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: Practical Questions (How to do X in Linguistics 7)

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

The scientific culture in linguistics has been changing recently, and more and more papers are published with code and data accompanying them. What is still often forgotten, however, is that code and data should also be shared with the reviewers during the first submission of a paper in order to guarantee a maximally transparent review process that includes also a thorough inspection of the data and the code. This calls for attention from two sides: Reviewers should make sure that they receive data and code if they are needed to replicate the results reported in a paper, while authors should make sure to submit them to the reviewers in a way that they can easily inspect them. In this new blog post series, I want to summarize what authors should keep in mind when preparing their data and code for submission to a journal. On the one hand, I hope that this post will increase awareness among colleagues that data and code should be shared upon submission. On the other hand, I hope it also provides active help to all colleagues who plan to submit an article to a journal and are not sure how to share their data in the best form.

When to Share Data and Code?

Obviously, you should share data and code with the reviewers in all those cases where your paper makes active use of data and code in order to arrive at the results. This is by now even acknowledged by the big journals, who have been slow in adjusting to open science (Nature 2018). Generally speaking, all statistical analyses, all plots of a data you make, and all operations that constitute a part of your scientific work presented in a study should be shared in such a way that your colleagues can test them in order to make sure that the results can be replicated.

Computer-Assisted Language Comparison in Practice
Volume 4, Number 6
https://calc.hypotheses.org/2782/

Published on 02/6/2021

In the past, I have ran into situations where I asked for the code as a reviewer, but the scholars who had submitted their paper were surprised, emphasizing that there was no code they had used. What they had used were some web services in order to calculate correlations, such as SocialScienceStatistics.com, and they considered it so trivial to run a correlation that they thought it was not worthwhile to provide information on this. However, since there are many ways to compute correlations, and since many colleagues still do not know how to compute them, they cannot be called standard knowledge. As a result, I would still demand that detailed information on what was done to arrive at a correlation value is provided, even if this includes an appendix that provides an Excel sheet and instructions on how to paste the data from the sheet into the website. Since sharing code is also important for the education of our colleagues, and since all people who know how to code have learned a large part about it from others who have shared their code, I consider it as very problematic if scientists refuse to be transparent in this regard.

Supplementary Material and Supplementary Information

If you share data and code, they should be easy to find in your paper (since findability is the basis of FAIR data, see Wilkinson 2016). I recommend to make your supplements findable by adding an extra section to your paper directly before the references, which you should call Supplementary Material, and which provides a link to a repository, where one can download the data. I recommend to distinguish the supplementary material from the supplementary information, which is typically additional information in the form of plots, text, explanations, which would otherwise be called an appendix, if a paper would get otherwise too long. Scholars still often confuse the supplementary information with the supplementary material and think that it is enough to share a table with data points in the form of a PDF document (and sometimes, journals even force them to do so). But the true supplementary material is usually a folder that one can download, which contains code, data, and a README file that explains the colleagues how they should interpret the data.

You can use standard wording to announce how your data is shared. Compare, for example, the way I announce my supplementary material in a study from 2019 on the detection of sound correspondence patterns (List 2019):

> The supplementary material accompanying this article contains the code and all instructions needed to repeat the experiments described in this article. The original package for correspondence pattern detection is publicly available from GitHub under https://github.com/lingpy/lingrex (Version 0.1.0). The package providing the supplementary

> material with results and instructions for running the code is also available via GitHub under https://github.com/lingpy/correspondence-pattern-paper (Version 1.1.1) and has been archived with Zenodo at https://doi.org/10.5281/zenodo.1544949.

So a very general template could look like:

> The supplementary material accompanying this article contains the code, the data, and additional instructions on how to use them in order to repeat the experiments described in this study. Data have been curated on REPOSITORY (URL, VERSION) and archived with PROVIDER (DOI). Code has been curated on REPOSITORY an darchived with PROVIDER (DOI, VERSION).

How to Share Code in Practice?

People are often afraid to share their code during peer review since they think it would unmask their anonymity. Maybe this is also the reason why code on NLP conferences is still often not submitted (which is a pity, since especially here, there are experts who should have a look at the code and check it). Thanks to the Open Science Framework (https://osf.io) and recently also Zenodo (https://zenodo.org), there are no more hurdles to share code anonyously, and I recommend all authors who need to supply data and code with their papers, since they want to support open science and transparent practices, to really make use of the great public repositories which we have by now at our disposal.

I do not recommend to submit code along with the editorial management system of the journal where you submit your article to, since journals often mess up code, and in the end you may have to resend it to the editor via email, so rather put your data on the Open Science Framework, get a read-only link in anonymized form, and place this link in an extra section of your paper, which you call “Supplementary Material and Source Code”.

If you run into problems with copyright issues when sharing your data, or your data contains sensitive information that would not be ethical to share, you should still allow reviewers to have access to the full data, because who, if not the reviewers, should have this direct access? Otherwise, if nobody else but you yourself has seen your data, you could claim to have found algorithms for anything you want. So in the case of copyright issues and sensitive information, I recommend to make sure that the data are stored somewhere independently of your computer, ideally on a public repository with closed access. This would guarantee long-term reproducibility, as sensitive data often lose their sensitivity after some decades, and even political institutions tend to open their archives after a certain amount of time. If the copyright-relevant parts of your data are not essential for the replication of the study itself, you can also share your data in anonymized form. This is what I did for a blog post on rhymed poetry, which contained many copyrighted song texts. Since only the rhyme words were important for the study at hand,

I shared the rhyme words and replaced the remaining letters by question marks (see List 2020, code and data shared here: https://gist.github.com/LinguList/159392d371a7015a307846f76f1b6c07).

Summary

Let me briefly summarize what I have tried to cover so far. If data and code play a crucial part in one's study, one must share them to conform to good scientific practice. It is crucial to share data and code already during the review process, to make sure that reviewers can also review how data and code are presented. To share data and code anonymously, there are many solutions available by now, and it is generally recommended to always deposit data and code at a repository independent from the journal, since journals have failed to provide good solutions for the archiving of data so far. To make sure reviewers and potential readers find the link to the supplementary material, you should add a section called "Supplementary Material" at the end of your paper, right before the references.

Outlook

When I began to write on this topic, I thought I could handle all aspects of data and code sharing in a single blog post. Later, I realized that the details of data and code sharing easily exceed one blog post, and — as a result — now decided to make a series, in which I will try to explain some aspects which I consider important. Having given a practical overview on the steps to undertake this time, I want to go a bit more into detail in future posts. In the following months, I therefore plan to discuss both general cases, such as dependencies and versions of data and code, and very concrete cases, such as my recommendations for scholars who share comparative data in the form of cognate sets and sound correspondences or scholars who want to share their data in the form of a CLDF dataset. People interested to weigh in are cordially invited to contribute to our blog.

References

List, Johann-Mattis (2019): Automatic inference of sound correspondence patterns across multiple languages. Computational Linguistics 1.45. 137-161.

Johann-Mattis List (2020-08-24): Constructing rhyme networks (From rhymes to networks 5). The Genealogical World of Phylogenetic Networks 9.8. http://phylonetworks.blogspot.com/2020/08/constructing-rhyme-networks-from-rhymes.html

Nature, Editorial Board (2018): Referees´rights. Nature 560. 409.

Wilkinson, Mark D. and Dumontier, Michel and Aalbersberg, Ilsbrand J. and Appleton, Gabrielle and Axton, Myles and Baak, Arie and Blomberg, Niklas and Boiten, Jan-Willem and da Silva Santos, Luiz B. and Bourne, Philip E. and others (2016): The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3.