DATA NOTE

REVISED A comparative wordlist for the languages of The Gran Chaco, South America [version 2; peer review: 2 approved]

Nicolás Brid[1], Cristina Messineo[1,2], Johann-Mattis List[3]

[1]Universidad de Buenos Aires, Buenos Aires, Argentina
[2]CONICET, Buenos Aires, Argentina
[3]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany

v2 **First published:** 21 Jul 2022, **2**:90
https://doi.org/10.12688/openreseurope.14922.1
Latest published: 06 Dec 2022, **2**:90
https://doi.org/10.12688/openreseurope.14922.2

Abstract

Home to more than twenty indigenous languages belonging to six linguistic families, the Gran Chaco has raised the interest of many linguists from different backgrounds. While some have focused on finding deeper genetic relations between different language groups, others have looked into similarities from the perspective of areal linguistics. In order to contribute to further research of areal and genetic features among these languages, we have compiled a comparative wordlist consisting of translational equivalents for 326 concepts — representing basic and ethnobiological vocabulary — for 26 language varieties. Since the data were standardized in various ways, they can be analyzed both quantitatively and qualitatively. In order to illustrate this in detail, we have carried out an initial computer-assisted analysis of parts of the data by searching for shared lexicosemantic patterns resulting from structural rather than direct borrowings.

Keywords

South American languages – Gran Chaco – comparative wordlist – structural borrowing

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 06 Dec 2022		✓ view
version 1 21 Jul 2022	✓ view	? view

1. **Joshua Birchall**, University of New Mexico, Albuquerque, USA

2. **Rik van Gijn**, Leiden University Centre for Linguistics, Leiden, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Horizon 2020 gateway.

This article is included in the Languages and Literature gateway.

This article is included in the European Research Council (ERC) gateway.

This article is included in the Linguistic Diversity collection.

Corresponding authors: Nicolás Brid (bridnicolas@gmail.com), Johann-Mattis List (mattis_list@eva.mpg.de)

Author roles: Brid N: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Messineo C**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing; **List JM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715618).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Brid N, Messineo C and List JM. **A comparative wordlist for the languages of The Gran Chaco, South America [version 2; peer review: 2 approved]** Open Research Europe 2022, **2**:90 https://doi.org/10.12688/openreseurope.14922.2

First published: 21 Jul 2022, **2**:90 https://doi.org/10.12688/openreseurope.14922.1

REVISED **Amendments from Version 1**

In this revised version, we have not modified the data, but rather tried to take the suggestions of the reviewers into account. As a result, the text contains some additional paragraphs in which we try to be a bit more transparent regarding the shortcomings of the data collection procedure we used in order to collect this dataset.

Any further responses from the reviewers can be found at the end of the article

(Plain language summary)

In this data note we present a list of words in indigenous languages of the Gran Chaco region in South America. These languages belong to arguably six established language families, whose deeper relationship is under discussion. Five of those language families are found only in the Gran Chaco, while one of them, the Tupi-Guarani family, is found across all of South America. In order to make it easy to compare the words in the wordlist, we standardized the data in several ways. We illustrate how the data can be analyzed by providing examples for cases in which words in unrelated languages show similar structures without being directly borrowed from each other.

Introduction

The Gran Chaco is a South American eco-region that extends through north-central Argentina, eastern Bolivia, western Paraguay and southern Brazil. It is located north of the Salado river, east of the Andes mountains, south of the Amazon, from which it is separated by the Chiquitania, and west of the Paraguay and Paraná rivers. Apart from languages that have entered the region through conquest and colonization, such as Spanish, German and Paraguayan Guaraní, the region is home to indigenous languages of six different families: Guaicuruan, which includes Toba, Western Toba, Pilagá, Mocoví, Kadiwéu and extinct Abipón; Matacoan or Mataguayan, which includes Wichí, Maká, Nivaclé, and Chorote; Enlhet-Enenlhet, which includes Enlhet, Enxet, Enenlhet, Guaná, Sanapaná and Angaité; Zamucoan, which includes Ayoreo and Chamacoco; Lule-Vilela, which includes only Lule and Vilela; and Tupi-Guarani, which in the Gran Chaco includes Tapiete, Ava, and Guaraní Izoceño but which also extends all through South America (Campbell & Grondona, 2012; Durante, 2018; Fabre, 2005; Golluscio & Vidal, 2010). For many of these languages there are also different geographic varieties.

The linguistic diversity of the Gran Chaco and the striking similarities in the features of some apparently unrelated languages have attracted the attention of numerous linguists, who have approached the topic from various theoretical and methodological frameworks. On the one hand, much research has focused on genetic relations among the languages. Recently, for instance, it has been stated that Vilela and extinct Lule are related and the family has been named Lule-Vilela (Viegas Barros, 2001), or that Guaicuruan and Matacoan languages have a common genetic origin and belong to one family, termed Guaicuruan-Matacoan (Viegas Barros, 1993; Viegas Barros, 2013a). Previous work had proposed even greater language family groupings (Kaufman, 1990; Mason, 1950). On the other hand, similarities among Chaco languages, not only Guaicuruan and Matacoan, have been analysed from the perspective of areal linguistics. Such similarities include phonological traits such as the presence and absence of certain phonemes, as well as grammatical features like the presence of possessive classifiers and noun determiners (Comrie *et al.*, 2010).

