Appendix: Guidelines for better replicability in typology

1 Transparent documentation of the language sample

In full agreement with Harris, Hyman & Staros (2006) (and others, for that matter), we think that providing the full sample including all relevant annotations and references is key for transparency and replicability. We wish to emphasize that the full sample needs to be provided; in case later subsampling is carried out in the original study as Dryer (2018) or Berg (2020), access to the original full sample is necessary to understand and evaluate all of the later sampling decisions. As for the annotations, it may also be helpful to provide an extra document with explanations for each of the variables in case the annotations are more complex.

Moreover, full transparency about the language varieties in the sample is crucial for replicability. It is not sufficient to identify a variety as, e.g., "Otomi", since this is a cover term for a number of varieties that may differ in the relevant linguistic features (Glottolog identifies 7 child varieties under "Otomi"). A simple solution is to include glottocodes as identifiers for each language of the dataset. One may argue that some of the nomenclature and phylogeny decisions in Glottolog are questionable. Yet, it is the best classification of language varieties that we currently have, being maintained and updated according to suggestions from experts in the community. Another advantage of the inclusion of glottocodes is that it makes it easy to include additional information about the language when replicating the study or re-using the dataset. In the present study, for instance, we added geographic information on the languages in the datasets using the glottocodes. Another example would be the addition of information from WALS (Dryer & Haspelmath 2013), APiCS (Michaelis et al. 2013) or Grambank (Skirgård et al. 2023), all of which is facilitated when glottocodes are provided in the original dataset.

The last point to be addressed is the format in which the dataset, i.e. the sample with annotations and references, is made available. In theory, the CLDF format (Forkel et al. 2018) is currently viewed as the standard of dataset formats in quantitative typology. This is certainly the case for larger typological databases that have been compiled in collaborative projects over the course of years and that have been published as databases to be used by the typological community. However, we doubt that this format is used much in smaller typological studies that involve a sample built by individual researchers for a particular study. In that case, it is more practical to simply provide the dataset (sample, annotations and references) in a tabulated format, e.g. as a ".csv" or ".tsv" file.[1] Such files are stored as comma-separated values (csv) or tab-separated values (tsv) and do not depend on specific operating systems or spreadsheet software such as Microsoft Excel. Tabulated formats are practical in two more ways. They are easier to access for quick visual inspection, and they can easily be imported to programming environments such as R, which is probably the environment used by

[1] This does not mean that we argue against publishing data using the CLDF format; rather, we propose that a tabulated format should be included in addition to allow quick and barrier-free access to the dataset.

most typologists for statistical analysis.[2] This last point relates to another advantage that sharing datasets in a tabular format has over sharing word or pdf files, which is still done in publications as recent as 2023. Tabular formats ensure that when re-using the dataset for another study, additional manual steps are kept to a minimum, which avoids potential errors in the otherwise manual transfer process.

2 Transparent linguistic analysis & annotation

Already Corbett (2005) raised an important point regarding the transparency of data annotation in typology, arguing that providing the sample together with the annotation values is not sufficient. He notes the following about how replicability is ensured in the Surrey Databases:

> [...] most important, the cells of the databases do not just contain values. They do not specify only that a language has this or that phenomenon. Rather they give a link to the data. This means that the user can reproduce the link from data to claim. If it is claimed in the database that language X has suppletion in situation Y, the user can see the data on which that claim is made. If s/he has reason to analyse the data differently, then that type can be eliminated from further analysis [...] the reasoning behind a key decision is made explicit so that the reader can make informed use of a typological resource. (Corbett 2005: 19-20)

We agree that full data transparency in typological studies should go beyond providing the final annotations together with the sample. Ideally, an independent researcher should be able to inspect the original sample and evaluate not only the annotations but also the decision leading to these annotations. In order to do that, a typological study would need to include at least one example for each of the annotation decisions made in the paper.

We are aware of the fact that this may not be realistic in every case. Still, we think that this should be the default goal for typological studies in general, with compromises made as necessary. Especially dissertations and other monographs that are built around a typological study with a language sample should aim at including examples of all the languages in the sample, e.g. as an appendix. This also forces the author to be more rigorous in the annotation process, i.e. to exclude annotation choices that cannot be backed up with a clear example from the literature. However, it is not yet the standard of large-scale typological studies to include an appendix with examples from all the languages of the sample, monographs included. Besides Corbett (2005), the few examples that we are aware of are Becker (2021), Haspelmath (1997), Hein (2020), which are three typological monographs that include examples from each language of the sample to justify the annotation decisions.[3] It remains to be seen to what extent this is also a realistic goal for studies of the size of journal articles, but the latter three examples show that it certainly is for monographs.

3 Transparent and accessible code for the statistical analysis

The last component necessary for transparency and replicability is the code of a statistical analysis, if carried out. While this point has not been brought up in earlier discussions, statistical tests and

[2] Tabulated formats like ".csv" or ".tsv" can easily be imported to Python as well, for that matter.

[3] Haspelmath (1997) provides complete examples for a 40 languages sub-sample of his study.

modeling are becoming increasingly common in typology. In order to be fully replicable, the code used for the statistical analysis needs to include:

- which operating system, including which version, was used,
- which software, including which version, was used,
- which other, additional software packages, including their respective versions, were used.

Information on the operating systems, software and versions thereof is crucial for replicability, as they can all influence how the code for the statistical analysis is executed. Specific functions of a software package may work well in one operating system but not in another, which could lead to code that cannot be evaluated by another researcher, with not much of a basis for trouble shooting. Functions can also do different things across operating systems and software versions, which can affect complete replication as well. With regards to this issue, Roberts (2018: 9) notes: “Roberts et al. (2015) found that results could differ substantially between running the stats on different operating systems, due to small bugs in the code for the *lme4* package (since fixed, see Roberts et al., 2015).”

There are several solutions to tackle this particular issue. R libraries like R-env or Packrat allow researchers to easily produce a project specific state (including libraries with fixed versions), which can be shared to facilitate replication.[4] Another alternative is to use a system like docker, which in addition to sharing the R libraries, allows researchers to share the exact operating system settings used to produce the analysis. Docker is a more complete system, but, in our experience, much harder to set up than R-env and Packrat.[5]

References

Becker, Laura. 2021. *Articles in the world’s languages* (Linguistic Studies 577). Berlin: De Gruyter.

Berg, Thomas. 2020. Nominal and pronominal gender: Putting Greenberg’s Universal 43 to the test. *STUF-Language Typology and Universals* 73(4). 525–574.

Corbett, Greville. 2005. Suppletion in personal pronouns: theory versus practice, and the place of reproducibility in typology. *Linguistic Typology* 9(1). 1–23.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798–833.

Dryer, Matthew & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

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. *Scientific Data* 5(1). 180205.

Harris, Alice, Larry Hyman & James Staros. 2006. What is reproducibility? *Linguistic Typology* 10(1). 69–73.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

Hein, Johannes. 2020. *Verb doubling and dummy verb: Gap avoidance strategies in verbal fronting*. De Gruyter.

Michaelis, Susanne, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). 2013. *APiCS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

[4] We provide lock files for our R-env state in the supplementary materials.

[5] We planned to provide a docker container but there seems to be a bug that leads to a memory leak when running Stan models, which is difficult to debug.

Roberts, Seán. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology* 9.

Skirgård, Hedvig et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.