Types of classifier systems

Although classifiers constitute one of the prototypical illustrations of linguistic diversity, our knowledge of their expression is constrained by the absence of data. Among the different types of classifiers, i.e., numeral, noun, possessive, verbal, deictic, and locative, only the distribution of numeral classifiers is illustrated in the *World Atlas of Language Structures Online* (Gil 2013) and the *World Atlas of Classifier Languages* (*WACL*) (Her et al. 2022), while only qualitative assessments are available of the other types, e.g., in Aikhenvald (2000). In this paper we report on the preliminary findings of an ongoing project that aims to determine the distribution of types of classifiers found in the world's languages, the semantic categories they express and the correlation between classifier type and semantics.

The database of classifier languages and types of classifiers will be built on sources such as *WACL*. The Gramfinder software, which gathers information from over 7000 language grammars and sketches of over 3000 languages (Virk et al. 2020), will then be used to access the published materials for each language and assess automatically and manually which types of classifiers are found in each language (Allassonnière-Tang et al. 2021; Hammarström et al. 2021). Quantitative analyses controlling for geographic area, language family, and cultural traits (Kirby et al. 2016) will be conducted at the global scale to identify the universality and specificity of each classifier type.

We have conducted a preliminary analysis of 986 languages based on a reduced sample of the available grammars. First, Gramfinder was used to count the occurrences of the term 'classifier' in the available sources. The result is displayed in Figure 1a below, where we find 651 classifier languages. Within the sources found for each language, we automatically extracted the immediately preceding word for each occurrence of the term 'classifier' to identify the main types of classifiers, as the terms for classifier types are generally used in tight combination with the word 'classifier', e.g., 'numeral classifier', 'possessive classifier'. To visualize the preliminary output of this method, Figures 1b-d show the geographic distribution of languages with the mention of the following terms in the available sources: 'numeral classifier' (Figure 1b), 'noun classifier' (Figure 1c), and 'possessive classifier' (Figure 1d). The distribution matches existing descriptions, with numeral classifiers mostly found in Asia and the Pacific, and in the Americas, while possessive classifiers are mostly found in Oceania. However, the distribution of noun classifiers is fuzzy, which infers a further need for manual checking.

While several potential pitfalls need to be considered, e.g., the need for a systematic manual check of the available sources due to the diversity of terms used for classifier types (as shown by the fuzzy distribution of noun classifiers in Figure 1c), we show that the use of available sources as corpora combined with NLP methods is a suitable tool with high potential for identifying classifier types in the world's languages.

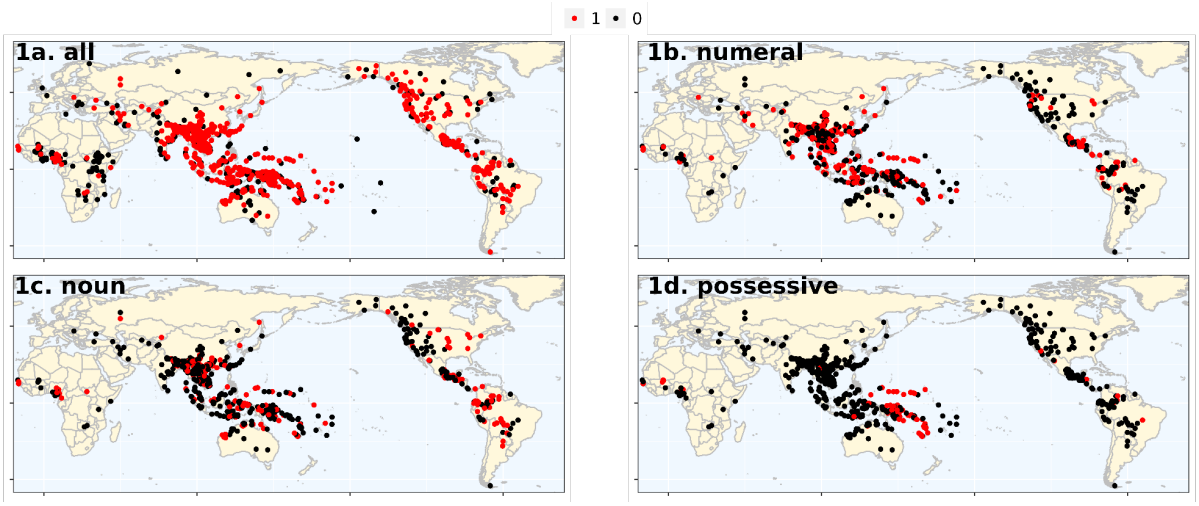

Figure 1a. Languages included in the preliminary study. Among the 986 languages, there are 651 (66.02%) classifier languages. Points in red indicate classifier languages and points in black indicate languages without classifiers. Figures 1b-d. The languages for which the terms 'numeral classifier', 'noun classifier', and 'possessive classifier' are detected within available sources. Points in red indicate classifier languages with the mention of 'numeral classifier' (43.9%, 286/651), 'noun classifier' (29.3%, 191/651), and 'possessive classifier' (11.7%, 76/651) and points in black indicate classifier languages without the mention of these terms in available sources.

References

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

Allassonnière-Tang, Marc, Olof Lundgren, Maja Robbers, Sandra Cronhamn, Filip Larsson, One-Soon Her, Harald Hammarström & Gerd Carling. 2021. 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. http://doi.org/10.1057/s41599-021-01003-5.

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. https://wals.info/chapter/55.

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 Peter Ljunglöf, Simon Dobnik & Richard Johansson (eds.), *Selected contributions from the Eighth Swedish Language Technology Conference (SLTC-2020), 25-27 November 2020*. Linköping: Linköping University Electronic Press. https://doi.org/10.3384/ecp184172.

Her, One-Soon, Harald Hammarström & Marc Allassonnière-Tang. 2022. Defining numeral classifiers and identifying classifier languages of the world. *Linguistics Vanguard* 8(1). 151-164. https://doi.org/10.1515/lingvan-2022-0006.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale & Michael C. Gavin. 2016. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11(7). e0158391. https://doi.org/10.1371/journal.pone.0158391.

Virk, Shafqat Mumtaz, Harald Hammarström, Markus Forsberg & Søren Wichmann. 2020. The DReaM Corpus: A multilingual annotated corpus of grammars for the world's languages. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 878-884. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.110.