scientific data

OPEN

DATA DESCRIPTOR

A comparative wordlist for investigating distant relations among languages in Lowland South America

Frederic Blum[1]✉, Carlos Barrientos[1,2], Roberto Zariquiey[3] & Johann-Mattis List[1,4]

The history of the language families in Lowland South America remains an understudied area of historical linguistics. Panoan and Tacanan, two language families from this area, have frequently been proposed to descend from the same ancestor. Despite ample evidence in favor of this hypothesis, not all scholars accept it as proven beyond doubt. We compiled a new lexical questionnaire with 501 basic concepts to investigate the genetic relation between Panoan and Tacanan languages. The dataset includes data from twelve Panoan, five Tacanan, and four other languages which have previously been suggested to be related to Pano-Tacanan. Through the transparent annotation of grammatical morphemes and partial cognates, our dataset provides the basis for testing language relationships both qualitatively and quantitatively. The data is not only relevant for the investigation of the ancestry of Panoan and Tacanan languages. Reflecting the state of the art in computer-assisted approaches for historical language comparison, it can serve as a role model for linguistic studies in other areas of the world.

Background & Summary

Much of the human history in South America is unknown, and linguistics can be one of many tools to investigate the human past. Yet, the linguistic history in South America is poorly understood, and despite the comparably recent human settlement, many genetic relationships between language families remain hypotheses without too much evidence[1–3]. In addition to that, most languages on the continent are severely endangered and the linguistic window to human history is closing[4,5]. New possibilities arise through the growth and application of computational methods, which in recent years not only inspired new research questions, but also offer a new perspective on unanswered cases. Part of this perspective has been made possible through transparent annotation of data and larger datasets becoming available[5–10]. Computational methods have become a valuable contribution to the study of linguistic history[11–13]. By combining those methods with the detailed work in documentary and historical linguistics, we aim to re-evaluate long-distance genetic relationships which have been proposed in the 20th century by applying the newly arisen methodologies.

One such case is the hypothesized Pano-Tacanan language family. Panoan and Tacanan are two language families currently spoken in Lowland South America[14,15], which have long been hypothesized to be genetically related[16,17]. Both language families have also been claimed to be related to other languages in the area, such as Mosetén[18,19], Chipaya, and Movima[20]. Even though there is a considerable amount of evidence in favor of the 'Pano-Tacanan hypothesis'[21–23], no fully accepted large-scale reconstruction has yet been carried out. The Panoan language family was first proposed by de la Grasserie in 1889[24]. A preliminary reconstruction of the common ancestor was carried out by Shell[25], which, however, lacked data from the Northern branch of the family and of Kaxararí[26]. Recently, a new reconstruction has been proposed by Oliveira[27], which still needs further revisions. The Tacanan languages on the other hand were proposed by Brinton in 1891[28] and reconstructed by Key[29] and later Girard[17]. Based on this reconstruction and the 'Reconstructed Pano' from Shell, Girard also proposed a reconstruction for the ancestral language, Proto-Pano-Tacanan. Given the problems of the sampled

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [2]University of Leipzig, Leipzig, Germany. [3]Pontificia Universidad Católica del Perú, Lima, Perú. [4]Chair of Multilingual Computational Linguistics, University of Passau, Passau, Germany. ✉e-mail: frederic_blum@eva.mpg.de

languages for Shell's Panoan reconstruction, however, this reconstruction is not generally accepted as a proof for the Pano-Tacanan family, and some doubts remain. More recently, Valenzuela & Zariquiey[23] provide a new reconstruction of Proto-Pano-Tacanan, but this work is limited with respect to the amount of lexical coverage. It does, however, provide a first detailed account of grammatical morphemes that appear to be cognates between the Panoan and the Tacanan language family. Cognates are lexical roots and morphemes from two genetically related languages that descend from the same ancestral form etymologically[30].

This dataset aims to present lexical data that can be used as a new starting point for investigating the past of Panoan, Tacanan, and other languages. Using state-of-the-art methods for computer-assisted historical language comparison[13], our dataset presents lexical data for 501 concepts of basic vocabulary across 21 languages. Basic vocabulary refers to stable lexical terms that are assumed to be more resistant to lexical borrowing than others, and thus more reliable than other parts of the vocabulary for establishing sound correspondences. Together with grammatical evidence, they are generally accepted as providing evidence for genetic relationships between languages[31].

Of the 21 languages, 17 are directly part of either Panoan or Tacanan, and four languages are included that have previously been claimed to be related to Pano-Tacanan. In total, the dataset comprises data from five different genealogical entities: Panoan, Tacanan, Chipaya, Mosetén-Tsimane, and Movima. The data is annotated for morphemes and partial cognacy, which opens the path for a detailed computer-assisted analysis of the languages involved, both from a qualitative and a quantitative perspective. Examples for such a workflow are provided as Technical Validation and Usage Notes. The dataset is intended to work as a role model for future studies on other long-distance genetic relationships, which can orient themselves at the standards and details of annotation offered in this dataset.

Methods

Wordlists for Historical Language Comparison. The original goal for developing the dataset was an analysis of the genetic relationship of the languages involved. In order to conclusively establish such a relationship, a detailed phonological and morphological reconstruction that extends to the grammatical structure of the languages is necessary. However, the long way of proving the genetic relationship between languages tends to start with a comparative wordlist of basic vocabulary that is not specific to any culture or geographic region[32,33]. While the traditional wordlists mostly include 100 or 200 concepts[34,35], the low lexical coverage has been criticized for several reasons. For purposes of language documentation, Dockum and Bowern[36] argue that an average of 400 lexical items is necessary to identify all phonemes of a language. This, of course, is also a pre-requisite for an accurate historical analysis. Automated methods have been shown to benefit from wordlists of at least 300 items[37]. Other scholars argue that a minimum number of 500 etymological concepts is necessary in order to find sufficient recurring sound correspondence patterns to work on the phonological reconstruction using the comparative method[2]. For phylogenetic studies, a sample size of 33 cognate classes per classified language has been suggested[38]. In any case, this means that in order to capture all relevant sound correspondences for Pano-Tacanan (and beyond), the first necessary (but not sufficient) step is to create a large-scale lexical dataset.