Fewer studies, however, have focused on shared semantic features that are visible in the lexicon in the form of similar lexical motivation patterns (Campbell & Grondona, 2012; Messineo *et al.*, 2010). In that sense, we consider that a big-scale dataset for further comparison of the Gran Chaco languages is a necessary tool that we have been lacking. Even though there have been many valuable works that compare different languages of the region, some of the criteria are inconsistent, and they seldom deal with the entirety of the indigenous languages of the Gran Chaco in a human and machine-readable way. Such an enterprise should be a starting point for a project that includes genetic comparison and concrete investigation of both lexical and pattern borrowing across Chaco languages of different families.

Materials and methods

Materials

Two different datasets were first individually compiled and later combined for this study. The first one comprised a list of 502 concepts reflecting basic vocabulary terms translated into 23 language varieties spoken in the Chaco area and two language varieties from other regions. The second one consisted of 825 ethnobiological concepts translated into 16 Chaco varieties. While the coverage for the basic dataset was rather high, with most languages showing word forms for 80% and more of the data, the coverage for the ethnobiological dataset was rather low, since the terms are highly specific and it was often difficult to find translations for all terms in resources available for the respective varieties. In order to allow for a more targeted comparison of the languages with respect to lexical structures, we then decided to combine them. This decision was motivated by the fact that — although previous research showing interesting cases of pattern borrowing in flora and fauna vocabulary had sparked our interest in that domain — we realized that the lexical motivation for the formation of individual terms still depends to a large degree on words and morphemes that can primarily be found in the realm ofbasic vocabulary. Thus, a combined list, albeit imperfect, permits a detailed study on pattern borrowing while taking lexical motivation patterns into account. For this purpose, we selected 224 concepts from the basic vocabulary lists, and 100 ethnobiological concepts, resulting in a total of 324 concepts for 23 language varieties (see Table 1), which are geographically distributed across and around the Chaco area (see Figure 1).

The collection of basic words was compiled from various sources, mainly dictionaries, but in some cases also from grammatical descriptions. One of the largest contributors was the Intercontinental Dictionary Series (IDS), (Key & Comrie, 2021).

Table 1. Languages and data points covered in our study.

#	Variety	Family	F	C	B	E	Co	Sources
1	Abipón	Guaicuruan	216	155	155	0	0.48	Najlis, 1966
2	Ava Guaraní	Tupian	263	215	215	0	0.66	Dietrich, 2021 (IDS)
3	Ayoreo	Zamucoan	377	228	212	16	0.70	Benz & Salinas Jacai Picanerai, 2020; Briggs, 2021 (IDS); Schmeda-Hirschmann, 1998
4	Chamacoco	Zamucoan	251	163	162	1	0.50	Ulrich & Ulrich (2000)
5	Enlhet	Enlhet-Enenlhet	438	252	216	36	0.78	Arenas, 1981; Unruh & Kalisch, 1997
6	Enxet Sur	Enlhet-Enenlhet	334	209	189	20	0.65	Rojas & Curtis, 2017
7	Guaraní Paraguayo	Tupian	325	238	214	24	0.73	Carol, 2018; Guasch & Ortiz, 1986; Seelwische, 1980
8	Iyojwa'ja Chorote	Matacoan	360	274	216	58	0.85	Drayson, 2009; Scarpa, 2010
9	Iyo'wujwa Chorote	Matacoan	254	190	176	14	0.59	Carol, 2018
10	Kadiweo	Guaicuruan	225	158	157	1	0.49	Griffiths, 2002; Sándalo, 1995
11	Lule	Lule-Vilela	296	174	174	0	0.54	Machoni & Larsen, 1877
12	Maká	Matacoan	282	243	199	44	0.75	Arenas, 1983; Gerzenstein, 1999
13	Mapudungun	Araucanian	256	207	207	0	0.64	Fernández Garay *et al.,* 2021
14	Mbya	Tupian	223	168	168	0	0.52	Cadogan, 1992
15	Mocoví	Guaicuruan	298	216	213	3	0.67	Buckwalter & Ruiz, 2021; Rosso, 2010
16	Nivaclé	Matacoan	376	250	217	33	0.77	Seelwische, 1980
17	Pilagá	Guaicuruan	287	248	211	37	0.77	Buckwalter & Suárez, 2021; Filipov, 1993; Vidal, 2010 and Vidal, 2013
18	Quichua Santiagueño	Quechua	235	176	162	14	0.54	Bravo, 1975
19	Tapiete	Tupian	272	202	194	8	0.62	González, 2005; González, 2011
20	Toba	Guaicuruan	471	273	216	57	0.84	Buckwalter & Litwiller de Buckwalter, 1980; Buckwalter & Sánchez, 2021; Cúneo & Porta, 2009; Martínez, 2009
21	Toba de Cerrito	Guaicuruan	180	154	154	0	0.48	Messineo, 2009
22	Toba-pilagá	Guaicuruan	368	255	192	63	0.79	Arenas, 1993; Tebboth, 1943
23	Wichí	Matacoan	388	241	209	32	0.74	Braunstein, 2021 (IDS), DIWICA (2021); Suárez, 2010 and Suárez, 2014

Column F refers to the forms in the data, column C refers to the concepts that are covered, columns B and E refer to the number of concepts covered from basic and ethnobiological vocabulary, and column Co refers to the coverage (number of attested concepts divided by number of concepts in the whole wordlist).

