Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Past and Future of Computer-Assisted Language Comparison in Practice

Johann-Mattis List / Annika Tjuka
Chair of Multilingual Computational Linguistics / DLCE
University of Passau / Max Planck Institute for Evolutionary Anthropology

Our blog “Computer-Assisted Language Comparison in Practice” goes into its seventh year. We reflect on the role the blog played in the past and present and new goals and concrete ideas for the future. The most drastic innovation we initiated is to turn the blog into an open journal, which means that all future and successively also past contributions will be archived in PDF format with digital object identifiers.

1 Introduction

Our blog “Computer-Assisted Language Comparison in Practice” was first published in June 2018 and has since then been publishing at least one contribution per month with a total of 87 contributions during the last six years. After such a long time of constant blogging, we thought it would be useful to review our past experience with the blog and to set new goals for its future.

When we started “Computer-Assisted Language Comparison in Practice”, we had no concrete goals in mind with respect to what we wanted to achieve with the blog. We felt that there was some need for ways to share scientific knowledge and thoughts beyond the scope of lengthy journal publications on the one hand and short tutorials or notes that one might post to different forums, on the other hand.

The introductory post stated the goal that the blog should serve to help those who wanted to learn about “coding (in Python and R), data curation and analysis, theory of diversity linguistics, and methodology of historical language comparison” (List 2018: 1). 87 contributions later, one can say that this original promise of providing tutorials on coding and data analysis in the context of comparative linguistics and computer-assisted language comparison has been kept. But at the same time, we have considerably sharpened the mission of the blog, which by now consists of three major goals:

Computer-Assisted Language Comparison in Practice
Volume 7, Number 1
URL: https://calc.hypotheses.org/6651 DOI: 10.15475/calcip.2024.1.1

Published on 17/01/2024

"Computer-Assisted Language Comparison in Practice" should offer an open space for preliminary ideas (1), it should provide a venue for young scholars to gain first publication experience (2), and it should serve as an open place for communication (3).

2 Open Space for Preliminary Ideas

With our blog, we want to give scholars a means to publish small-scale studies and ideas that can be quoted and reused by others without having to go the long way of submitting full-fledged articles to a journal. Past experience with scientific publishing has shown that the format of a web blog fulfills a special need that journals cannot and also do not want to fulfill.

We need room for ideas that are too small or too preliminary to be pursued with the energy with which one would pursue a submission to a journal with peer review. We also need room to share our particular knowledge on certain topics without having to give it away for free by posting it in some online forum where people and chatbots would use it without giving credit. Finally, we also need room to share thoughts, data, and code quickly, making them available to others to use and criticize or to serve as a proof of concept in grant applications and full-fledged follow-up studies in journals. In the past, several contributions were published that qualify as small or preliminary ideas, which turned into bigger projects later on.

As an example of the development of small ideas, Schweikhard (2018) discussed the idea of semantic promiscuity, following up on an idea shared in articles by Geisler (2018) and List et al. (2016), according to which some morphemes (or lexical roots) are more frequently used than others to coin new words. The topic was subsequently discussed as one of 10 open problems in computational historical linguistics in a series of blog posts from 2019 (List 2019, see List 2023a). Under the term lexical root productivity, the concept plays an important role in the research project "Productive Signs" (List 2023b), and in a recent study on partial colexifications, a first method to measure lexical root productivity indirectly was proposed (List 2023c).

3 Publication Experience for Young Scholars

The second goal that we identified for our blog is to allow young scholars to gain first publication experience in a friendly atmosphere. Young scholars who are in the phase of transitioning from a level where they have to write term papers to a level where they are supposed to write publications that withstand rigorous and at times hostile peer review, profit a lot from an intermediate playground where they can make first guided experience in paper writing.

In our blog, the writing of individual blog posts is guided by the editorial team who invest time in discussing wording, graphics, data, code, and style of contributions with

young authors. In contrast to anonymous peer reviewers, our team does not reject contributions right from the beginning but makes an initial assessment regarding their fit with the blog and its specific goals. If these are met, we support our team of writers in producing a nice contribution to our blog. During this process, young scholars get the unique chance to receive concrete advice on writing that would be hard to get otherwise.