In its current version, the concept list of our dataset contains data for 501 concepts. The list is largely based on a rarely used concept list proposed by Kaufman (https://www.ailla.utexas.org/islandora/object/ailla%3A246899) of more than 1000 individual entries. This list was originally gathered by analyzing 35 comparative studies that involve reconstruction of ancestor languages. The most frequent of the 2100 meanings compiled from those studies have been selected by the original author for his final list. The author claims that those are the most stable etymologies, and are part of a 'universal basic vocabulary' that, 'if applied to a set of related languages, will yield more true cognates than any other list of its size' (Kaufman 1973, p.29)[39]. As the original list was only recently made available in an archive, it has not found wide distribution among scholars earlier, despite its potential for historical language comparison. The approach presented by Kaufman contrasts with other approaches to historical language comparison, where the dictionaries are searched in a targeted way for specific items that are assumed to be cognate. However, this way of investigating cognacy among putatively related languages comes at the danger of cherry-picking the desired data. The advantage of the larger wordlist is that it is more realistic to find all the relevant sound correspondences compared to the small wordlists, while not cherry-picking the data.

For creating the first version of our wordlist, we chose a 450-concept subset of the concepts provided by Kaufman. For example, we have removed concepts relating to grammatical concepts (e.g. verbal inflection markers and case-marking) and those that relate for flora and fauna. As one step during the pre-processing of the data, we added the concept list to Concepticon[40] (https://concepticon.clld.org/contributions/Kaufman-1973-1028). This makes it possible to compare our concept list with concept lists that had been used in previous attempts of reconstructing Panoan (Shell, Oliveira) and Tacanan (Girard) in order to arrive at the most important concepts for this study. For this purpose, the concepts reconstructed for Proto-Panoan[27] and for Proto-Tacanan[17] were also added to Concepticon. The addition to Concepticon helps in mapping the individual concepts despite the different languages of the original publications (English, Spanish, Portuguese) and to easily integrate comparisons to other sources, such as the Swadesh lists for basic vocabulary[34,35]. In the selection of the final subset, we have oriented ourselves at the intersection of concepts between the Kaufman-conceptlist and the conceptlists from Swadesh, Oliveira, and Girard. The concepts from Oliveira and Girard that have not been used are primarily terms for flora and fauna, while the more basic terms have been preserved.

Language Sample. The language sample includes three groups of languages, namely a) twelve (of 18 extant) Panoan, b) five (of seven extant) Tacanan languages, and c) four languages from three small language families, which have previously been argued to be related to Panoan or Pano-Tacanan. All relevant languages for which reliable linguistic material exists have been included in the sample. Given the large amount of concepts in this dataset and the need for high mutual coverage, only languages with published dictionaries have been selected.

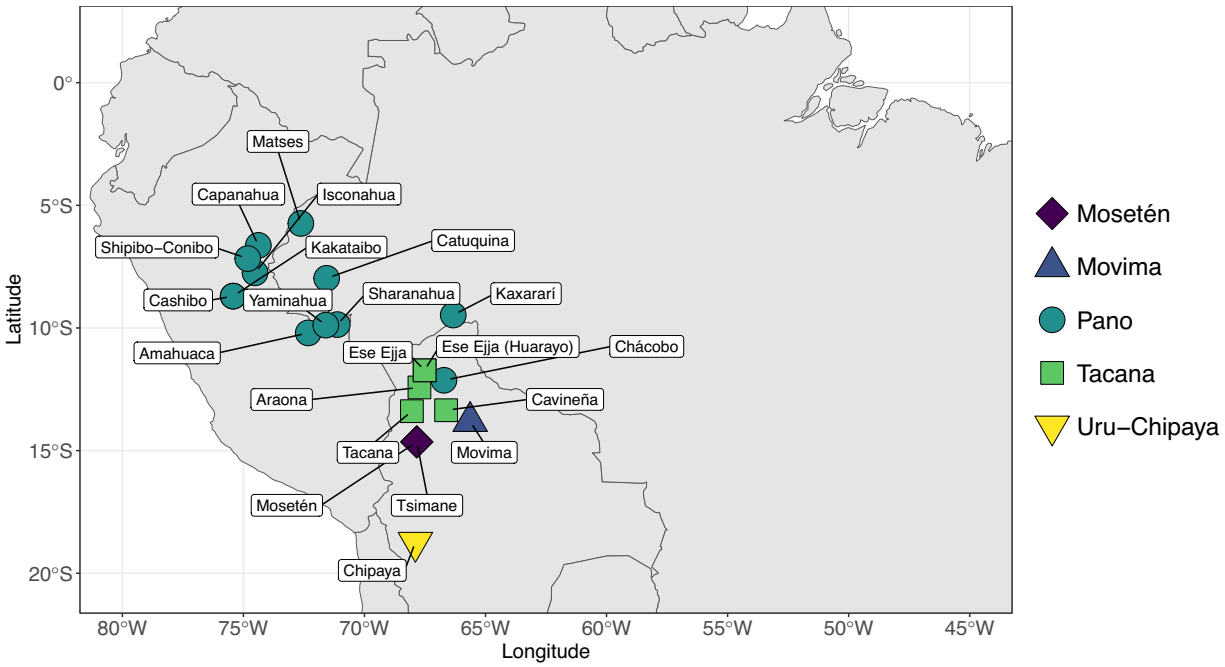

Fig. 1 Location of sampled languages in the dataset.

