Semantics of classifier systems

Marcin Kilarski & Marc Allassonnière-Tang
(Adam Mickiewicz University in Poznań & Lab Ecological Anthropology UMR 7206; Centre national de la recherche scientifique; Muséum national d'histoire naturelle; Université Paris Cité)

Keywords: linguistic typology, classifier systems, classifier types, semantics, corpora

While classifiers are well-known for the remarkable diversity in semantics and means of expression, our knowledge about them is constrained by the absence of data. Among the different types, i.e., numeral, noun, possessive, verbal, deictic, and locative, only the distribution of numeral classifiers is illustrated in Gil (2013) and Her et al. (2022), while other studies provide only qualitative assessments (e.g., Aikhenvald 2000). Here we report on the preliminary findings of an ongoing project aiming to determine the distribution of semantic values in classifier systems and the correlation between semantic values and classifier type.

The database of classifier languages is now constructed on sources such as Her et al. (2022). Next, Gramfinder will be used to assess the distribution of semantic values (Allassonnière-Tang et al. 2021; Hammarström et al. 2021), based on c.7000 descriptions of c.3000 languages (Virk et al. 2020). Quantitative analyses controlling for geographic area, language family, and cultural traits (Kirby et al. 2016) will be conducted to identify the distribution of semantic values, followed by an assessment of the interaction among semantic values as well as between semantic values and classifier types.

The analysis of 986 languages is based on a sample of the available grammars. Gramfinder was used to count occurrences of the term 'classifier'; the result is displayed in Fig. 1a, where we find 651 classifier languages (66.02%). Within the sources for each language, the preceding word was extracted for each occurrence of the term 'classifier' as such terms generally precede the word 'classifier', e.g., 'general classifier'. The results for 'general classifier' (24.6%, 160/651) (Fig. 1b) and 'human classifier' (10.5%, 68/651) (Fig. 1c) show that not all classifier languages have such classifiers. This unexpected distribution may result from the diversity of terms found in the literature and the presence of other phraseological contexts with relevant information, e.g., 'classifier for humans', which thus requires a systematic manual check of the sources. While such pitfalls need to be considered, we show that the use of available sources as corpora combined with NLP methods is a suitable tool for identifying the semantic values of classifiers.

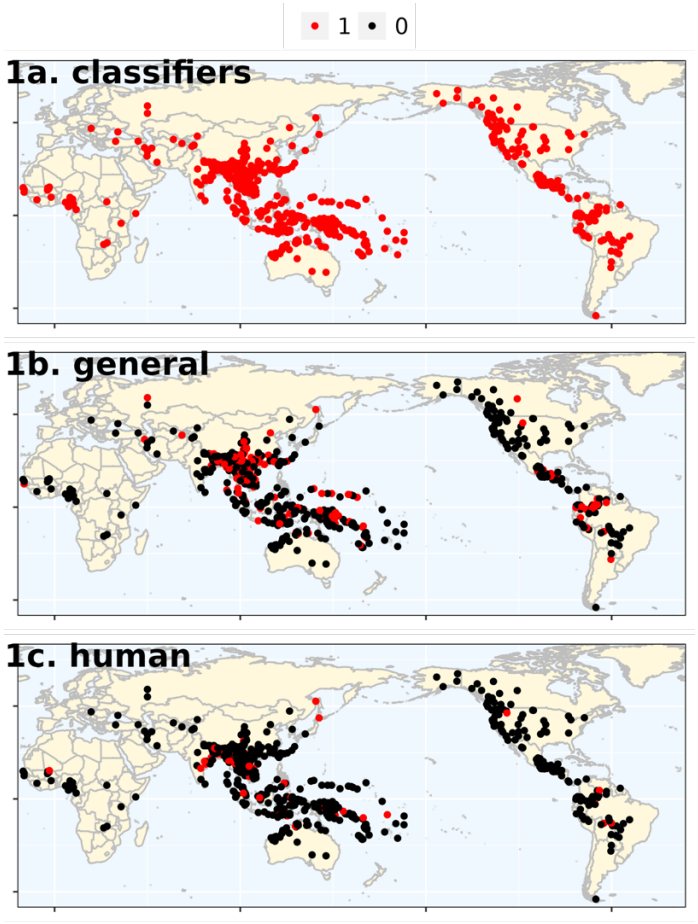

Figure 1a. Languages in the sample (red = classifier languages; black = languages without classifiers). Figures 1b-c. Languages for which the terms 'general classifier' and 'human classifier' are detected within available sources (red = classifier languages with the mention of 'general classifier' and 'human classifier'; black = classifier languages without the mention of these terms).

Acknowledgments

The project is funded by the National Science Centre, Poland (UMO-2022/47/B/HS2/02999).

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 M. S. Dryer & M. Haspelmath (eds), (2013), *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 P. Ljunglöf, S. Dobnik & R. Johansson (eds), (2021), *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*, Marseille: European Language Resources Association, 878-884. https://aclanthology.org/2020.lrec-1.110.