Other material came from individual sources available for the respective varieties, mainly dictionaries, wordlists, and compilations of different Chaco languages. In these cases, translational equivalents for the basic words were carried out manually. The collection of ethnobiological terms was typically compiled from specific lists of ethnobiological vocabulary, taken from articles and books dedicated to the topic, but in some cases, unified resources for basic vocabulary and ethnobiological terms were available and could be used.

Methods

In creating our resource, we had two major goals in mind. On the one hand, we wanted to create a resource that is both human- and machine-readable at the same time, allowing us to analyse the data and annotate particular findings step by step in future work (this process is ongoing work and might be featured in studies to be published in the future). On the other hand, we wanted to create a resource that can be easily compared with other lexical resources, both on a world-wide and a regional

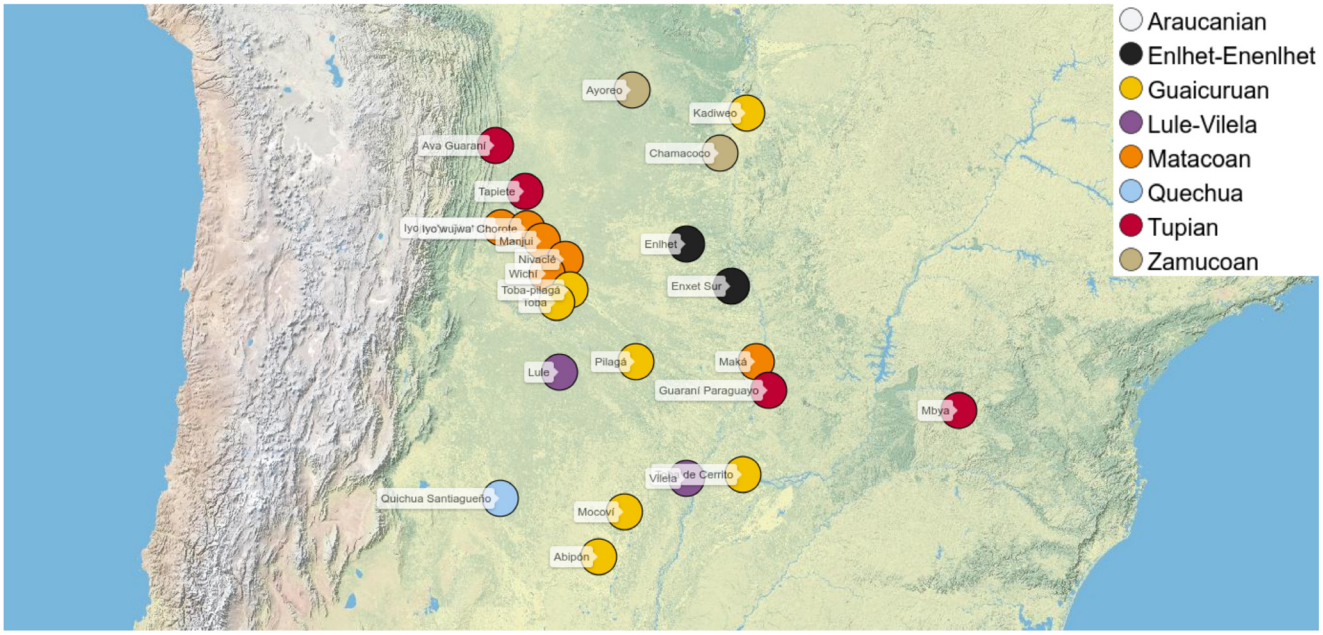

Figure 1. Languages covered in our study (with exception of Mapudungun, which is located further in the South).

scale. This allows us to make use of additional information or to compare our findings with those reported for other areas of the world in our future work. In order to achieve the first goal, we used an internal representation of the data for analysis and annotation, based on the Etymological Dictionary Edictor (EDICTOR, Version 2.0, List, 2021a), in which we curate the data manually, annotating the data for various aspects, such as cognacy, borrowings, or borrowed patterns (loan translations) shared across the Chaco languages. In order to achieve the second goal, we converted our data to Cross-Linguistic Data Formats (CLDF, Forkel *et al.*, 2018), using the Lexibank workflow for the curation of lexical data in CLDF (List *et al.*, 2022a). While data curation and annotation with the help of the EDICTOR tool were largely done in a manual fashion, the conversion to CLDF was mostly done automatically, providing additional steps that helped us to identify potential problems in our data.