A more detailed account of the sources will be presented in Table 2 in the Data Records section. The map of sampled languages is presented in Fig. 1.

The data was gathered using both traditional and computational methods. The most important single source is the IDS dataset[41]. This digital publication provided the data for four Panoan (Cashibo, Catuquina, Shipibo-Conibo, Yaminawa) and two Tacanan (Ese Ejja, Huarayo) varieties. As there is a considerable overlap between the IDS conceptlist and ours, it was possible to quickly integrate the data. The availability from IDS also contributed to the integration of non-Panoan languages, as Movima and Mosetén were already digitally available. To complement this, the extensively documented Tsimané variety of Mosetén was added manually to complement the sample. For some languages, we were able to extract the data directly from digital published dictionaries. In two cases (Isconahua, Kakataibo), data recorded by one of the co-authors (RZ) and stored using ToolboX, a language documentation software, were parsed and manually checked for integration into the wordlist. Two other dictionaries, Matses[26] and Sharanahua[42], were parsed from their PDF source. For the remaining languages, we went through the dictionaries manually to extract the relevant data.

A recurring problem for the manual sampling is synonymy. There are many cases where the authors give more than two forms for a meaning, which results in a problematically high synonymy. Where possible, we included only the most frequently used item for a specific concept. Such information is often provided in dictionaries, where archaic terms or less commonly used terms are provided after the most frequently used one. In cases of allomorphic or phonological alternations, we have added this as notes to the data entry. In other cases, the most general form has been used, if such information was available. For example, the entry for 'to eat' might be accompanied by terms for 'to eat fish' or 'to eat fruit'. In this case, we have chosen the first, most general entry, that provides the best fit for our target concept.

There is one exception to the goal of high coverage, and that is the sparsely documented language Kaxararí. Due to several phonological characteristics absent in other Panoan languages, like the presence of a lateral consonant /l/, Kaxararí is argued to be of great importance for the reconstruction of Proto-Panoan[27]. For this reason, we extracted the data for Kaxararí from another dataset which digitized the Proto-Pano reconstruction by Oliveira[27,43]. In his work, Oliveira includes data on Kaxararí from several different sources, presenting 171 forms in total. Due to the lack of a detailed grammatical description of the language, the exact quality of the available data for this language cannot be confirmed without further documentation work with the speakers. Excluding Kaxararí, the total coverage is at ∼85%.

Annotation in EDICTOR: Morphemes, Cognates, Alignments. The data is annotated with respect to different levels of linguistic analysis using computer-assisted methodology. For a detailed historic analysis, we need to detect and segment morphemes, and assign partial cognacy to all elements[44]. Furthermore, we need to exclude known borrowings from the data. As only words that descend from the proto-language should be considered cognate, all known borrowings from Spanish and Quechua that could be found in the data were annotated as such.

As a starting point for annotating the data, automated cognate judgements are carried out using the LexStat algorithm from LingPy (v2.6.9[45]). For carrying out the manual annotations and to correct the automated cognate judgements, the data is imported to EDICTOR (v2.0.0)[46], a visual tool for annotating data in historical linguistics.

Concept	Doculect	Segments	Partial cognacy	Morphemes
EYE	Amahuaca	β i + r o	681 608	round_object + eye
EYE	Capanahua	β i + r o	681 608	round_object + eye
EYEBROW	Cashibo	β i + s k o	681 860	round_object + eyebrow
EYEBROW	Chacobo	β i + s k o	681 860	round_object + eyebrow
TEAR (OF EYE)	Isconahua	β i + r o + i n i	681 608 315	round_object + eye + water

Table 1. Segmented and annotated morphemes various concepts related to eye in Panoan languages.

In a first step, affixes are separated from their roots. They are assigned a different ID of cognacy, as they do not relate etymologically to the lexical root. As part of this step, we have included morpheme glosses[47,48]. In this step, non-salient morphemes are tagged explicitly in order to be excluded from further analysis[44]. This includes for example verbal derivational markers or instrumental nominalizers whose presence is mostly due to artifacts in the process of language documentation. For example, in some traditions verbs are always presented in the first-person singular form, while others may give a base form. Excluding such kind of artifacts from the data is thus essential to assure comparability across forms. After the segmentation of morphemes, partial cognacy is analyzed within each concept[44]. In a second step, this analysis is carried out across similar meanings. For example, many of the languages in the dataset colexify the terms for GREEN and UNRIPE. Hence, the forms are represented in the same cognate set. Other examples include body-part roots, which are widespread across the Panoan languages[49,50]. This is showcased in Table 1, where several languages share the same body-part root for small, round objects (e.g. 'eye', 'seed'), but the formatives differ with respect to the affix they combine with to arrive at different concepts. A network visualization of full colexifications will be presented as part of the Technical Validation.

In the final step, all cognate sets that have been found are aligned phonetically. During this step, all cognate sets are checked for validity, and erroneous cognate judgements have been fixed. The correspondence patterns can now be extracted and analyzed computationally as well as in manual fashion. An example for this automated extraction is added as a script within the data repository and briefly presented in the Data Records.

