Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

# Data Gathering in Times of a Pandemic: Upcycling Constenla Umaña's Data on the Chibchan, Lencan and Misumalpam Language Families

Frederic Blum
Humboldt-Universitat zu Berlin

While searching for the topic of a small research project about the linguistic history of South America, I realized that a lot of data that is crucial for assessing central arguments is not openly available, but new data is difficult to come by these days. And when it is, it is not usually presented in data format that allows for easy reuse. Guided by these thoughts, I decided to turn towards the upcycling of previously published data (also called **retro-standardization**, see for example Geisler et al. (forthcoming) on the upcycling of the TPPSR dataset, https://tppsr.clld.org). The dataset I chose was previously published by Adolfo Constenla Umaña (2005). In this article, the author investigated comparatively the long-claimed genealogical relationship of three families of Central and South America, Chibchan, Lencan and Misumalpam (Lehmann 1920).

The main conclusion from the original article is that the languages are indeed part of one larger family. Apart from the explicit discussion of some cognates regarding their importance to the family relationships, Constenla Umaña presents a 110-item concept list and provides cognates within those concepts for the 25 languages of the study. In a concluding chapter, the author presents a lexicostatistical and glottochronological analysis of his data and proposes a corresponding genealogical tree of the families involed based on this results.

*Computer-Assisted Language Comparison in Practice*
Volume 4, Number 5
https://calc.hypotheses.org/2751/

Published on 17/05/2021

Even though some of his claims are contested by more recent studies (Pache 2018), the dataset forms a valuable point of reference for discussing the history of the Chibchan language family. Given the geography within the bottleneck between North and South America, the hope to identify crucial aspects of South American history by assessing the past of the Chibchan linguistic history and related population movements is not unjustified.

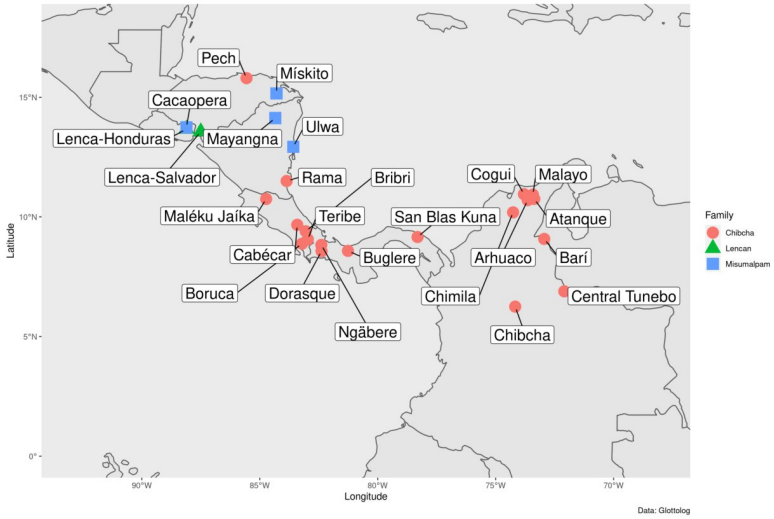


Figure 1: A map of the languages in the dataset.

In the remainder of this post, I will present my experience of upcycling this dataset to CLDF and discuss some challenges and future outlooks regarding this project. All in all, I am quite happy with the dataset. A preliminary version has been released on GitHub and I am looking forward to further working with this data.

## The Original Dataset

The dataset provided by Constenla Umaña is suited perfectly for applying the CLDFBench-workflow (Forkel and List 2020), as it adhered to certain points of FAIR standards from the very start. Beginning with the explicit publishing of all of his data, the author gave a close phonetic representation of each form where ever possible. Additionally, cognate sets for each concept were presented along the form and some of the decisions have been made transparent by discussing the arguments for or against the addition of a form to a certain set of cognates. For representing the dataset in CLDF, the following steps were necessary:

1. Creating a CSV-file out of the original list of forms
2. Linking the data to Glottolog (Hammarstrom et al. 2020)
3. Uploading the concept-list to Concepticon (List et al. 2020)

4. Creating orthography profiles for each language (see Moran and Cysouw 2018 on the specifics of orthography profiles)
5. Running CLDFBench (Forkel and List 2020)

I will now lead through each of those steps.