Data curation with EDICTOR

Basic vocabularies and ethnobiological vocabularies were first collected separately. Only later, when we realized that both can be better analyzed in combination, we decided to combine them. For this purpose, we decided for a combined list of 324 items, with 224 basic vocabulary items and 100 ethnobiological items in total. Both datasets were combined to form a single TSV file in the format required by the EDICTOR tool and converted to an SQLITE database, using the PyEdictor package (List, 2021b, Version 0.4), which we use to allow for the convenient online editing of the data.

Our main intention for the analysis was to annotate structural borrowings, that is, cases of borrowings in which it is not the word form that is being transferred, but rather the lexical motivation by which certain objects can be denoted. As an example, consider the English term "(computer) mouse", which is reflected as *ratón de computadora* (literally "mouse or rat of the computer") in Spanish.

In order to annotate structural borrowings in the Chaco data, we made use of existing annotation schemes that were developed for the handling of partial cognates (Hill & List, 2017) and later extended to handle more complex cases of language-internal cognates and semantic shift (Schweikhard & List, 2020) and ultimately implemented in Version 2.0 of the EDICTOR tool (List, 2021a). The main idea of these annotation schemes is to provide what we call 'morpheme glosses' for each word form in the data and combine these with identifiers for partial cognates (see List *et al.*, 2016).

As an example, consider the words for "beak" and "lip" for Maká and Chorote (both from the Matacaon language family) and Pilaga (from the Guaicuruan family) in Table 2. As can be seen from the table, all three language varieties express the word for "beak" by using the entire word or a part of the word for "lip". Since Pilaga is not related with Chorote and Maká, and the form that expresses the concept "lip" in Pilaga ([a s e p], according to our annotation) is not cognate with the form [p a s] in Chorote and Maká, we assign these forms different cognate set identifiers (2 for [a s e p] and 4 for [p a s]). But since we judge the pattern as identical, consisting of a possessive marker (marked as :poss in our morpheme glosses) and the reuse of the form "lip" to denote the concept "beak", we assign them the same pattern identifier, indicating that we have a shared structure here. Whether this structural commonality is due to language contact or due to independent processes of lexical change cannot be said at this point, since the pattern annotation

Table 2. Example of our extended annotation of cognate sets, with morpheme glosses and structural similarities with respect to the motivation structure of individual word forms.

Family	Language	Concept	Form	Cognates	Structure	Morpheme Glosses
Guaicuruan	Pilaga	beak	n - a s e p	1 2	1 2	:poss **lip**
Guaicuruan	Pilaga	lip	n - a s e p	1 2	1 2	:poss **lip**
Matacoan	Chorote	beak	x i - p a s - a t	3 4 5	1 2 3	:poss **lip** :suff
Matacoan	Chorote	lip	x i - p a s - a t	3 4 5	1 2 3	:poss **lip** :suff
Matacoan	Maká	beak	ɬ a - p a s	6 4	1 2	:poss **lip**
Matacoan	Maká	lip	p a s	4	2	l**ip**

is work in progress and has not been done for all of the data. Assembling more of these patterns in our data, however, will eventually allow us to find out whether these scenarios might result from contact or not.

Table 2 shows words for "beak" and "lip" across three varieties from two language families. While word forms are not cognate across the two language families, and also not borrowed directly, we find structural similarities with respect to the motivation. In all three varieties, our annotation assumes that the word for "beak" is derived from the word for "lip". We indicate this structural commonality with the help of identifiers that reflect the abstract structure (column Structure) and with the help of morpheme glosses, that provide an analysis of the underlying motivation (column Morpheme Glosses). Note that our analysis is not the only possible one for the given data. One could likewise argue or speculate that the word for "beak" was primary and that the word for "lip" was derived from it. In this case, the morpheme glosses would have to be modified. In order to avoid being forced to make a decision on the primary word form, one can — finally — also use neutral morpheme glosses like "beak/lip" which would explicitly avoid to make any judgment regarding primary or secondary word forms in the data.

Data Sharing with CLDF

Whenever substantial changes to the data have accumulated and we decide to release a new version, we export the dataset and convert it automatically to CLDF. In doing so, we carry out several consistency checks of the data and make sure that the individual datapoints are maximally comparable across datasets from different sources. The CLDF conversion is carried out with the help of the CLDFBench toolkit that offers a command line interface that facilitates the conversion of language data to CLDF formats (Forkel & List, 2020, https://pypi.org/project/cldfbench). Since we are working with lexical data, we additionally use the PyLexibank plugin for CLDFBench (Forkel *et al.*, 2021), which offers extended functionality (see List *et al.*, 2022a). The conversion to CLDF makes sure that our concepts are regularly linked to the most recent version of the Concepticon reference catalogue (List *et al.*, 2022b), that all languages, where possible, are linked to Glottolog (Hammarström *et al.*, 2022), and that the transcriptions follow the standards proposed by the Cross-Linguistic Transcription Systems reference catalogue (List *et al.*, 2021). Since the CLDF standard currently does not (yet) offer standards to annotate structural borrowings, we define custom formats for now (see Table 2), which we will propose for the inclusion in future versions of CLDF. In the following, we discuss the integration of our data with the three reference catalogs of (Concepticon, Glottolog, and CLTS) in more detail.