Data Records

The dataset in its current version is stored on Zenodo (v0.2)[51]. It is published under a CC-BY 4.0 license and curated on GitHub (https://github.com/pano-tacanan-history/blumpanotacana/tree/v0.2). The data follows a specific template of CLDF[7], namely that of Lexibank[8]. The main data intended for re-use is stored in the 'cldf'-folder, while the two additional folders 'raw' and 'etc' are mainly used for the conversion of the raw data into CLDF. In the main directory, the 'lexibank_blumpanotacana.py'-script manages the conversion from raw data to CLDF. This includes a download of the most up-to-date version of the data from EDICTOR. A 'metadata.json' file stores all the relevant metadata for the dataset, namely its ID, a short description, the adequate citation, the license, and the link to the Concepticon wordlist. Further technical details of CLDF will be described in the section on Technical Validation.

The 'cldf'-folder consists of csvw-files ('csv on the web') whose metadata is stored in the 'cldf-metadata.json' file. The individual lexemes are stored in 'forms.csv', with columns for the entry ID, the language ID, a parameter ID, value and form of the entry, tokenized segments, additional comments that have been added during analysis, the source, the cognate set ID's, and information about borrowing. The 'cognate.csv' file stores additional information about the cognate sets, such as the detection method ('expert', because it has been done manually by the first author) and the phonetic alignments of the tokenized entries. A 'languages.csv' file includes the necessary information about the languages in the dataset: ID, Name, Glottocode, the Macroarea, Latitude and Longitude of the language, and the language subgroup. Similarly, the 'parameters.csv' file stores information about the concepts, their ID, Name, their ID and glosses on Concepticon, as well as a translation to Spanish and Portuguese, common languages for dictionaries which have been used as source for the dataset. This file provides the translational equivalents to the English concepts. The 'sources.bib' file contains all the sources that contributed to the dataset in BiBTeX-format. The 'requirements.txt' and 'README.md' files round off the folder for reproducibility of the CLDF conversion.

The original raw data is represented in a csv-file within the 'raw'-folder. A metadata 'etc' folder includes the tsv-files that are necessary for linking the data to other large-scale linguistic datasets. This includes the mapping of the languages to Glottolog ('languages.tsv')[52] and orthography profiles that map the graphemes in all languages to sounds in CLTS[53]. Those are included within a subfolder that contains the individual orthography profiles for all languages. We have included a folder 'analysis/' which includes all scripts as presented in the Usage Notes. This includes the automated extraction of correspondence patterns ('s_patterns.py') using the LingRex package in Python[54,55], as well the code for all figures that are part of this data descriptor. The main README.md file containts a walk-through for all scripts. The coverage, synonymy, and sources of all languages are presented in Table 2.

Technical Validation

Integration with Reference Catalogues. The final data is presented using the Cross-Linguistic Data Format (CLDF)[7]. The conversion into CLDF includes several control measures, such as the linking to several linguistic reference catalogues to retrieve information about the concepts (Concepticon v3.1.0)[40,56], the languages (Glottolog v4.8)[52], and the phonemes in the data (CLTS v2.2.0)[53,57]. This includes the mapping of graphemes to tokenized phonemes through orthography profiles[58], ensuring that all representations for analysis are based on

Variety	Glottocode	Language Family	Original Author	Coverage	Synonymy
Amahuaca	amah1246	Pano	Hyde[63]	0.85	1.15
Capanahua	capa1241	Pano	Loos & Loos[64]	0.91	1.36
Cashibo	cash1251	Pano	Key & Comrie[41]	0.79	1.14
Catuquina	pano1254	Pano	Key & Comrie[41]	0.91	1.06
Chácobo	chac1251	Pano	Zingg[65]	0.93	1.05
Isconahua	isco1239	Pano	Zariquiey[66]	0.77	1.13
Kakataibo	cash1251	Pano	Zariquiey[67]	0.88	1.27
Kaxararí	kaxa1239	Pano	Oliveira[27]	0.21	1.07
Matses	mats1244	Pano	Fleck *et al.*[26]	0.86	1.28
Sharanahua	shar1245	Pano	Scott[42]	0.93	1.30
Shipibo-Conibo	ship1254	Pano	Key & Comrie[41]	0.85	1.38
Yaminahua	yami1256	Pano	Key & Comrie[41]	0.89	1.17
Araona	arao1248	Tacana	de Pitman[68]	0.82	1.03
Cavineña	cavi1250	Tacana	Guillaume[69]	0.74	1.20
Ese Ejja	esee1248	Tacana	Key & Comrie[41]	0.81	1.36
Ese Ejja (Huarayo)	esee1248	Tacana	Key & Comrie[41]	0.74	1.07
Tacana	taca1256	Tacana	Ottaviano & Ottaviano[70]	0.86	1.07
Chipaya	chip1262	Uru-Chipaya	Cerrón-Palomino & Ballón Aguirre[71]	0.87	1.29
Movima	movi1243	isolate	Key & Comrie[41]	0.90	1.00
Mosetén	mose1249	Mosetén-Tsimane	Key & Comrie[41]	0.87	1.14
Tsimane	mose1249	Mosetén-Tsimane	Gill[72]	0.87	1.20
Overall: 21				**0.82**	**1.17**

Table 2. Source, synonymy, and coverage of all language varieties in the dataset.

sounds, and not on orthography. The phonemes are linked directly to CLTS, which contains further information about the individual sounds. Similarly, all concepts in the data are linked to the list on Concepticon[40]. A metadata file for languages includes information such as the glottocode for linking to Glottolog, and information on the language family subgroup. In the two cases of overlapping Glottocodes (Mosetén-Chimané, EseEjja-Huarayo), the two ID's in the dataset include two recognized varieties that are not represented as such in Glottolog.

Quality Measures. CLDF comes with a variety of quality measures. The data is converted using cldfbench with the pylexibank plug-in[59]. This step involves detailed quality checks, such as whether all sounds are represented according to the CLTS standard, that all concepts are represented in the conceptlist, and that all languages are part of the languages-metadata file.