With the initialization of "Computer-Assisted Language Comparison in Practice", several doctoral students have had the chance to share preliminary ideas that they could later develop into full articles or publish tutorials that were shared alongside an article. For example, two tutorials followed the publication of the Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe, Tjuka et al. 2022) describing how data can be added and compared (Tjuka 2021a, 2021b). The idea of offering help to students is not restricted to doctoral students. We managed also to successfully recruit some Master students to contribute their work by making blog posts out of term papers (Grond and Tufekci 2021) or by writing small data notes (Blum 2021) or tutorials (van Zantwijk 2023) independently of any university courses. In the future, we hope to recruit more young scholars, not only from the pool of Bachelor and Master students that we teach in our courses and from the number of doctoral students supervised in our projects, but also from the readers of our blog.

4 Open Communication

Several blog posts that were published in our blog in the last years have been written as open answers to informal questions of colleagues asking for advice. Answering questions that were brought up in personal communication has several advantages over answering only in person. First, answers may often not only be useful for the person asking, but also for a broader range of people. Second, by providing an answer in the form of a blog post, it can be readily quoted by those who make use of the solution in their own work, which may increase the motivation of the persons who answer to share their knowledge. Third, since it is quite common to trade know-how against collaborations in science, writing a quick solution for partial problems of a larger project allows scholars to provide help without having to formally agree with the results of a study in which they officially collaborate.

An example of a very important contribution in the blog, Tjuka (2020) explained in due detail, how concept lists can be added to the Concepticon project (https://concepticon.clld.org, List et al. 2016). With the publication of this blog post, the introduction by Tjuka has become the standard starting point for student assistants that help us in expanding and correcting entries in the Concepticon project. It would, of course, have also been possible to write the tutorial for internal purposes only. However, publishing the tutorial with "Computer-Assisted Language Comparison in Practice"

emphasized the inclusive idea of the Concepticon project, and individual feedback we have received on the blog post has shown that it was a wise decision to go for an open tutorial.

5 New Chances and Challenges for CALCiP

Scientific blogging has many advantages and fulfills an important need in open science, by offering scholars alternative ways to present and test their ideas, gaining first-publication experience, and providing help to their colleagues by sharing their skills.

A particular problem of blog posts, however, is still the question of how to recognize them. In our work, we have seen a broad range of attitudes towards the value of blog posts. While some scholars were very excited about ideas shared in some posts and asked for permission to use some of our material shared in the form of figures in talks and teaching, there were also situations where reviewers explicitly asked to delete references to blog posts, expressing concern about their scientific quality.

Given that the awareness, that scientific work is not only limited to published articles and books, and new ways of increasing the possibility of recognizing scientific contributions are now being actively discussed (Burton et al. 2023), we thought it is time to address the problem of recognizability concerning "Computer-Assisted Language Comparison in Practice" more concretely. While our past solution consisted of posting PDF versions of the blog with open preprint archives, such as Humanities Commons (https://hcommons.org), we thought that it might be a good idea to go a step further from 2024 on.

As a result of these considerations, from this year on, "Computer-Assisted Language Comparison in Practice" will be an open journal. Supported by the University Library of the University of Passau, we will use the Open Journal Systems software framework to host all articles that are written each year in two issues that will be published in the form of floating releases. From then on, all contributions to our blog will also be contributions to the CALC-IP journal. They will be publicly archived via the University Library of the University of Passau and digital object identifiers will be registered. This guarantees that articles can be properly quoted and also easily found through search engines and other means of internet query.

Past contributions to our blog will also be published as part of the newly founded journal. Starting with the two issues from 2023, which are already available, we will try to add all past issues throughout 2024, proceeding from the most recent ones to the earliest contributions, year by year.

With this step, we hope to strengthen the role that small publications play in scientific endeavors, contributing actively to the further development of open and transparent research, while at the same time providing the basis for young scholars to test their ideas

in a friendly atmosphere. The blog "Computer-Assisted Language Comparison in Practice" will exist in the same form in which it was published until now. In addition, however, we will link the digital object identifiers of all individual contributions as well as PDF versions of all blog posts to the website of the new and free open access journal "Computer-Assisted Language Comparison in Practice" (https://ojs3.uni-passau.de/index.php/calcip/).