Concept linking. The concept list underlying our study was linked to the Concepticon reference catalogue (Version 2.6, List *et al.*, 2022b). Concepticon offers unique identifiers for various concepts that are frequently used in questionnaires for language documentation and historical language comparison. Since Concepticon is by now more and more often used as a common standard reference for lexical datasets, also underlying large collections such as the Database of Cross-Linguistic Colexifications (CLICS) (Rzymski *et al.*, 2020) or the Lexibank repository of standardized wordlists in CLDF formats (List *et al.*, 2022a), we also made sure to link the concepts in our data to Concepticon, where possible. For the very specific plant and animal names in our data, however, the Concepticon does not offer concept identifiers. Here, we therefore linked our data to the Global Biodiversity Information Facility (GBIF).

Language mapping. Another way of linking the data with already existing sources consists in the linking of language varieties to the Glottolog project (Hammarström *et al.*, 2022). Glottolog provides unique identifiers for several language varieties, including dialect points and ancient varieties along with additional information regarding the language families to which the respective languages belong. For two varieties in our data, no Glottocode could be found. These are Manjui, which is a variety of Chorote spoken in the territory of Paraguay, and Toba de Cerrito, also spoken in the Paraguayan Chaco. These have not been identified as separate varieties on Glottolog yet, but might be added in future versions.

Most of the languages in our dataset are spoken in the Gran Chaco region of South America, in the territories of Argentina, Bolivia, Brazil, and Paraguay. In addition, we have chosen three languages spoken in adjacent regions, which we hope to use as control cases in future analyzes, namely Mapudungun (Araucanian), spoken in southern Chile and Argentina, Mbyá (Tupí-Guaraní), spoken in Argentina, Brazil and Paraguay, and Quichua Santiagueño (Quechuan), spoken in north-central Argentina. Although we are aware that these languages are spoken in the vicinity of the Gran Chaco, their inclusion as control languages responds to the fact that we intend to find shared semantic patterns that are not even found in adjacent territories. However, while some patterns have been observed in our data only in the Gran Chaco languages, others do appear also in the control languages. While it is true that areal influence does not end abruptly, and thus those coincidences could also be due to language contact exceeding the limits of the Gran Chaco, this could also be explained by the fact that not all shared semantic patterns are equally ubiquitous, with some patterns being more likely shared due to common typological traits in the world's languages. This point, and the need for a hierarchy on pattern borrowing in order to rank the evidence by strength, is discussed in the conclusion. Even so, future studies should include control languages spoken in additional locations (in and out of South America) in order to render the results more robust. Finally, Paraguayan Guaraní is usually not considered a Chaco language in origin, but it has an undeniable influence on indigenous communities of the Gran Chaco, especially in the territory of Paraguay, where it is the second and sometimes the first language of many indigenous people who are multilingual in other languages.

When searching for the translational equivalents of individual concepts in our concept lists in the different sources for the varieties we included in our sample, it is often difficult to decide which word corresponds best to a given concept, specifically in cases where one has to choose from several variants. Variants may result from several reasons. On the one hand, two translations for the same concept may correspond to different varieties that have been included in the same resource. For example, we have added a document for a variety of Toba spoken in Paraguay, Toba de Cerrito. However, this variety has two subvarieties, one spoken in the village of Rioverde and the other spoken in the village of Rosario. In those cases in which these subvarieties display different forms, we indicate in a comment which form corresponds to which variety. In future versions of the database, we plan to find more principled ways of handling this kind of dialectal variation. On the other hand, different resources may give different forms for the same concept but no indication in which regard the forms differ (e.g., regarding their usage, specific semantic nuances, etc.). In these cases we indicated the different sources in our comments, but hope to find a more principled way to handle these cases of variation in future versions of our database.

This study includes Lule and Abipón, two extinct varieties of which no speakers are known to have survived until today. The original sources of these varieties were written by missionaries in the eighteenth and nineteenth centuries. Since transcription practices differed largely in the past, we cannot fully account for the accuracy of the transcriptions we used. Including the varieties in the study has proven useful, however, since it allowed us to check whether certain kinds of semantic patterns existed already 300 or 200 years before.

Phonetic transcriptions. After having compiled the vocabulary in the corresponding sheets, the forms were converted, into a broad version of the International Phonetic Alphabet, called B(road)IPA, the central transcription system underlying the five transcription systems provided in the CLTS reference catalog. For the initial conversion, we made use of orthography profiles (Moran & Cysouw, 2018), which are integrated into the Lexibank workflow for the curation of lexical data, which we used for our study (List *et al.*, 2022a). In this workflow, original forms are preserved, and for the target phonetic transcriptions used for cross-linguistic comparison, automatic tests are carried out to make sure they only reflect sounds defined in the CLTS reference catalog.