The standardization of the data makes it easy to conduct further computational measures that assure the quality of the data. Based on a computational implementation that measures the regularity of correspondence patterns in cognate sets in the data[60], we can analyze the proposed cognate sets. In Fig. 2, we present the proportion of shared cognate sets between Panoan languages. Even though this measure is not an explicit phylogenetic representation, it closely resembles the currently accepted family tree for the Panoan languages, with Matses forming an outgroup[26].

Visualizing Colexifications in the Data. The standardization of the data makes it possible to study the colexifications in the dataset. A network visualization of those cross-semantic colexifications can be used to verify the semantic relationship between concepts in the language families of the dataset. In Fig. 3, we present the colexifications around the concepts WATER, EYE, and FACE.

Usage Notes

Through the standardization of the data, we can exploit the consistency of the annotations in several ways. As the data is stored in tabular files (csv), the tables are readable on all computing platforms. You can install all the necessary packages that we use by cloning into the repository and installing the dependencies in the command line.

```
> git clone https://github.com/pano-tacanan-history/blumpanotacana
> cd blumpanotacana
> pip install –e.
```

The data can be accessed both manually and computationally. For a manual inspection of the data within a single file, we provide a 'd_blumpanotacana.tsv' in the 'analysis'-folder. This file includes the cognates and alignments, and can be uploaded to a local version of EDICTOR ('http://lingulist.de/EDICTOR/'). This is especially useful for linguists who want to manually assess the quality of the alignments provided in the dataset. Of course, you can also open this file with any office application or use it for inspection with other tools of programming.

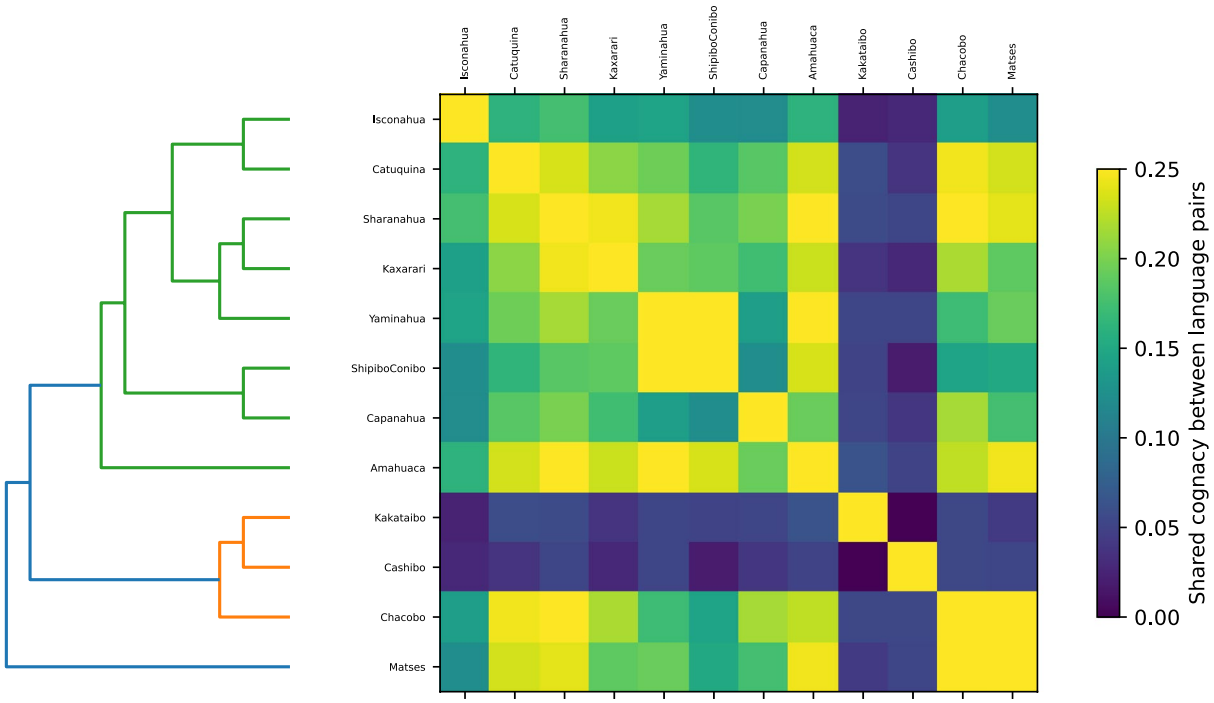

Fig. 2 Proportion of shared cognate sets between Panoan languages.

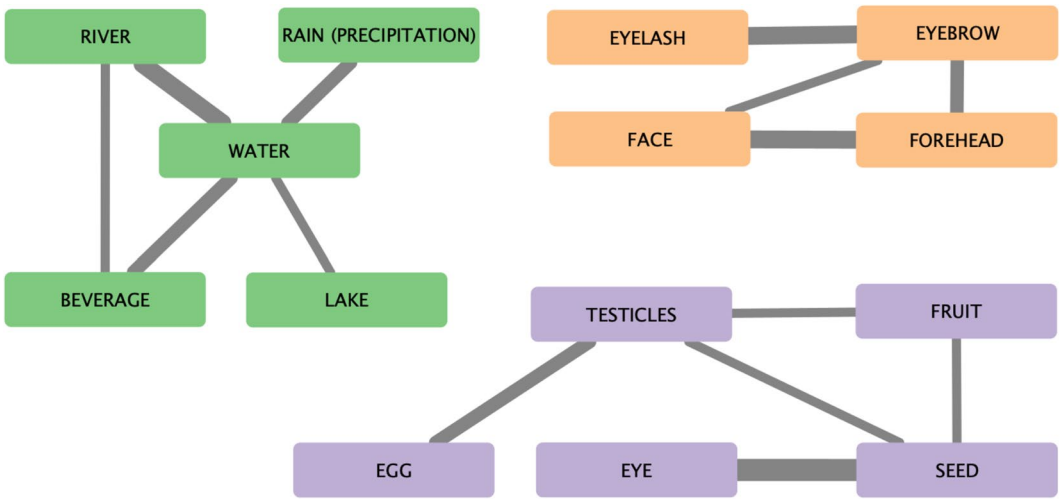

Fig. 3 Colexification of concepts in the dataset. The width of the network edges corresponds to the amount of languages in which the concepts are colexified.

