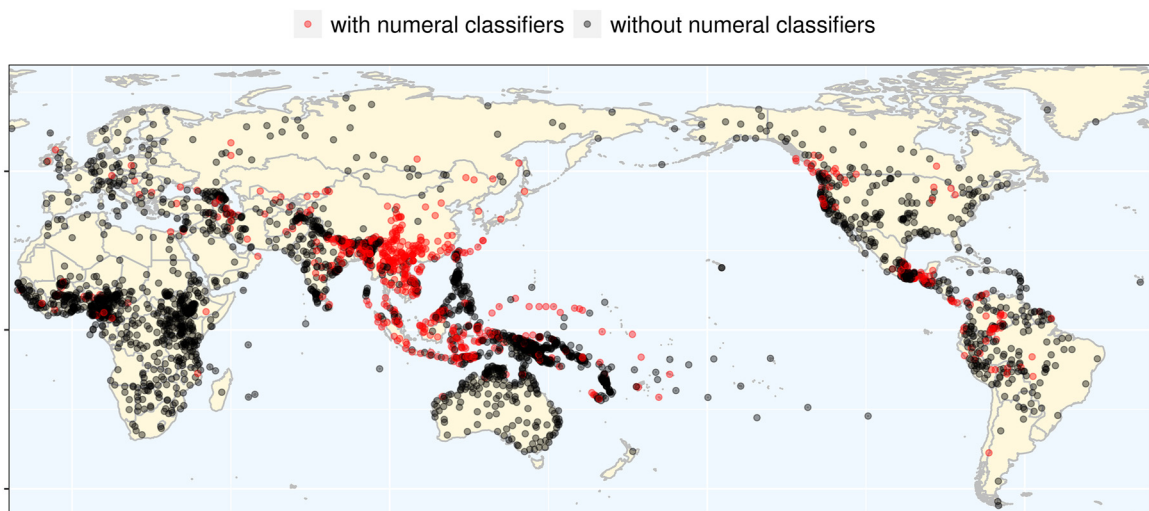

with numeral classifiers   without numeral classifiers

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

Table 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 (723/3,338), because only languages with identified coordinates are mentioned in this table.

Continent	Proportion on total		Proportion per 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%

necessary to understand how frequent numeral classifier languages are amongst the 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 many 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 divergences with observations from previous studies. As an example, some studies (Nichols 1992; Sinnemäki 2019) observe that numeral classifiers are more commonly found in the Pacific than elsewhere. We do not engage with 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 consist only 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.

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 contents 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 are needed to further enlarge and deepen our survey. Therefore, following the

