Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Adding Standardized Transcriptions to Panoan and Tacanan Languages in the Intercontinental Dictionary Series

John Miller[1] and Johann-Mattis List[2]
[1] Artificial Intelligence, Pontificia Universidad Católica del Perú
[2] Chair for Multilingual Computational Linguistics, University of Passau

In this study, we illustrate how standardized phonetic transcriptions can be added to the data for Panoan and Tacanan languages provided by the Intercontinental Dictionary Series. The result is presented as a new dataset that keeps reference to the original data and adds phonetic transcriptions for each word form in Panoan languages, Tacanan languages, as well as Spanish and Portuguese.

1 Introduction

Data-driven disciplines such as historical linguistics and linguistic typology have a long tradition in modifying existing datasets in various forms in order to reuse them in new analyses. For computational and computer-assisted approaches to historical and typological language comparison, the retrostandardization of legacy datasets has become one of the most important sources to increase the amount of digitally available datasets that can be analyzed or used to test existing and new methods (Geisler et al. 2021, Forkel et al. 2024). With the recent publication of the Lexibank repository (List et al. 2022, https://lexibank.clld.org) detailed proposals for the standardization of lexical datasets in concordance with the Cross-Linguistic Data Initiative (Forkel et al. 2018, https://cldf.clld.org) have been made and already been applied for the targeted collection of datasets for particular regions (Blum et al. 2024).

The core idea of the standards proposed by Lexibank is to link the major dimensions of the linguistic sign — its form, its meaning, and its language (Gevaudan 2007) — to three reference catalogues — Cross-Linguistic Transcription Systems (CLTS) for phonetic transcriptions (List et al. 2024a, https://clts.clld.org), Concepticon for lexical

Computer-Assisted Language Comparison in Practice
Volume 7, Number 2
URL: https://calc.hypotheses.org/7503 DOI: 10.15475/calcip.2024.2.3

Published on 02/09/2024

glosses (List et al. 2024b, https://concepticon.clld.org), and Glottolog for languages (Hammarstrom et al. 2024, https://glottolog.org). By linking individual data points of published lexical datasets (typically provided in the form of a wordlist in which a list of concepts has been systematically translated into one or more language varieties) consistently to these three catalogs, data can be easily aggregated from various sources that have been originally compiled independently of each other.

With a concept list of up to 1310 items translated into more than 300 different language varieties, the Intercontinental Dictionary Series (IDS) by Key and Comrie (last updated in 2023, originally published much earlier, see https://ids.clld.org) is one of the largest multilingual wordlist collections currently available. While the dataset has proven useful and essential to build the Database of Cross-Linguistic Colexifications, its major drawback has been for a long time that the data was not standardized regarding the word forms provided as translations for the individual concepts. The database mixes different kinds of representations for word forms, ranging from standard orthographies via phonemic transcriptions up to phonetic transcriptions in different transcription systems.

In List (2023), a new version of the IDS data was presented in which standardized phonetic transcriptions were semi-automatically provided for all word forms with the help of a large orthography profile (see Moran and Cysouw 2018) that would segment the original forms into units representing individual sounds and later convert these units into phonetic transcriptions following the standardized version of the IPA proposed by the CLTS initiative. While useful for the original purpose of the study, the new data itself proved problematic in various ways, since many wrong transcriptions had been slipped into the derived dataset, since data in standard orthography was not consistently distinguished from data in different kinds of phonetic or phonemic transcriptions.

Given that the IDS can be divided into subsets of language varieties compiled by individual scholars for languages of particular regions, it became clear that the approach used by List (2023) in trying to standardize the transcriptions by one individual orthography profile alone, was not sufficient. What one would have to do instead would be to select individual languages from the data and to provide language-specific orthography profiles (as we have illustrated recently for the Linguistic Survey of India, see Forkel et al. 2024, https://lsi.clld.org).

In the following, we show how this approach can be used to add standardized phonetic transcriptions to a subset of South-American language varieties of the IDS, mostly taken from Panoan and Tacanan languages.