The file is created using the 'pyedictor'-package[61] that comes as part of the repository dependencies, using the following command from the commandline interface:

> cd analysis/
> edictor wordlist --name = d_blumpanotacana --data = ./cldf/cldf-metadata.json > --preprocessing = s_realign.py --addon = "language_subgroup:subgroup", > "cognacy:cogid", "partial_cognacy:cogids", "borrowing:borrowing".

The call to pyedictor includes the output file ('--name'), the input CLDF metadata ('--data'), a script for pre-processing that can be adopted to other purposes ('--preprocessing'), and columns from the different CLDF tables with the syntax 'cldf-name:column-name'. The same workflow can also be used to create similar files from other Lexibank-datasets[8].

Having installed the requirements, the dataset can now easily be converted to a SQLite dataset using a command from the pycldf-package in the command line[7].

```
> cldf createdb cldf/cldf-metadata.json blumpanotacana.sqlite
```

This dataset can then be queried with all common programming and dataset tools. Given the linking to other reference catalogues in linguistics, the data is easily comparable with information from other datasets. For example, we can use SQLite queries to integrate the data with other datasets, such as Grambank[5], the largest currently available dataset on grammatical information of languages, which is equally linked to the reference catalogues. By creating the SQLite dataset for Grambank in the same way as we did for the lexical dataset, we can retrieve the information in Grambank for all the languages in the dataset. This examplifies the utility advantage for integrating datasets by using CLDF and SQLite. The following SQLite commands showcases the integration of Grambank-data based on the glottocodes of the languages in the current dataset. An example script for this process that uses SQLite is provided within the 'analysis/'-folder.

```
> attach 'blumpanotacana.sqlite' AS db1;
> attach 'grambank.sqlite' AS db2;

> SELECT *
> FROM db1.LanguageTable AS a
> INNER JOIN db2.ValueTable AS b
> ON a.cldf_glottocode = b.cldf_languageReference;
```

Code availability

All code that has been used during the creation of this dataset is published on Zenodo (v0.2)[51] and curated on GitHub (https://github.com/pano-tacanan-history/blumpanotacana). For converting the data to CLDF[7], we have used the Python tools cldfbench (v1.13.0)[59] using the pylexibank plugin (v3.4.0)[62]. The dataset is linked to Concepticon (v3.1.0)[40], Glottolog (v4.7)[52], and CLTS (v2.2.0)[53,57]. The code for integrating data with other datasets via SQL is presented in the main README.md. The scripts that were used to create the plots and to compute the coverage and synonymy is part of the 'analysis' folder, where another README.md file leads through the replication of all necessary steps. The code for the initial addition of IDS data is added to 'raw/archive/' for documentation. This list was then filtered while finalizing the concept list. All the orthography profiles that are used during the conversion of graphemes are part of 'etc/orthography'.

Received: 30 October 2023; Accepted: 5 January 2024;
Published online: 18 January 2024

References

1. Campbell, L. Review Article: Language in the Americas. By Joseph H. Greenberg. Stanford, California: Stanford University Press, 1987, Pp. x.,438. *Language* **64**, 591–615 (1988).
2. Kaufman, T. Language history in South America: What we know and how to know more. In Payne, D. L. (ed.) *Amazonian linguistics: Studies in Lowland South American languages*, 13–67 (University of Texas Press, Austin, 1990).
3. Michael, L. The Classification of South American Languages. *Annual Review of Linguistics* **7**, 329–349, https://doi.org/10.1146/annurev-linguistics-011619-030419 (2021).
4. Adelaar, W. F. H. Threatened Languages in Hispanic South America. In *Language Diversity Endangered*, 9–28, https://doi.org/10.1515/9783110197129.9 (Mouton de Gruyter, 2007).
5. Skirgård, H. *et al*. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* **9**, eadg6175, https://doi.org/10.1126/sciadv.adg6175 (2023).
6. Dryer, M. S. & Haspelmath, M. WALS Online (v2020.3). *Zenodo* https://doi.org/10.5281/zenodo.7385533 (2013).
7. Forkel, R. *et al*. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* **5**, 180205, https://doi.org/10.1038/sdata.2018.205 (2018).
8. List, J.-M. *et al*. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* **9**, 1–16, https://doi.org/10.1038/s41597-022-01432-0 (2022).
9. Wichmann, S., Holman, E. W. & Brown, C. H. The ASJP database (version 20) (2022).
10. Heggarty, P., Anderson, C. & Hans-Jörg, B. CLDF dataset derived from Heggarty, Paul and; Anderson, Cormac and Scarborough, Matthew's "Indo-European Cognate Relationships database project" (IE-CoR) from 2022. *Zenodo* https://doi.org/10.5281/ZENODO.8089434 (2023).
11. Wichmann, S. & Saunders, A. How to use typological databases in historical linguistic research. *Diachronica* **24**, 373–404, https://doi.org/10.1075/dia.24.2.06wic (2007).
12. Jäger, G. Computational historical linguistics. *Theoretical Linguistics* **45**, 151–182, https://doi.org/10.1515/tl-2019-0011 (2019).
13. Wu, M.-S., Schweikhard, N. E., Bodt, T. A., Hill, N. W. & List, J.-M. Computer-Assisted Language Comparison: State of the Art. *Journal of Open Humanities Data* **6**, 2, https://doi.org/10.5334/johd.12 (2020).
14. Fleck, D. W. Panoan languages and linguistics. *Anthropological Papers of the American Museum of Natural History* **99**, https://doi.org/10.5531/sp.anth.0099 (2013).
15. Valenzuela, P. & Guillaume, A. Estudios sincrónicos y diacrónicos sobre lenguas Pano y Takana: una introducción. *Amerindia* **39**, 1–49 (2017).
16. Schuller, R. The Language of the Tacana Indians. *Anthropos* **28**, 99–116 (1933). 463—484.
17. Girard, V. *Proto-Takanan Phonology*. (University of California Press, Berkeley, Los Angeles, London, 1971).
18. Suárez, J. A. Moseten and Pano-Tacanan. *Anthropological Linguistics* **11**, 255–266 (1969).
19. Suárez, J. A. Macro-Pano-Tacanan. *International Journal of American Linguistics* **39**, 137–154, https://doi.org/10.1086/465258 (1973).
20. Greenberg, J. H. *Language in the Americas* (Stanford University Press, Stanford, 1987).
21. Wichmann, S. A classification of Papuan languages. In Hammarström, H. and van den Heuvel, W. (eds.) *History, contact and classification of Papuan languages*, 313–386 (Linguistic Society of Papua New Guinea, Port Moresby, 2013).

