Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

How to Share Data and Code when Submitting Papers to a Journal: Transparent Data (How to do X in Linguistics 8)

Johann-Mattis List
Department of Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology

When sharing data and code when submitting papers to a journal, you need to make sure that the reviewers can test and inspect your data as conveniently as possible. In between reviews, you should also be maximally transparent on any changes that have been made to the data or the code underlying your study. When using data that was published elsewhere, this means you should pay specific attention to the versions you have used and make sure they are readily accessible.

Describe Your Data Transparently

It is not enough to just provide a dump of all the data you used when you submit your studies to a journal. In order to make it easy for everybody to reproduce your studies, you should always describe your data carefully. Typically, you can do this in a README file which you provide along with your data. In such a file, you should exhaustively describe all files that are used, explain the format, and also make clear how they relate to the research questions. Specifically when using tabular data, it is always better to be verbose rather then sparse with descriptions. Ideally, each column of your comma-separated-values (CSV) file is described in your README file. It would even be better if you provided metadata along with CSV files, such as it is recommended by the CSVW initiative, which provides guidelines how to share tabular data on the web (W3C 2015).

In any case, you should never take it for granted that people just look at your data and understand the structure immediately without any further description. All too often, however, data is exactly presented in this form: as if it was obvious what abbreviations in CSV files refer to, or as if it is clear what separator is being used in a comma-separated value file. So far, more standardization in this regard never really hurts, so I recommend everybody producing data for computational studies to take the documentation of data formats (and specifically the semantics underlying spreadsheet data) serious. You will see that the extra time you spend on explaining your data structure to other people may also help yourself, especially in those situations where you intend to reuse the data you coded a couple of years ago yourself. Similar to programming, where documentation of code is an investment in the future, documenting the structure of datasets may prevent future data loss.

Computer-Assisted Language Comparison in Practice
Volume 4, Number 8
https://calc.hypotheses.org/2877/
Published under a Creative Commons Attributions 4.0 LICENSE
Published on 01/08/2021

Using Explicit Data Versions in Your Studies

It is incredible how many different versions of certain datasets are being used and produced without scholars paying due attention to it. The famous Dyen database, for example, typically attributed to Dyen et al. (1997), which was underlying many early studies on phylogenetic reconstruction in historical linguistics, was originally presented in such a strange format (see Geisler and List 2010), that many people parsed binary cognate set information independently into nexus files, thereby creating slightly different versions that were later reused, modified, and reused, until the data finally dissapeared from the website where it was originally posted. Since scholars barely parsed the data along with lexical entries, but rather only extracted the numerical cognate information, errors could easily slip in, and it is impossible to identify where these errors happened, since the connection to the original data has been lost when people decided to throw parts of the data away.

Plans to improve on the data culminated in the IELex databse (Dunn 2012), which itself has gone offline now, but was reused in parts again in a couple of studies, each of which would use a slightly different derived version (Bouckaert et al. 2012, Chang et al. 2015). I myself created an early version for the purpose of automated cognate detection (List 2014), which has been reused several times now (Jäger et al. 2017), so that by now it is no longer clear to which degree modifications were done during publications.

Luckily, we are now in a situation where the version control system GIT is underlying all larger code repositories and scholars have started to versionize the data they produce very clearly. This means, that it is now also time for those who use the data to stop shuffling text files around, and to start working with clear versions of individual datasets.

When working with datasets not distributed by yourself, you should resist the initial temptation to suck the datasets into your private coding ecosystem and convert them to the formats you prefer for your applications. If you need to do so, you should first make sure to check if the data has an official version, and if not write down the date when you conducted your conversion. But even more importantly, you should make the conversion transparent, by providing the script that you used in order to retrieve the data in your internal formats from the original source.