---

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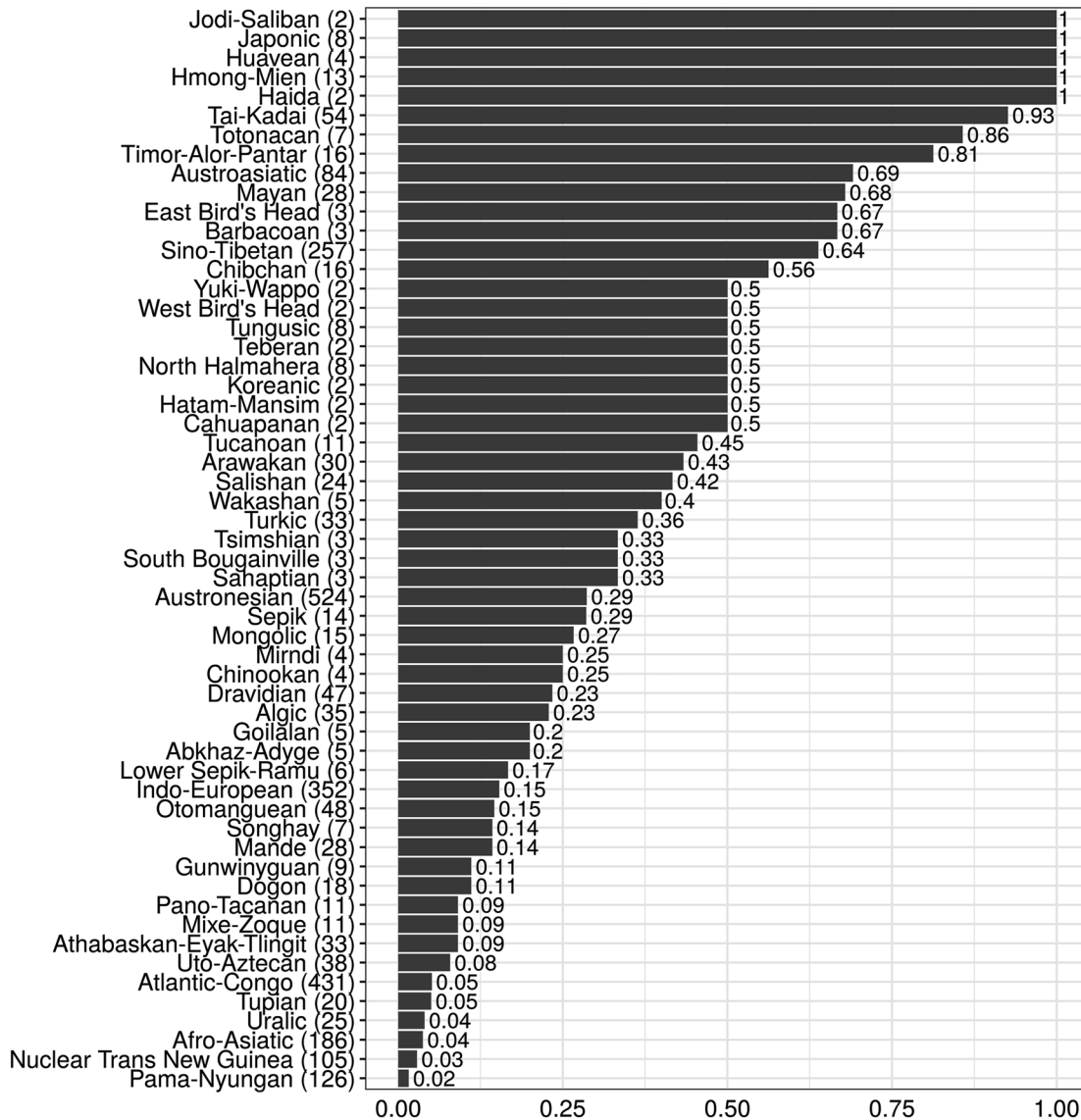
Figure 2: The proportion of numeral classifier languages per family. The numbers in parenthesis refer to the number of languages included in the data for each family. Families without numeral classifier languages or families for which there is only one data point are not listed here.

FAIR principles (Findable, Accessible, Interoperable, and Reusable), 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 contents 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 locations https://wacl.clld.org/ and http://wacl.thu.edu.tw/one. 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.

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 https://doi.org/10.1515/lingvan-2022-0006. All these files will be freely available under an open-source license. The contents of the database will also have updated releases, which will be stored in Zenodo and assigned a DOI.

References

Adams, Karen Lee. 1989. *Systems of numeral classification in the Mon-Khmer, Nicobarese and Aslian Subfamilies of Austroasiatic*. Canberra: Pacific Linguistics.

Adams, Karen Lee & Nancy F. Conklin. 1973. Toward a theory of natural classification. In Claudia Corum, Thomas C. Smith-Stark & Ann Weiser (eds.), *Papers from the ninth regional meeting of the Chicago Linguistic Society*, 1–10. Chicago: University of Chicago.

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

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

Aikhenvald, Alexandra. 2003. 4: Numeral classifiers. In *Classifiers*, 98–124. Oxford: Oxford University Press.

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

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

Allassonnière-Tang, Marc & One-Soon Her. 2020. Numeral base, numeral classifier, and noun: Word order harmonization. *Language and Linguistics* 21(4). 511–556.

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

Allassonnière-Tang, Marc, Dunstan Brown & Sebastian Fedden. 2021. Testing semantic dominance in Mian gender: Three machine learning models. *Oceanic Linguistics* 60(2). 302–334.

Audring, Jenny. 2016. Gender. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press.

Basirat, Ali & Marc Tang. 2018. Lexical and morpho-syntactic features in word embeddings: A case study of nouns in Swedish. In *Proceedings of the 10th international conference on Agents and Artificial Intelligence*, vol. 2, 663–674.

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

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

Bickel, Balthasar & Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In Peter Austin, Helen Dry & Peter Witternburg (eds.), *Proceedings of the international LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002. ISLE and DOBES*. Nijmegen.

Bisang, Walter. 1999. Classifiers in East and Southeast Asian languages: Counting and beyond. In Jadranka Gvozdanović (ed.), *Numeral types and changes worldwide, vol. 118 of trends in linguistics: Studies and monographs*, 113–186. Berlin: Mouton de Gruyter.