22. Zariquiey, R. & Valenzuela, P. M. Body-part nouns, prefixation, incorporation, and compounding in Panoan and Takanan: Evidence for the Pano-Takanan hypothesis? In Zariquiey, R. and Valenzuela, P. (eds.) *The Grammar of Body-Part Expressions*, 441–466, https://doi.org/10.1093/oso/9780198852476.003.0017 (Oxford University Press, Oxford, 2022).
23. Valenzuela, P. & Zariquiey, R. Language classification in Western Amazonia: Advances in favor of the Pano-Takana Hypothesis. *LIAMES: Línguas Indígenas Americanas* **23**, e023002, https://doi.org/10.20396/liames.v23i00.8670150 (2023).
24. de La Grasserie, R. *De la famille linguistique Pano* (Maisonneuve & C. Leclerc, Paris, 1889).
25. Shell, O. A. *Pano Reconstruction*. Ph.D. thesis, University of Pennsylvania (1965).
26. Fleck Zuazo, D. W., Uaqui Bëso, F. S. & Jiménez Huanán, D. M. *Diccionario Matsés - Castellano* (Tierra Nueva, Iquitos, 2012).
27. Oliveira, S. C. S. d. *Contribuições para a reconstrução do Protopáno*. Ph.D. thesis, Universidade de Brasília, Brasília (2014).
28. Brinton, D. G. *The American Race* (N. D. C. Hodges Publisher, New York, 1891).
29. Key, M. R. *Comparative Tacanan Phonology* (Mouton, The Hague, Paris, 1968).
30. Trask, R. L. *The Dictionary of Historical and Comparative Linguistics* (Edinburgh University Press, 2000).
31. Campbell, L. & Poser, W. J. *Language classification: History and method* (Cambridge University Press, Cambridge, 2008).
32. Campbell, L. How to Show Languages are Related: Methods for Distant Genetic Relationship. In Joseph, B. D. & Janda, R. D. (eds.) *The Handbook of Historical Linguistics*, 262–282, https://doi.org/10.1002/9781405166201.ch4 (Blackwell Publishing, 2017).
33. Rankin, R. L. The comparative method. In Joseph, B. D. & Janda, R. D. (eds.) *The Handbook of Historical Linguistics*, 181–212, https://doi.org/10.1002/9781405166201.ch1 (Blackwell Publishing, 2017).
34. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* **96**, 452–463 (1952).
35. Swadesh, M. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* **21**, 121–137, https://doi.org/10.1086/464321 (1955).
36. Dockum, R. & Bowern, C. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. In Austin, P. K. (ed.) *Language Documentation and Description*, vol. 16, 35–54 (EL Publishing, London, 2018).
37. List, J.-M. Investigating the impact of sample size on cognate detection. *Journal of Language Relationship* **11**, 91–102, https://doi.org/10.31826/jlr-2014-110111 (2014).
38. Rama, T. & Wichmann, S. Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies. In Bender, E. M., Derczynski, L. & Isabelle, P. (eds.) *Proceedings of the 27th International Conference on Computational Linguistics*, 1578–1590 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018).
39. Kaufman, T. Kaufman's basic concept list on historical principles. https://www.ailla.utexas.org/islandora/object/ailla%3A246899 (1973).
40. List, J. M. *et al*. Concepticon v3.1.0. A Resource for the Linking of Concept Lists, https://doi.org/10.5281/zenodo.7777629 (2023).
41. Key, M. R. & Comrie, B. *Intercontinental Dictionary Series* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021).
42. Scott, M. *Vocabulario Sharanahua - Castellano*, vol. 53 of *Serie Lingüística Peruana* (Instituto Lingüístico de Verano, Lima, 2004).
43. Blum, F. & Barrientos, C. A New Dataset with Phonological Reconstructions in CLDF. *Computer-Assisted Language Comparison in Practice* **6** (2023).
44. Wu, M.-S. & List, J.-M. Annotating cognates in phylogenetic studies of Southeast Asian languages. *Language Dynamics and Change* 1–37, https://doi.org/10.1163/22105832-bja10023 (2023).
45. List, J.-M. & Forkel, R. *LingPy. A Python library for quantitative tasks in historical linguistics (v2.6.9)* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2022).
46. List, J.-M. *EDICTOR (v2.0.0). A web-based tool for creating, editing, and publishing etymological datasets* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2022).
47. Hill, N. W. & List, J.-M. Challenges of annotation and analysis in computer-assisted language comparison: A case study on burmish languages. *Yearbook of the Poznan Linguistic Meeting* **3**, 47–76, https://doi.org/10.1515/yplm-2017-0003 (2017).
48. Schweikhard, N. E. & List, J.-M. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics* **17**, 2–26 (2020).
49. Fleck, D. W. Body-Part Prefixes in Matses: Derivation or Noun Incorporation? *International Journal of American Linguistics* **72**, 59–96, https://doi.org/10.1086/505279 (2006).
50. Zariquiey, R. *et al*. Untangling the evolution of body-part terminology in Pano: conservative versus innovative traits in body-part lexicalization. *Interface Focus* **13**, https://doi.org/10.1098/rsfs.2022.0053 (2022).
51. Blum, F., Ugarte, C. M. B., Zariquiey, R. & List, J.-M. CLDF dataset derived from Blum et al.'s "A Comparative Wordlist for Investigating Distant Relations Among Languages in Lowland South America (v0.2). *Zenodo*, https://doi.org/10.5281/zenodo.10450408 (2023).
52. Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. Glottolog database (v4.8). *Zenodo* https://doi.org/10.5281/ZENODO.8131084 (2022).
53. List, J.-M., Anderson, C., Tresoldi, T. & Forkel, R. CLTS. Cross-Linguistic Transcription Systems (2.2.0). *Zenodo* https://doi.org/10.5281/zenodo.5583682 (2021).
54. List, J.-M. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* **45**, 137–161, https://doi.org/10.1162/coli_a_00344 (2019).
55. List, J.-M. & Forkel, R. *Linguistic Reconstruction with LingPy. [Computer software, Version 1.4.0]. With contributions by Frederic Blum and Mei-Shin Wu*. (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2023).
56. Tjuka, A., Forkel, R. & List, J.-M. Curating and extending data for language comparison in concepticon and NoRaRe. *Open Research Europe* **2**, https://doi.org/10.12688/openreseurope.15380.3 (2023).
57. Anderson, C. *et al*. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting* **4**, 21–53, https://doi.org/10.2478/yplm-2018-0002 (2018).
58. Moran, S. & Cysouw, M. *The Unicode Cookbook For Linguists: Managing Writing Systems Using Orthography Profiles* (Language Science Press, Berlin, 2018).
59. Forkel, R. & List, J.-M. CLDFBench: Give your cross-linguistic data a lift. In *12th Conference on Language Resources and Evaluation*, 6995–7002 (European Language Resources Association, Marseille, France, 2020).
60. Blum, F. & List, J.-M. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, 52–64 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).
61. List, J.-M. PyEDICTOR: A Small Python Package that Integrates LingPy, EDICTOR, and CLDF. *Computer-Assisted Language Comparison in Practice* **5** (2022).
62. Forkel, R. *et al*. pylexibank for CLTS 2.0. *Zenodo* https://doi.org/10.5281/ZENODO.4629131 (2022).
63. Hyde, S. *Diccionario Amahuaca* (Instituto Lingüístico de Verano, Yarinacocha, 1980).
64. Loos, E. & Loos, B. *Diccionario Capanahua Castellano* (Instituto Lingüístico de Verano, Yarinacocha, Pucallpa, Perú, 1998).
65. Zingg, P. *Diccionario Chacobo-Castellano, Castellano Chacobo: Con Bosquejo de la Gramatica Chacobo y con apuntes culturales* (Ministerio de Desarrollo Sostenible y Planificación, Ministerio de Educación, Cultura y Deportes, Confederación de Pueblos Indígenas de Bolivia, La Paz, 1998).
66. Zariquiey, R. *Vocabulario Iskonawa-Castellano-Inglés* (Department of Romance Languages, Tufts University, Boston, 2016).
67. Zariquiey, R. *A Grammar of Kakataibo* (De Gruyter Mouton, Berlin, Boston, 2018).

