Research Article

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

# Defining numeral classifiers and identifying classifier languages of the world

https://doi.org/10.1515/lingvan-2022-0006
Received January 21, 2022; accepted May 3, 2022

**Abstract:** This paper presents a precise definition of numeral classifiers, steps to identify a numeral classifier language, and a database of 3338 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 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 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

## 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: 120; Delahunty & Garvey 2010: 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: 1). This function is generally referred to as 'individualization' (Bisang 1999: 120) or 'unitizing' (Enfield 2004: 132).[1] 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.[2] As demonstrated in (1), semantically unbounded mass nouns such as 'water' cannot apply sortal classifiers (1a) and can only be quantified with mensural classifiers (1b). See Tang & Her (2019) for a theoretical and quantitative analysis on the subject matter.

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

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

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

**One-Soon Her,** Department of Foreign Languages and Literature, Tunghai University, Taichung, Taiwan/ 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
***Corresponding author: Marc Allassonnière-Tang,** Lab Ecological Anthropology, CNRS/MNHN/University Paris City, Paris, France, e-mail: marc.allassonniere-tang@mnhn.fr

| 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 have found that numeral classifiers and grammatical plural markers[3] (Tang & Her 2019). follow a complementary-like distribution cross-linguistically. Thus, different hypotheses have been developed to explain this observation (Ghomeshi & Massam 2012: 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 1977; Bale & 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 & Ouwayda 2010; Cowper & Hall 2012; Doetjes 2012; Greenberg 1990; Mathieu 2012; Nomoto 2013; Sanches & Slobin 1973; T'sou 1976; Wu & 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.

# 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 to be 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.[4] 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

[3] Grammatical plural (also called grammatical number) involves grammatical agreement outside the noun phrase. It is distinguished from semantic plural, which is only marked on the noun and relates to collective and associative marking

[4] One example is the Pipil grammar by Campbell (1985: 61): “It should be noted that unlike Proto-Uto-Aztecan (Langacker 1977: 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).”.

being described.[5] Furthermore, there is 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: 29-30). The overall generality of a document $i$ constitutes a proportion $\alpha_i$ which we hypothesize to be akin to the ratio between "signal" and "noise" in this document (Hammarström et al. 2021: 29-30). For languages where we have only one document, we may simply take average $\alpha_i$ 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 1 below with grammars and grammar sketches of Wutun [wuh]. The grammar of Sandman (2016) has an $\alpha_i$ of 0.91 and contains a total of 100624 tokens.

**Tab. 1:** 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 | $\alpha_i$ | $t$ | # 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 |

If we subtract $(1-0.91)\cdot 100624 \approx 9056$ tokens from the least frequent types, this leaves only types with frequency of four or more, defining the frequency threshold $t=4$. Each grammar has a corresponding $\alpha$ purity level as described above, the total number of tokens, and the frequency threshold $t$ induced by $\alpha_i$. 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 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, or 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, ISO639-3: nal), 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).

[5] For example, Lorenzino's (1998) 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.

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 (2a). 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'.

(2) Classifier-like structures in Nalik (Volker 1998: 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; (2b) 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 (2c) 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.[6]

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 2.

[6] 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.

**Tab. 2:** 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 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 & Li, in press).

## References

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

Bale, Alan & Jessica Coon. 2014. Classifiers are for numerals, not for nouns: Consequences for the mass/count distinction. *Linguistic Inquiry* 45(4). 695–707. 10.1162/LING_a_00170. http://www.mitpressjournals.org/doi/10.1162/LING_a_00170.

Bisang, Walter. 1999. Classifiers in East and Southeast Asian languages: Counting and beyond. In Jadranka Gvozdanović (ed.), *Numeral Types and Changes Worldwide*, vol. 118 Trends in Linguistics: Studies and Monographs, 113–186. Mouton de Gruyter.

Borer, Hagit. 2005. *Structuring Sense, part I*. Oxford: Oxford University Press.

Borer, Hagit & Sarah Ouwayda. 2010. Men and their apples: Dividing plural and agreement plural. In *Handout of a talk presented at GLOW Asia 8*, Beijing.

Bugaeva, Anna. 2012. Southern Hokkaido Ainu. In Nicolas Tranter (ed.), *The Languages of Japan and Korea*, 461–509. New York: Routledge.

Campbell, Lyle. 1985. *The Pipil language of El Salvador*. Berlin: De Gruyter Mouton.

Cathcart, Chundra A., Andreas Hölzl, Gerhard Jäger, Paul Widmer & Balthasar Bickel. 2020. Numeral classifiers and number marking in Indo-Iranian: A phylogenetic approach. *Language Dynamics and Change* 1–53. 10.1163/22105832-bja10013. https://brill.com/view/journals/ldc/aop/article-10.1163-22105832-bja10013/article-10.1163-22105832-bja10013.xml.

Chiarelli, Valentina, Radouane El Yagoubi, Sara Mondini, Patrizia Bisiacchi & Carlo Semenza. 2011. The syntactic and semantic processing of mass and count nouns: An ERP study. *PLoS ONE* 6(10). 1–15. 10.1371/journal.pone.0025885.