The conversion of transcription systems used by individual scholars to standardized transcriptions that conform to CLTS can be considerably tedious, especially when different transcription systems are underlying the data from every source. The conversion therefore required an intensive study of the phonological descriptions of all language varieties in our sample, for which often information often could only be found in broader grammatical descriptions. Inspecting the data also revealed that our initial conversion to phonetic transcriptions with orthography profiles was at times not optimal or contained occasional errors, which we then had to refine manually by modifying the data in the EDICTOR application. For the two extinct languages in our collection, Lule and Abipón, no reliable phonological descriptions available. In the case of Abipón, we followed the description of on phonology in Viegas Barros (2013b), based on comparison with other Guaicuruan languages. For Lule, we followed Zamponi's analysis from 2008.

Implementation

Having set up the data in its current form, our workflow for data curation and analysis now consists of two steps. In a first step, the data is analyzed using the EDICTOR tool. Figure 2 shows how the data appear in the Wordlist panel of the EDICTOR interface. In order to share the data publicly, we then used the Lexibank workflow (List *et al.*, 2022a) to convert the data automatically into Cross-Linguistic Data Formats, which can be triggered from the commandline. The conversion automatically checks various aspects of the data, including the transcriptions as reflected in a given version of the CLTS reference catalog, the mapping to a given Glottolog version and a given Concepticon version, and the formal correctness of currently available annotations.

Conclusion

Although we consider the collection of the dataset reported here as preliminary, it has reached a stage where we can start with the concrete analysis of individual patterns in the data (Brid *et al.*, 2022). In the future, we plan to enhance the current

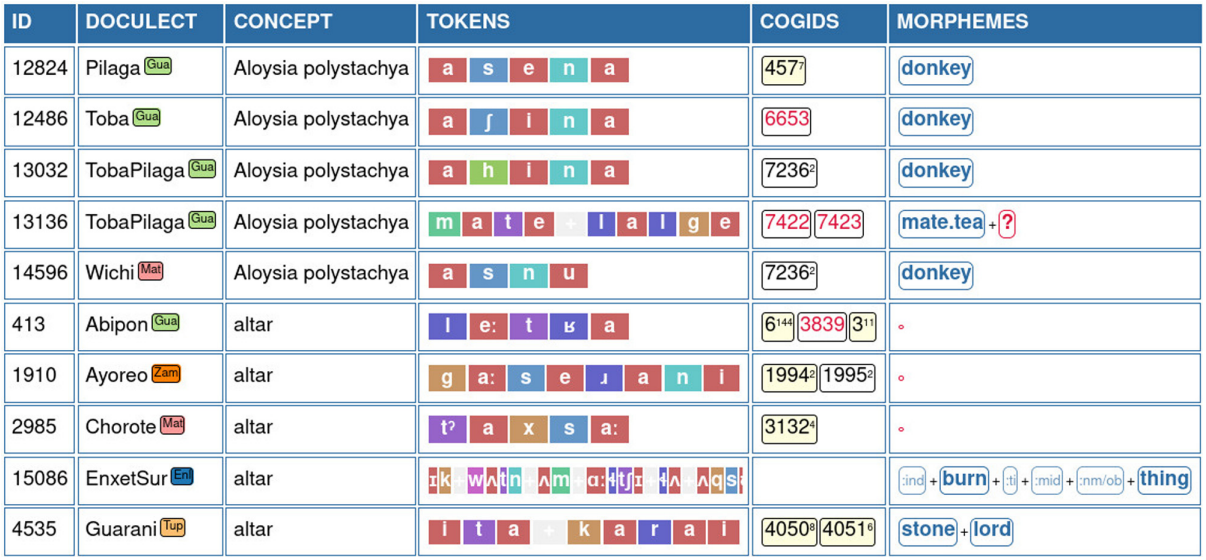

Figure 2. Curating the data with the help of the EDICTOR interface. The screenshot shows the Wordlist panel view of the EDICTOR tool. Word forms are rendered by coloring speech sounds according to their major sound class.

dataset further and also extend the annotation of cognate words and structural borrowings.

Although we consider the dataset as good enough to publish it at this point, we should make clear that we are not fully content with all decisions we undertook in the past when collecting our data. By explicitly pointing to these points of dissatisfaction, we hope that we can warn readers of this study to avoid our mistakes when conducting similar works.

Firstly, we warn future researchers against mixing multiple sources for the same language varieties with no overt indication. For instance, our Chorote, Wichí, and Ayoreo data come from different sources. Although it may be important to include multiple sources, it would be advantageous to include a reference to the source in the database, perhaps in a separate column. This would make a discussion of the data and the underlying decisions which led to their creation more transparent. Also, it may turn out that a source differs from another source because it is based on a different language variety, perhaps more in contact with another language of the region. In suchhh a case, having that information at one's disposal would be highly relevant for the results.

Even if sources are overtly indicated, a future reader would have to find the entries in the source. However, at present our data is not visible in its original orthography. For that reason, we encourage similar projects in the future to keep the original transcription in a separate column. This would enable users to copy-paste the original form in order to look it up in the original source. We plan to solve these two issues in the future, but at this stage, our data curation process had advanced too much to allow us for handling these problems efficiently.