2 Materials

The starting point of the derived dataset presented here is the CLDF version of the Intercontinental Dictionary Series which is curated on GitHub

(https://github.com/intercontinental-dictionary-series/ids), archived with Zenodo (Version 4.3, https://doi.org/10.5281/zenodo.7701635), and deployed as a CLLD application (https://ids.clld.org). From this dataset 22 language varieties were selected, including 6 Panoan languages, 5 Tacanan languages, four isolates and four varieties from other language families spoken in closer proximity to the Panoan and Tacanan languages, as well as Portuguese and Spanish. The letter two were selected in order to allow us to search systematically for borrowings from dominant donor languages in the region, using methods we had tested on smaller datasets before (Miller and List 2023).

3 Methods

We use CLDFBench to create datasets in CLDF (Forkel and List 2020, https://pypi.org/project/cldfbench) along with the PyLexibank plugin that allows us for a targeted handling of phonetic transcriptions and lexical glosses. The conversion workflow consists of two major stages. In the first stage, we extract the relevant data from the CLDF version of the IDS dataset in CLDF (this phase is also called "download" in CLDF data conversion workflows, supported by a command with the same name in the CLDFBench package). In the second stage, the extracted IDS data — restricted to those parts that provide data for the 22 varieties in our sample — is converted to CLDF using the Lexibank workflow, in which orthography profiles are systematically applied to add standardized phonetic transcriptions to the data (this stage is the proper CLDF creation, supported by a command called makecldf in CLDFBench).

Orthography profiles were created semi-automatically. In a first step, we used a command provided by the PyLexibank package to derive a draft profile. This command (called `initprofile` in PyLexibank) uses LingPy (Forkel and List 2024, https://pypi.org/project/lingpy) to segment the input forms automatically and to assemble identical segments in a draft orthography profile along with their potential counterparts in IPA transcriptions supported by CLTS. In a second step, we used custom code to extract individual profiles for all individual languages in our sample. These were then systematically corrected in a third step. While language-specific orthography profiles could be determined in a rather straightforward way for the individual South American languages in the sample, transcriptions for Spanish and Portuguese were elicited from the PONS dictionary (PONS Company 2023) and added for all forms in the sample.

The resulting data and code are curated on GitHub (https://github.com/intercontinental-dictionary-series/keypano) and archived on Zenodo (https://doi.org/10.5281/zenodo.13234493). To test and run the conversion, all one has to do is to install the required dependencies, to download the reference catalogs, and to run the conversion code from the command line.

```
$ git clone https://github.com/intercontinental-dictionary-series/keypano
$ git clone https://github.com/concepticon/concepticon-data
$ git clone https://github.com/glottolog/glottolog
$ git clone https://github.com/cldf-clts/clts
$ cd keypano
$ cldfbench lexibank.makecldf --concepticon=../concepticon-data
--concepticon-version=v3.3.0 --clts=../clts/ --clts-version=v2.3.0
--glottolog=../glottolog --glottolog-version=v5.0
```

Running this code will recreate the CLDF data and store the data in the folder cldf of the `keypano` folder, overwriting previous data in the same `cldf` folder. All orthography profiles for individual languages can be found in the folder `etc/orthography/` in the CLDF dataset. The language selection can be found in the file `etc/languages.tsv`.

4 Examples

Having converted the data to a CLDF dataset, we can test some of the basic properties of the data. We can, for example, use the csvkit package (https://pypi.org/project/csvkit) to count the number of unique segmented forms in the data. The first of the following two commands counts the unique number of words in segmented form, the second counts the number of unique forms (not standardized) in the data.

```
$ csvcut cldf/forms.csv -c Segments | sort -u | wc -l
18882
$ csvcut cldf/forms.csv -c Form | sort -u | wc -l
19097
```

We can see from the output, that the difference is not that great between the two perspectives on the data. With the standardized segments, we have 215 unique entries less, which indicates that the segmentation contributes to the representation of the data across languages, while showing at the same time, that the differences are not grave, and that the original distinctiveness of linguistic forms has most likely been preserved, while we win a much richer annotation of phonetic sequences.