Chierchia, Gennaro. 1998. Plurality of mass nouns and the notion of semantic parameter. In Susan Rothstein (ed.), *Events and grammar*, 53–104. Dordrecht: Kluwer.

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

Cowper, Elisabeth & Daniel Currie Hall. 2012. Aspects of individuation. In Diane Massam (ed.), *Count and mass across languages*, 27–53. Oxford: Oxford University Press.

Delahunty, Gerald P & James J Garvey. 2010. *The English language: From sound to sense*. West Lafayette: Parlor Press.

Doetjes, Jenny. 2012. Count/mass distinctions across languages. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning, part III*, 2559–2580. Berlin: Mouton de Gruyter.

Enfield, Nick J. 2004. Nominal classification in Lao: a sketch. *STUF - Language Typology and Universals* 57(2-3). 117–143. 10.1524/stuf.2004.57.23.117. http://www.degruyter.com/view/j/stuf.2004.57.issue-2-3/stuf.2004.57.23.117/stuf.2004.57.23.117.xml.

Ghomeshi, Jila & Diane Massam. 2012. The mass count distinction: Issues and perspectives. In Diane Massam (ed.), *Count and mass across languages*, 1–8. Oxford: Oxford University Press.

Gillon, Brendan S. 1999. The lexical semantics of English count and mass nouns. In Evelyne Viegas (ed.), *Breadth and depth of semantic lexicons*, 19–37. Dordrecht: Springer.

Greenberg, Joseph H. 1990. Generalizations about numeral systems. In Keith Denning & Suzanne Kemmer (eds.), *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.].

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

Hammarström, Harald. 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*, 81–89. Stroudsburg, PA: Association for Computational Linguistics (ACL).

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.

Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg & Bettina Speckmann. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation* 12. 359–392.

Hammarström, Harald, Shafqat Mumtaz Virk & Markus Forsberg. 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* 71–75.

Hansen, Chad. 1983. *Language and logic in ancient China*. Ann Arbor: University of Michigan Press.

Her, One-Soon & Bing-Tsiong Li. 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*, Amsterdam: John Benjamins.

Janhunen, Juha, Marja Peltomaa, Erika Sandman & Dongzhou Xiawu. 2008. *Wutun* (Languages of the world 466). Germany: Lincom Europa.

Jolly, Lesley. 1989. *Aghu Tharrnggala, a language of the Princess Charlotte Bay region of Cape York Peninsula*. Brisbane University of Queensland MA thesis.

Krifka, Manfred. 1995. Common nouns: A contrastive analysis of Chinese and English. In Gregory N Carlson & Francis J Pelletier (eds.), *The generic book*, 398–411. Chicago: University of Chicago Press.

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

Lee-Smith, Mei W. & Stephen A. Wurm. 1996. The Wutun language. In Stephen A. Wurm, Peter Mühlhäusler & Darrell T. Tryon (eds.), *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*, vol. II.2, 883–897. Berlin: Mouton de Gruyter. 10.1515/9783110819724.3.883. http://www.degruyter.com/view/books/9783110819724/9783110819724.3.883/9783110819724.3.883.xml.

Link, Godehard. 1998. *Algebraic semantics in language and philosophy*. Stanford: CSLI.

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

Mathieu, Eric. 2012. On the mass-count distinction in Ojibwe. In Diane Massam (ed.), *Count and mass across languages*, 172–198. Oxford: Oxford University Press.

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

Ojah, Deepali. 1995. *A critical study of Barpeta dialect*. Assam: Gauhati University PhD Dissertation.

Quine, Willard van Ormine. 1960. *Word and object*. Cambridge: MIT Press.

Sanches, Mary & Linda Slobin. 1973. Numeral classifiers and plural marking: An implicational universal. *Working Papers in Language Universals* 11. 1–22.

Sandman, Erika. 2016. *A grammar of Wutun*. Helsinki: University of Helsinki PhD Dissertation.

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. 10.1515/flin-2019-2013.

T'sou, Benjamin K. 1976. The Structure of Nominal Classifier Systems. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic Studies Part II* Oceanic Linguistics Special Publication, 1215–1247. Honolulu: University Press of Hawaii.

Virk, Shafqat Mumtaz, Lars Borin, Anju Saxena & Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In Kamil Ekštein & Václav Matoušek (eds.), *Text, speech, and dialogue: 20th international conference, tsd 2017, prague, czech republic, august 27-31, 2017, proceedings*, vol. 10415 Lecture Notes in Computer Science, 111–119. Berlin: Springer.

Virk, Shafqat Mumtaz, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal & Nazia Khurram. 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)*, 1247–1256. Varna, Bulgaria: NCOMA Ltd.

Volker, Craig A. 1998. *The Nalik language of New Ireland, Papua New Guinea*. Bern: Peter Lang.

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*, 40–71. Routledge 1st edn. 10.4324/9781315166308. https://www.taylorfrancis.com/books/9781351679602.

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

Zhang, Niina Ning. 2012. Countability and numeral classifiers in Mandarin. In Diane Massam (ed.), *Count and mass across languages*, 220–237. Oxford: Oxford University Press.