Finally, it would also be desirable to rank the evidence for borrowingby strength. This means that, in order to address the topic of areal influence on shared semantic patterns, one would like to be able to tell the difference between patterns that may be shared due to typological traits common to the world's languages and patterns that are more likely shared due to areal influence. This requires a theoretical and methodological apparatus that permits to suppose some kind of hierarchy on pattern borrowing. Since — to the best of our knowledge — such an apparatus does not exist at the moment, we can only hope on future research to provide us with additional tools to enhance the analysis of our datasets.

Ethics and consent statement

Ethical approval and consent were not required.

Data and software availability

Data and Software available from: https://github.com/lexibank/chacolanguages/

Archived source code and data at time of publication: https://doi.org/10.5281/zenodo.6660368

Acknowledgements

We thank Paola Cúneo for helpful comments on our data analysis, as well as Temis Tacconi for help with the organization of the Maká material. We also thank our two reviewers for very helpful and thoughtful comments.

References

Arenas P: **Etnobotánica lengua-maskoy [Lengua-Maskoy ethnobotanics]**. Buenos Aires: Fundación para la Educación, la Ciencia y la Cultura. 1981.

Arenas P: **Nombres y usos de las plantas por los indígenas Maká del Chaco Boreal [Names and uses of plants by the Maká Indians of the Chaco Boreal].** In: *Parodiana.* Buenos Aires: Asociación Parodiana. 1983; **2**(2): 131–229.
Reference Source

Arenas P: **Fitonimia toba-pilagá [Toba-Pilagá phytonymy].** In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco V.* Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 1993; 75–100.

Benz EA, Salinas Jacai Picanerai J: **Diccionario Ayoeode Uuode – Español – Español – Ayoeode Uuode [Ayoreo – Spanish dictionary]**. Asunción: Fondo Nac6ional de la Cultura y las Artes.

Braunstein J: **Wichî dictionary**. In: Key Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Bravo D: **Diccionario quichua santiagueño-castellano [Santiago del Estero Quichua – Spanish dictionary]**. Buenos Aires: Editorial Universitaria de Buenos Aires. 1975.
Reference Source

Brid N, List JM, Messineo C: **Las lenguas del Chaco desde la perspectiva de la semántica léxica. Análisis preliminar de patrones léxicos compartidos en el dominio etnobiológico [The languages of the Gran Chaco from the perspective of lexical semantics. Preliminary analysis of shared lexical structures in the ethnobotanical domain].** *LIAMES.* **2022; 22:** e022005**, 1–21.**
Publisher Full Text

Briggs J: **Ayoreo dictionary**. In: Key, Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Buckwalter A, Litwiller de Buckwalter L: **Vocabulario toba.** Buenos Aires: Talleres Gráficos Grancharoff. 1980.
Reference Source

Buckwalter A, Sánchez O: **Toba dictionary**. In: Key Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Buckwalter A, Ruiz R: **Mocovî dictionary**. In: Key Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Buckwalter A, Suárez J: **Pilagá dictionary**. In: Key Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Cadogan L: **Diccionario mbyá guaraní – castellano [Mbya Guarani - Spanish dictionary]**. Asunción: CEADUC. 1992.
Reference Source

Campbell L, Grondona V: **Languages of the Chaco and Southern Cone.** In: *The indigenous languages of South America: A comprehensive guide.* Berlin: De Gruyter Mouton. 2012; **2**: 625–667.
Publisher Full Text

Carol J: **Inamtes jleeizi' Inkijwas ji'lij - Kiláyi ji'lij: Diccionario Bilingüe Manjui - Castellano [Manjui –Spanish bilingual dictionary]**. Asunción: Paraguái Ñe'ẽnguéra Sãmbyhyha. 2018.

Comrie B, Golluscio L, Vidal A, *et al.*: **El Chaco como área lingüística [Chaco as a linguistic area].** In: *Estudios de lenguas amerindias.* Hermosillo, Sonora: Editorial Unison. 2010; **2**: 85–130.
Reference Source

Cúneo P, Porta A: **Vocabulario toba sobre peces y aves [Toba vocabulary of fish and birds].** In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco.* Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 2009; **VIII**: 237–252.

Dietrich W: **Chiriguano dictionary**. In: Key Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

DIWICA: **Wichi-siwele lhayhilh / Diccionario wichí-castellano [Wichí –Spanish dictionary]**. Formosa: INILSyT. 2021.
Reference Source

Drayson N: **'Niwak Samtis: Diccionario Iyojwa'ja 'Lij-Kilay 'Lij (Chorote-Castellano) [Chorote - Spanish dictionary]**. In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco*. Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 2009; **VIII**: 91–174.