References

Blum, Frederic (2021): Data gathering in times of a pandemic: Upcycling Constenla Umaña's data on the Chibchan, Lencan and Misumalpam language families. Computer-Assisted Language Comparison in Practice 4.5. 2751. URL: https://calc.hypotheses.org/2751

Burton, Kath and Cocks, Catherine and Russell, Bonnie (2023): Recognizing and harnessing the transformational power of persistent identifiers (PIDs) for publicly-engaged scholars. Information Services & Use 43.3–4. 381–386. DOI: 10.3233/isu-230212

Geisler, Hans (2018): Sind unsere Worter von Sinnen? Uberlegungen zu den sensomotorischen Grundlagen der Begriffsbildung. In: Kazzazi, Kerstin and Luttermann, Karin and Wahl, Sabine and Fritz, Thomas A. (eds.): Worte uber Worter. Festschrift zu Ehren von Elke Ronneberger-Sibold. Tubingen:Stauffenburg. 131-142.

Grond, Fiona R. and Tufekci, Ayten (2021): Computer-assisted comparison fo Gelong and Hlai using Cross-Linguistic Data Formats. Computer-Assisted Language Comparison in Practice 4.7. 2827. URL: https://calc.hypotheses.org/2827

List, Johann-Mattis and Cysouw, Michael and Forkel, Robert (2016): Concepticon. A resource for the linking of concept lists. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation. 2393-2400. URL: https://concepticon.clld.org

List, Johann-Mattis and Pathmanathan, Jananan Sylvestre and Lopez, Philippe and Bapteste, Eric (2016): Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. Biology Direct 11.39. 1-17. DOI: 10.1186/s13062-016-0145-2

List, Johann-Mattis (2019): Typology of semantic promiscuity (Open problems in computational diversity linguistics 10). The Genealogical World of Phylogenetic Networks 6.11. DOI: 10.17613/s5r7-wy64

List, Johann-Mattis (2023a): Open problems in computational historical linguistics [version 1; peer review: awaiting peer review]. Open Research Europe 3.201. 1-13. DOI: 10.12688/openreseurope.16804.1

List, Johann-Mattis (2023b): Productive Signs. A computer-assisted analysis of evolutionary, typological, and cognitive dimensions of word families [Research Project, 2023-2027]. Passau: Chair of Multilingual Computational Linguistics. DOI: 10.3030/101044282

List, Johann-Mattis (2023c): Inference of partial colexifications from multilingual wordlists. Frontiers in Psychology 14.1156540. 1-10. DOI: 10.3389/fpsyg.2023.1156540

Schweikhard, Nathanael E. (2018): Semantic promiscuity as a factor of productivity in word formation. Computer-Assisted Language Comparison in Practice 1.11. URL: https://calc.hypotheses.org/1169

Tjuka, Annika (2020): Adding concept lists to Concepticon: A guide for beginners. Computer-Assisted Language Comparison in Practice 3.1. URL: https://calc.hypotheses.org/2225

Tjuka, Annika (2021a): Adding data sets to NoRaRe: A guide for beginners. Computer-Assisted Language Comparison in Practice 4.8. URL: https://calc.hypotheses.org/2890

Tjuka, Annika (2021b): Comparing NoRaRe data sets: Calculation of correlations and creation of plots in R. Computer-Assisted Language Comparison in Practice 4.11. URL: https://calc.hypotheses.org/3109

Tjuka, Annika and Forkel, Robert and List, Johann-Mattis (2022): Linking norms, ratings, and relations of words and concepts across multiple language varieties. Behavior Research Methods 54, 864–884. DOI: https://doi.org/10.3758/s13428-021-01650-1

van Zantwijk, Mathilda (2023): Five recommendations for creating spreadsheets. Computer-Assisted Language Comparison in Practice 6.2. 1-4. URL: https://calc.hypotheses.org/6573

Acknowledgements

We thank all those who have supported the blog Computer-Assisted Language Comparison in Practice, both by reading, recommending, and quoting our work, and by actively contributing blog posts.

Funding Information

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (Grant agreement No. 101044282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.