## Starting the Data Wrangling

The first and most laborious task was transferring the non-machine-readable list to a CSV-file for further processing. This is the downside of the explicit representation of phonemes, as there were various orthographic symbols which made an automatic processing impossible. As such, all 2643 data points had to be checked manually, making this was by far the most tedious part of the whole workflow. Working through this step illustrated to myself how useful the machine-readable publication of source data really is, first by opening up the possibility of working with the data, but then being limited in my work due to the image representation of the forms.

**1. Agua**

**LS** wal (a), **LH** was (a), **Ca** li (b),**Su** was (a),**Ul** was (a), **Mi** li (b), **Pa** àsò, **Ra** si: (c), **Gua** ti: (c), **Bor** díʔ (c), **Bri** dĭʔ (c), **Cab** díkɻ́ó (c), **Te** dí (c), **Mo** ɲʏ (c), **Boc** tʃi (c), **Dor** *yi* (c), **Cu** ti: (c), **Co** ni (c), **Ica** dʒe (c), **Da** 'dʒɪra (c), **Chi** ditake: (c), **Mu** sie (c), **Tun** 'ɾiʔa (c), **Ba** sĩ: mã (c)

**2. Amarillo**

**LS** ku, **LH** suninga, **Ca** maju, **Su** lalahni (a), **Ul** lalahka (a), **Mi** lalalhni (a), **Pa** sè: wa, **Ra** nuknukŋa, **Gua** ɬaxa:ɾa ɬu:tu iŋa:, **Bor** ʃòsát,**Bri** tsikidĭdĭ (b), **Cab** tsikidí (b), **Te** ʃòinóɾ, **Mo** ɯbrɯɾe, **Boc** moloi, **Dor** *utká*, **Cu** kollokwa, **Co** kaʃiku'ama (c), **Ica** 'tʃəmmi (ch), **Da** kɨʃkwama (c), **Chi** tʃoŋgragwattu, **Mu** tiβan (ch), **Tun** ta'w̃aja (ch), **Ba** kanikã: siũkdu˚.

Figure 2: Cognate assignment within concepts by the original author.

The second step was to create numeric identifiers for the cognate sets. In the original publication, letters from (a) to (f) were used to identify cognates within concepts. In order to make it possible to create individual handles, it was necessary to create an ID corresponding to each cognate set. I decided to do so by creating a lookup-table for each combination of letters and concepts using the programming language R, but any other programming language could have been used just as easy. I then combined this table with the original list to create my raw input for the CLDFBench-workflow.

## Linking the Data to Glottolog and Concepticon

This part was actually a lot easier than expected. While Glottolog already provided nearly all of the necessary data and only coordinates for some languages had to be added to create a complete CSV-file with information about all of the languages in the dataset, uploading and linking the concept list to Concepticon was facilitated by the corresponding workflow provided in a previous blogpost (Tjuka 2020) and some individual support by Mattis List. The blogpost also led me through the necessary steps for mapping the concept list to the Concepticon identifiers. This made it quite a pleasant experience experience working with CLDF, as all steps were accompanied by the relevant tutorials and detailed workflows.

The final step in the workflow was the creation of individual orthography profiles for each language. As the orthographical convention already corresponded reasonably well to IPA, not a lof of changes were necessary, even though they may have to be refined in the future after being compared with other sources.

Once all this was done, the path to my first CLDF-dataset was quite straightforward. Running the CLDF-workflow takes the raw-input (the CSV-file with cognate ID's), maps it against the Glottolog and Concepticon tables and creates the corresponding CLDF-files. It is now possible to run further analysis with the data, such as a comparative analysis of sound correspondences or a phylolinguistic analysis of the cognates provided by the author.