Blust, Robert. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics.

Boas, Franz. 1911. Chinook. In Franz Boas (ed.), *Handbook of American Indian Languages 1*, vol. 40 of Smithsonian Institution Bureau of American Ethnology Bulletin, 559–678. Washington, D.C.: Government Printing Office.

Chao, Yuenren. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

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

Contini-Morava, Ellen & Marcin Kilarski. 2013. Functions of nominal classification. *Language Sciences* 40. 263–299.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2013. Number of Genders. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world Atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, Greville G. & Sebastian Fedden. 2016. Canonical gender. *Journal of Linguistics* 52(3). 495–531.

Craig, Colette. 1986. *Noun classes and categorization*. Amsterdam: John Benjamins.

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

Csirmaz, Aniko & Eva Dekany. 2014. Hungarian is a classifier language. In Raffaele Simone & Francesca Masini (eds.), *Word classes: Nature, typology and representations*, 141–160. New York: John Benjamins.

Dekany, Eva & Aniko Csirmaz. 2017. Numerals and quantifiers. In Gabor Alberti & Tibor Laczko (eds.), *Syntax of Hungarian: Nouns and noun phrases*, 1044–1150. Amsterdam: Amsterdam University Press.

Denny, Peter. 1976. Papers from the 12th regional meeting of the Chicago Linguistic Society. In Mufwene Salikoko (ed.), *What are noun classifiers good for*, 122–132. Chicago: Chicago Linguistic Society.

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

Dixon, Robert M. W. 1986. Noun class and noun classification. In Colette Craig (ed.), *Noun classes and categorization*, 105–112. Amsterdam: John Benjamins.

Donohue, Mark. 2006. Review of the The world Atlas of language structures. *LINGUIST LIST* 17(1055). 1–20.

Dryer, Matthew S. & Martin Haspelmath. 2013. *WALS Online*. Place: Leipzig. Available at: https://wals.info/.

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

Evans, Nicholas. 2000. Word classes in the world's languages. In Geert Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphology: A handbook on inflection and word formation*, vol. 1, 708–732. Berlin: Mouton de Gruyter.

Fedden, Sebastian. 2011. *A grammar of Mian*. Berlin: Walter de Gruyter.

Fedden, Sebastian & Greville G. Corbett. 2017. Gender and classifiers in concurrent systems: Refining the typology of nominal classification. *Glossa: A Journal of General Linguistics* 2(1). 1–47.

Fedden, Sebastian & Greville G. Corbett. 2018. Extreme classification. *Cognitive Linguistics* 29(4). 633–675.

Forkel, Robert. 2014. The cross-linguistic linked data project. In Christian Chiarcos, John Philip McCrae, Petya Osenova & Cristina Vertan (eds.), *3rd Workshop on linked data in linguistics: Multilingual knowledge resources and natural language processing*, 60–66. Reykjavik, Iceland: European Language Resources Association (ELRA).

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Nature Scientific Data* 5(180205). 1–10.

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

Goddard, Cliff. 2005. *The languages of East and Southeast Asia: An introduction*. Oxford, NY: Oxford University Press.

Greenberg, Joseph H. 1972. Numeral classifiers and substantival number: Problems in the genesis of a linguistic type. *Working Papers on Language Universals* 9. 1–39.

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

University Press [First published 1972 in Working Papers on Language Universals 9. 1–39. Stanford, CA: Department of Linguistics, Stanford University.].

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

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

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

Grinevald, Colette. 2015. Linguistics of classifiers. In James D. Wright (ed.), *International encyclopedia of the social & behavioral sciences*, 811–818. Oxford: Elsevier.

Hammarström, Harald & Sebastian Nordhoff. 2011. LangDoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language* 3(2). 31–43.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.1.* Jena: Max Planck Institute for the Science of Human History. Available at: https://glottolog.org/.

Hammarström, Harald, One-Soon Her & Marc Tang. 2021. Term-spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In Simon Dobnik, Richard Johansson & Peter Ljunglöf (eds.), *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25–27 November 2020*, 27–34. Linköping: Linköping Electronic Press.

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

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

Her, One-Soon & Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11(3). 527–550.