68. de Pitman, M. *Diccionario Araona y Castellano* (Instituto Lingüístico de Verano, Riberalta, 1981).
69. Guillaume, A. *A Grammar of Cavineña* (De Gruyter Mouton, Berlin, New York, 2008).
70. de Ottaviano, A. B. & Ottaviano, J. S. *Diccionario Tacana-Castellano* (Summer Institute of Linguistics, Dallas, 1989).
71. Cerrón-Palomino, R. & Ballón Aguirre, E. *Chipaya: Léxico y Etnotaxonomía* (Fondo Editorial de la Pontificia Universidad Católica del Perú, Lima, 2011).
72. Gill, D. *Diccionario Tsimane' - Castellano y Castellano - Tsimane'* (Misión Nuevas Tribus, 1993).

Acknowledgements

This project was supported by the Max Planck Society Research Grant 'Beyond CALC: Computer-Assisted Approaches to Human Prehistory, Linguistic Typology, and Human Cognition (CALC^3)'(2022-2024, FB and JML) and the ERC Consolidator Grant ProduSemy (Grant No. 101044282, see https://doi.org/10.3030/101044282, JML). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for helpful comments and all people who share their data openly, so we can use it in our research.

Author contributions

F.B. planned the dataset and gathered the data. J.-M. List developed and provided the software to annotate the data as well as supervision to the project. C.B. contributed to the annotation of the alignments. R.Z. provided fieldwork data for two language varieties and assisted on the manual annotation of the data. F.B. wrote the first draft. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.B.