As a second example, we can quickly compute lexical similarities between all languages in the sample by searching automatically for cognates and plotting the results as a phylogenetic tree, using the UPGMA algorithm by Sokal and Michener (1958). This can be done with the help of LingPy and the SCA method for automated cognate detection (see List 2014 for details).

```
from lingpy import LexStat

lex = LexStat.from_cldf("cldf/cldf-metadata.json")
lex.cluster(method="sca", threshold=0.45, ref="cogid")
lex.calculate("tree", tree_calc="upgma")
print(lex.tree.asciiArt())

lex.output("tsv", filename="keypano")
```

The resulting rooted tree is shown in Figure 1. As can be seen from the figure, the Panoan and Tacanan languages are clustered into one group, as are Spanish and Portuguese, while isolates are scattered over the tree. However, the results should not be taken seriously, as they merely serve to illustrate the usefulness of standardized phonetic transcriptions for computer-assisted approaches. First, the method for cognate detection used is rather simple, ignoring regular sound correspondences. Second, the method for phylogenetic tree reconstruction is also very simple, being based on distances, assuming regular divergence among languages over time. Third, the number of concepts per language is very large and not restricted to basic vocabulary, which increases the number of borrowings being falsely reported as cognates and shows that it may be useful to complement such studies later with explicit methods for borrowing detection (such as shown in Miller and List 2023).

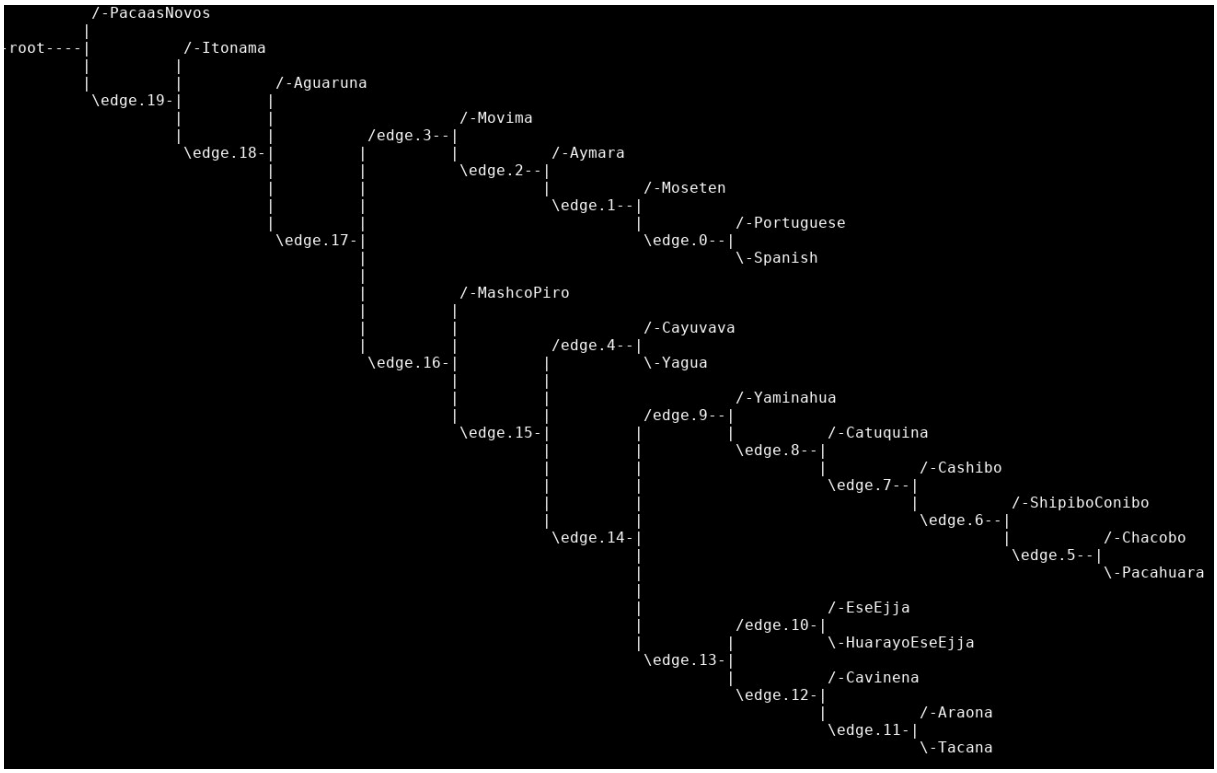

Figure 1: Phylogenetic tree inferred with the UPGMA algorithm.