## What's Missing

Now that the basic upgrade to CLDF is done, I can turn towards the more fine-grained details of the dataset. A possible next step will be the identification of derivational morphemes within the forms given by Constenla Umaña. While some are explicitly mentioned in his discussion, others will have to be identified individually. This will then open up the possibility of improving cognacy not only by relations within, but also between concepts. Additionally, the addition of further languages and concepts is something I have in mind for the future.

Another big topic that will need to be consulted with an expert of the languages involved is the status of borrowings. Not a single form has been claimed by the original author to represent a direct borrowing from another language, which seems quite surprising considering the size of the dataset. It will definitely be necessary to look deeper into this question.

## Interpreting the Data

Running a phylolinguistic analysis given the cognates by the original author clearly shows that one cannot rely on the family tree based on the earlier lexicostatistical analysis. While some known relationships have been captured fairly well by the data, most clades only have one thing in common: uncertainty. If the cognates stand, however, a reassessment of the family relationships between and within those languages might have important consequences on the structure of the family trees and especially the position of Pech, a language spoken in north-eastern Honduras. A big downside for phylogenetic analyses in historical linguistics right now is the lack of recognition of borrowings so far. However, the dataset could potentially be used as an additional starting point for future research on the family relationships discussed by Constenla Umaña.

After the upcycling and interpretation of the data, it become clear to me how the automated processing as well as the access to open and accessible datasets can aid our pursuit for knowledge and provide improvements for our workflows. But in the end, our scientific hypothesis stands or falls with the assessments about the data we make as linguists, and there is still much work to do in this area.

## Dataset

Constenla Umaña, Adolfo (2005). ¿Existe relación genealógica entre las lenguas misumalpas y las chibchenses? Estudios de Linguística Chibcha, Vol. 23, p. 7-85.

Link to dataset on GitHub:
https://github.com/lexibank/constenlachibchan/releases/tag/v0.2

## References

Forkel, R, and List, J.-M. (2020). “CLDFBench: Give your cross-linguistic data a lift.” Proceedings of The 12th Language Resources and Evaluation Conference.

Geisler, H., Forkel, R. and List, J.-M. (forthcoming). A digital, retro-standardized edition of the Tableaux Phonétiques des Patois Suisses Romands (TPPSR). In: Avanzi, M., N. LoVecchio, A. Millour, and A. Thibault (eds.): Nouveaux regards sur la variation dialectale. Éditions de Linguistique et de Philologie: Strasbourg. 1-21. https://doi.org/10.17613/x8yd-5y42

Hammarstrom, H., Forkel, R. and Haspelmath, M. & Bank, S. (2020). Glottolog 4.3. Jena: Max Planck Institute for the Science of Human History. https://glottolog.org

Lehmann, W. (1920). Zentral-Amerika. Berlin: Verlag Dietrich Reimer.

List, J.-M., C. Rzymski, S. Greenhill, N. Schweikhard, K. Pianykh, A. Tjuka, M. Wu, C. Hundt, T. Tresoldi, and Forkel, R. (2020). Concepticon. A resource for the linking of concept lists. Version 2.4.0. Version 2.4.0. Max Planck Institute for the Science of Human History: Jena. https://concepticon.clld.org

Moran, S. and Cysouw, M. (2018). The Unicode cookbook for linguists. Managing writing systems using orthography profiles. Language Science Press: Berlin. http://langsci-press.org/catalog/book/176

Pache, M. (2018). Contributions to Chibchan historical linguistics. Doctoral dissertation, Leiden University.

Tjuka, A. “Adding concept lists to Concepticon: A guide for beginners,” in Computer-Assisted Language Comparison in Practice, 29/01/2020, https://calc.hypotheses.org/2225.