Her, One-Soon & Wan-Jun Lai. 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, One-Soon & Marc Tang. 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, One-Soon & Li Bing-Tsiong. Nominal classification in Asia and Oceania: Functional and diachronic perspectives. In *A single origin of numeral classifiers in Asia and the Pacific: A hypothesis*. Amsterdam: John Benjamins, In press.

Her, One-Soon, Ying-Chun Chen & Nai-Shing Yen. 2017. Mathematical values in the processing of Chinese numeral classifiers and measure words. *PLoS One* 12(9). 1–9.

Her, One-Soon, Marc Tang & Bing-Tsiong Li. 2019. Word order of numeral classifiers and numeral bases. *STUF Language Typology and Universals* 72(3). 421–452.

Huffman, Franklin. 1970. *Modern spoken Cambodian*. New Haven: Yale University Press.

Hurd, Conrad. 1977. Nasioi projectives. *Oceanic Linguistics* 16(2). 111.

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

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

Kemmerer, David. 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, David. 2019. *Concepts in the brain: The view from cross-linguistic diversity*. Oxford: Oxford University Press.

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

Kilarski, Marcin. 2014. The place of classifiers in the history of linguistics. *Historiographia Linguistica* 41(1). 33–79.

Kilarski, Marcin & Marc Allassonnière-Tang. 2021. Classifiers in morphology. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*, 1–28. Oxford: Oxford University Press.

Krauss, Michael. 2015. *Eyak grammar*. Fairbanks: University of Alaska Unpublished PhD thesis.

Lakoff, George & Mark Johnson. 2003. *Metaphors we live by*. London: University of Chicago Press.

Li, Jinxi. 1924. *The grammar of Mandarin Chinese*. Beijing: Shangwu Chubanshe.

Lichtenberk, Frantisek. 1983. *A Grammar of Manam*. Honolulu: University of Hawaii Press.

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

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. *The AUTOTYP genealogy and geography database: 2013 release*. Published: Electronic database available. https://github.com/autotyp/autotyp-data (accessed 20 February 2019).

Nomoto, Hiroki. 2013. *Number in classifier languages*. Minneapolis: University of Minnesota PhD dissertation.

Nomoto, Hiroki & Hooi Ling Soh. 2019. Malay. In Alice Vittrant & Justin Watkins (eds.), *The Mainland Southeast Asia linguistic area*, 475–522. Berlin: De Gruyter Mouton.

Peyraube, Alain & Thekla Wiebusch. 1993. Le rôle des classificateurs nominaux en chinois et leur évolution historiqueun: un cas de changement cyclique. *Faits de langues* 1(2). 51–61.

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

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

Seifart, Frank. 2010. Nominal classification. *Language and Linguistics Compass* 4(8). 719–736.

Seiler, Hansjakob. 1986. *Apprehension: Language, object and order*. Tübingen: Gunter Narr.

Senft, Gunter. 2000. *Systems of nominal classification*. Cambridge: Cambridge University Press.

Simpson, Andrew, Hooi Ling Soh & Hiroki Nomoto. 2011. Bare classifiers and definiteness: A cross-linguistic investigation. *Studies in Language* 35(1). 168–193.

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

Sinnemäki, Kaius. 2019. On the distribution and complexity of gender and numeral classifiers. In Francesca Di Garbo, Bruno Olsson & Bernhard Walchli (eds.), *Grammatical gender and linguistic complexity*, 133–200. Berlin: Language Science Press.

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

Veeman, Hartger, Marc Allassonnière-Tang, Aleksandrs Berdicevskis & Basirat Ali. 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*, 265–275. Online. Association for Computational Linguistics.

Virk, Shafqat Mumtaz, Harald Hammarström, Markus Forsberg & Søren Wichmann. 2020. The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of the 12th language resources and evaluation conference*, 871–877.

Vittrant, Alice & Marc Allassonnière-Tang. 2021. Classifiers in Southeast Asian languages. In Paul Sidwell & Mathias Jenny (eds.), *The languages and linguistics of Mainland Southeast Asia*, 733–772. De Gruyter.

Wils, Jan. 1935. *De nominale klassificatie in de Afrikaansche Negertalen*. Nijmegen: Katholieke Universiteit Nijmegen PhD thesis.

Wu, Jiun-Shiung & One-Soon Her. 2021. Taxonomy of numeral classifiers. In Chungmin Lee, Young-Wha Kim & Byeong-Uk Yi (eds.), *Numeral classifiers and classifier languages: Chinese, Japanese, and Korean*, 1st edn. 40–71. London: Routledge.