That there are several borrowings in the data, both among South-American languages and specifically also from Portuguese and Spanish, can be easily seen when inspecting the data with the help of computer-assisted tools like the EDICTOR application that allows us to inspect standardized datasets interactively (List 2017, List and van Dam 2024a, https://edictor.org). Opening the file `keypano.tsv`, that we produced with the help of LingPy above, for example, allows us to inspect individual cognate judgments made by the SCA method and to check for the regularity of correspondence patterns attested in the data (see List and van Dam 2024b for details on the new functionalities in EDICTOR 3).

Thus, when opening the EDICTOR app at its base URL (https://edictor.org), and loading the file `keypano.tsv`, we can easily carry out phonetic alignments and search for correspondence patterns. Having opened the file by clicking on GETTING STARTED in the EDICTOR landing page, all we have to do is to drag the file to the button that allows to BROWSE the data. To carry out alignment analyses and to search for correspondence patterns, we then only need to add two columns to the data, by typing first ALIGNMENT into the "add column" field of the app, then pressing ENTER, and then repeating the same with PATTERNS.

Having done so, we can let EDICTOR automatically compute alignments first (by selecting COMPUTE → ALIGNMENTS) and then also search for correspondence patterns (selecting COMPUTE → CORRESPONDENCE PATTERNS). When inspecting these patterns, EDICTOR will automatically order the languages alphabetically, with Aguaruna (a Chicham languages that is the sole representative of its family in the sample), as the left-most language. Inspecting correspondence patterns will therefore start from Aguaruna and immediately illustrate that literally no regular patterns can be detected for this language variety compared to the rest of the languages. All we see are scattered patterns that only cover two or three languages for very few alignment sites.

COGNATES	INDEX	PATTERN	CONCEPTS	Cas	Cat	Cha	Pac	Shi	Yam	SIZE
998	1	ʂ / 1029	beak	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
8069	1	ʂ / 1029	green, unripe	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
14413	3	ʂ / 1029	raw	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
8014	1	ʂ / 1031	grease, fat	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
9099	1	ʂ / 1031	house	ʂ	ʃ	ʂ	ʂ	ʂ	Ø	7.17 / 9
9528	1	ʂ / 1031	itch	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
10448	1	ʂ / 1031	light (in weight)	ʂ	ʃ	ʂ	Ø	ʂ	ʂ	7.17 / 9
12617	1	ʂ / 1031	oil	ʂ	Ø	ʂ	Ø	ʂ	ʂ	7.17 / 9
13266	1	ʂ / 1031	peel	ʂ	ʃ	ʂ	Ø	ʂ	Ø	7.17 / 9

Figure 2: Correspondence pattern example for Panoan languages in the sample.

When restricting the selection to the six Panoan languages in the sample (which can be done via the language selection panel at the top left of the tool), we see, on the contrary, a completely different picture. There are numerous correspondence patterns recurring across multiple alignment sites involving all six languages. As an example, consider the correspondence patterns for sibilant retroflex fricatives in Figure 2.

On the contrary, when adding the unrelated languages Aguaruna and Spanish to the sample, we find only a much smaller amount of regularly recurring correspondence pattern, with those identified by the tool being clear borrowings, as can be seen from Figure 3, where patterns involving [p] in Aguaruna have been selected. As can be seen easily from the Figure, of the three correspondences with Spanish, three are borrowings, either from South-American languages into Spanish (*papa* "potato") or from Spanish into the South-American languages in our sample (*espejo* "mirror" and *zapato* "shoe").

COGNATES	INDEX	PATTERN	CONCEPTS	Agu	Cas	Shi	Spa	Yam	SIZE
15958	3	p / 526	shoe	s a p a t	Ø	Ø	θ a p a t o	p	3.00 / 4
13849	1	p / 526	potato	p a p a	p	p	p a p a	p	3.00 / 4
14687	1	p / 526	rib	p a g a ɨ	p	p	Ø	p	3.00 / 4
11432	3	p / 526	mirror	i s p i h u	Ø	p	ɛ s p e x o	Ø	3.00 / 4

Figure 3: Borrowing patterns between Spanish and South-American languages.

Although being done in an extremely rudimentary fashion, this sample analysis illustrates already the huge potential that the targeted modification of legacy data offers for computational and computer-assisted analysis.