If your data is shared in CLDF formats, which we have been propagating for some years now (Forkel et al. 2018), the data will typically be versionized, and the CLDF packaging format should allow you to access the data without conducting any more complex manipulations. As a result, there is typically no need to modify the data further. What you should do instead is to use the possibilities to parse CLDF data directly. If you use Python, you can for example use the pycldf software package. I have illustrated how the package can be used to access the data of the World Atlas of Language Structures (Dryer and Haspelmath 2013) in an earlier blog post (List 2021). If you use R, there is the rcldf package (Greenhill 2020), which you could give a try, or you could look at a recent study by Dediu (2021), where the CSV files of the CLDF package of the Phoible database (Moran and McCloy 2018) are accessed directly.

In any case, if you work with large datasets like WALS that have been used by many studies in the past, you should *never* make your own custom version and use it, claiming that you use the original data. Most of the larger datasets have now clear-cut versions, and an increasing amount of datasets is now also shared in the form of CLDF. So there is no excuse to use dumps extracted from scraping the CLLD websites, as it is unfortunately still done at times (Xu et al. 2020.

Summary

I often have the impression that people give much less importance to the beauty of their data than to the beauty of their wordings in their articles. This is a pity, since our data reflects our research in a much more direct way than our writings. Imagine chefs who cook incredibly tasteful meals but do not care to keep their kitchen clean. Would you still want to eat in their restaurants? And would you not rather like to know how they lead their kitchen, instead of trusting them blindly that they follow basic hygiene standards? However, similar to keeping a kitchen clean, sharing data transparently is never easy, and there may always be room for improvement.

References

Bouckaert, Remco and Lemey, Philippe and Dunn, Michael and Greenhill, Simon J. and Alekseyenko, Aalexander V. and Drummond, Alexei J. and Gray, Russell D. and Suchard, Marc A. and Atkinson, Quentin D. (2012): Mapping the origins and expansion of the Indo-European language family. *Science* 337.6097. 957-960.

Chang, Will and Cathcart, Chundra and Hall, David and Garret, Andrew (2015): Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis. *Language* 91.1. 194-244.

Dediu, Dan (2021): Tone and genes: New cross-linguistic data and methods support the weak negative effect of the "derived" allele of ASPM on tone, but not of Microcephalin. *PLOS ONE* 16.6. 1-60.

Dyer, Matthew and Haspelmath, Martin (2013): WALS Online. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://wals.info

Dunn, Michael (2012): Indo-European lexical cognacy database (IELex). http://ielex.mpi.nl/.

Dyen, Isidore and Kruskal, Joseph B. and Black, Paul (eds.) (1997): Comparative Indo-European database: File IE-data1. File IE-data1. http://www.wordgumbo.com/ie/cmp/iedata.txt.

Forkel, Robert and List, Johann-Mattis and Greenhill, Simon J. and Rzymski, Christoph and Bank, Sebastian and Cysouw, Michael and Hammarström, Harald and Haspelmath, Martin and Kaiping, Gereon A. and Gray, Russell D. (2018): Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5.180205. 1-10.

Geisler, Hans and List, Johann-Mattis (2010): Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. In: Hettrich, Heinrich (ed.): Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik. Wiesbaden:Reichert. https://hal.archives-ouvertes.fr/hal-01298493/document

Greenhill, Simon J. (2020): rcldf: Read Linguistic Data In The Cross Linguistic Data Format (CLDF). Jena: Max Planck Institute for the Science of Human History. https://rdrr.io/github/SimonGreenhill/rcldf/

Jäger, Gerhard and List, Johann-Mattis and Sofroniev, Pavel (2017): Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers. 1204-1215.

List, Johann-Mattis (2014): Sequence comparison in historical linguistics. Düsseldorf:Düsseldorf University Press.

List, Johann-Mattis (2021/02/24): How to work with WALS data in CLDF (How to do X in linguistics 5). *Computer-Assisted Language Comparison in Practice* 4.2.

Moran, Stephen and McCloy, Daniel (2019): PHOIBLE 2.0. Jena:Max Planck Institute for the Science of Human History. https://phoible.org/

W3C Consortium (2015-12-17): Model for Tabular Data and Metadata on the Web. W3C: .

Xu, Yang and Duong, Khang and Malt, Barbara C. and Jiang Serena and Srinivasan, Mahesh (2020): Conceptual relations predict colexification across languages. *Cognition* 201.104280.