5 Conclusion

In this study, we have presented a new modified version of several South American languages in the IDS in which phonetic transcriptions in standardized form have been added. The data assembled in this form can be used in multiple ways, to search for cognates and deeper relations among Panoan and Tacanan languages, to identify borrowings from dominant languages and how they influence the languages in the region, or as the starting point for phylogenetic analyses. At the same time, the study illustrates how datasets like the IDS can be enhanced by not standardizing them as a whole but by modifying particular subsets of the data in targeted studies.

References

Blum, F., C. Barrientos, R. Zariquiey, and J.-M. List (2024): A comparative wordlist for investigating distant relations among languages in Lowland South America. 11.92. https://doi.org/10.1038/s41597-024-02928-7

Forkel, Robert; List, Johann-Mattis; Greenhill, Simon J.; Rzymski, Christoph; Bank, Sebastian; Cysouw, Michael; Hammarstrom, Harald; Haspelmath, Martin; 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. https://cldf.clld.org

Forkel, Robert and List, Johann-Mattis (2020): CLDFBench. Give your Cross-Linguistic data a lift. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation. 6997-7004. https://pypi.org/project/cldfbench

Forkel, Robert and List, Johann-Mattis and Rzymski, Christoph and Segerer, Guillaume (2024): Linguistic Survey of India and Polyglotta Africana: Two Retrostandardized Digital Editions of Large Historical Collections of Multilingual Wordlists. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 10578-10583. https://lsi.clld.org

Geisler, Hans and Forkel, Robert and List, Johann-Mattis (2021): A digital, retro-standardized edition of the Tableaux Phonétiques des Patois Suisses Romands (TPPSR). In: M. Avanzi and N. LoVecchio and A. Millour and A. Thibault (eds.): Nouveaux regards sur la variation dialectale. Strasbourg:Éditions de Linguistique et de Philologie. 13-36. https://tppsr.clld.org

Gévaudan, Paul (2007): Typologie des lexikalischen Wandels. Bedeutungswandel, Wortbildung und Entlehnung am Beispiel der romanischen Sprachen. Tubingen: Stauffenburg.

Hammarstrom, Harald and Haspelmath, Martin and Forkel, Robert and Bank, Sebastian (2024): Glottolog [Dataset, Version 5.0]. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

Key, Mary Ritchie and Comrie, Bernard (2016): The Intercontinental Dictionary Series. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://ids.clld.org

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

List, Johann-Mattis (2017): A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations. 9-12.

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

List, Johann-Mattis and Forkel, Robert (2023): LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.13]. Passau: MCL Chair at the University of Passau. https://pypi.org/project/lingpy

List, Johann-Mattis; Anderson, Cormac; Tresoldi, Tiago; Rzymski, Christoph; and Forkel, Robert (2024a): Cross-Linguistic Transcription Systems [Dataset, Version 2.3.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://clts.clld.org

List, Johann-Mattis; Tjuka, Annika; van Zantwijk, Mathilda; Blum, Frederic; Barrientos Ugarte, Carlos; Rzymski, Christoph; Greenhill, Simon J.; and Robert Forkel (2024b): CLLD Concepticon [Dataset, Version 3.2.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, Johann-Mattis and van Dam, Kellen Parker (2024a): EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.0]. Passau: MCL Chair at the University of Passau. https://edictor.org

List, Johann-Mattis and van Dam, Kellen Parker (2024b): Computer-Assisted Language Comparison with EDICTOR 3. In: 5th International Workshop on Computational Approaches to Historical Language Change 2024. https://aclanthology.org/2024.lchange-1.1

Miller, John E. and List, Johann-Mattis (2023): Detecting lexical borrowings from dominant languages in multilingual wordlists. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Short Papers. Association of Computational Linguistics 2591-2597. https://aclanthology.org/2023.eacl-main.190/

Moran, Steven and Cysouw, Michael (2018): The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin: Language Science Press.

PONS (2023): PONS.eu Online-Worterbuch. Stuttgart: http://de.pons.com/

Sokal, Robert. R. and Michener, Charles. D. (1958): A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28. 1409-1438.

Supplementary Material
Data and code can be found at The data presented in this study has been curated on GitHub (https://github.com/intercontinental-dictionary-series/keypano) and archived with Zenodo (https://doi.org/10.5281/zenodo.13234493).
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.