Durante S: **La lengua ayoreo (familia zamuco), de la sintaxis al discurso: Documentación y descripción de una lengua amenazada [The Ayoreo language, from syntax to discourse: documentation and description of an endangered language]**. Buenos Aires: Facultad de Filosofía y Letras. 2018.
Reference Source

Fabre A: **Los pueblos del Gran Chaco y sus lenguas, primera parte: Los enlhet-enenlhet del Chaco Paraguayo [The Gran Chaco peoples and their languages, first part: the Enlhet-Enenlhet of the Paraguayan Chaco]**. In: *Centro de Estudios Antropologicos. Suplemento Antropologico*. Asunción: Universidad Católica Nuestra Señora de la Asunción. 2005; **40**(1): 503–569.
Reference Source

Fernández Garay A, Catrileo M, Ritchie Key M: **Mapudungun dictionary**. In: Key, Mary Ritchie and Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Filipov A: **Fitonimia pilagá [Pilaga phytonymy]**. In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco*. Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 1993; **V**: 101–119.

Forkel R, Greenhill SJ, Bibiko HJ, *et al.*: **PyLexibank. The Python Curation Library for Lexibank [Software, Version 2.8.2]**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

Forkel R, List JM: **CLDFBench. Give your Cross-Linguistic data a lift**. In: N. Calzolari, F. Béchet, P. Blanche, K. Choukri, C. Cieri, T. Declerck, et al. (Eds.) *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. Paris: European Language Resources Association (ELRA). 2020; 6997–7004.
Reference Source

Forkel R, List JM, Greenhill S, *et al.*: **Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics**. *Sci Data.* 2018; **5**: 180205.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gerzenstein A: **Diccionario etnolingüístico maká-español [Ethnolinguistic Maká - Spanish dictionary]**. Buenos Aires: Archivo de Lenguas Indoamericanas. 1999.
Reference Source

Golluscio L, Vidal A: **Recorrido sobre las lenguas del Chaco y los aportes a la investigación lingüística [The Chaco languages and their contribution to linguistic research]**. In: *Amerindia*. Paris: Association d'Ethnolinguistique Amérindienne. 2010; **33/34**: 3–40.
Reference Source

González H: **A grammar of Tapiete (Tupi-Guarani)**. Doctoral dissertation, University of Pittsburgh. 2005.
Reference Source

González HA: **Léxico etnobotánico tapiete (tupí-guaraní), lengua del Chaco argentino [Ethnobotanic vocabulary of Tapiete, a language of the Argentine Chaco].** *Indiana.* 2011; **28**: 255–288.
Publisher Full Text

Griffiths G: **Dicionário da língua Kadiwéu: Kadiwéu- Português, Português-Kadiwéu [Kadiwéu language dictionary]**. Cuiabá: Sociedade Internacional de Linguística. 2002.
Reference Source

Guasch A, Ortiz D: **Diccionario Guaraní-Castellano Castellano-Guaraní [Guarani - Spanish dictionary]**. Asunción: CEPAG. 1986.
Reference Source

Hammarström H, Forkel R, Haspelmath M, *et al.*: **Glottolog 4.6. Leipzig: Max Planck Institute for Evolutionary Anthropology.** 2022.
Publisher Full Text

Hill N, List JM: **Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages.** *Yearbook of the Poznań Linguistic Meeting.* 2017; **3**(1): 47–76.
Publisher Full Text

Kaufman T: **Language history in South America: What we know and how to know more**. In: Payne, D. (Ed.). *Amazonian Linguistics: Studies in Lowland South American Languages*. Austin: University of Texas Press. 1990; 13–67.
Reference Source

Key MR, Comrie B: **The Intercontinental Dictionary Series**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021.
Reference Source

List JM, Anderson C, Tresoldi T, *et al.*: **Cross-Linguistic Transcription Systems.** Version 2.1.0. Max Planck Institute for the Science of Human History: Jena. 2021.
Publisher Full Text

List JM, Forkel R, Greenhill S, *et al.*: **Lexibank, A public repository of standardized wordlists with computed phonological and lexical features**. *Sci Data.* 2022a; **9**(316): 1–31.
Publisher Full Text

List JM, Lopez P, Bapteste E: **Using sequence similarity networks to identify partial cognates in multilingual wordlists**. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin: Association of Computational Linguistics. 2016; **2**: 599–605.
Publisher Full Text

List JM: **EDICTOR. A web-based interactive tool for creating and editing etymological datasets.[Software, Version 2.0]**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021a.
Reference Source

List JM: **PyEDICTOR. A tool for the quick manipulation of CLDF datasets.** Leipzig: Max Planck Institute for Evolutionary Anthropology. 2021b.
Reference Source

List JM, Tjuka A, Rzymski C, *et al.*: **CLLD Concepticon [Dataset, Version 2.6.0]**. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2022b.
Publisher Full Text

Machoni A, Larsen JM: **Arte y vocabulario de la lengua lule y tonocoté: compuestos con facultad de sus superiores.** Buenos Aires: PE Coni. 1877.
Reference Source

Martínez G: **Fitonimia de los tobas bermejeños (Chaco Central, Argentina) [Phytonymy of the Bermejo Tobas of the Argentine Central Chaco]**. In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco*. Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 2009; **VIII**: 194–212.
Reference Source

Mason JA: **The Languages of South American Indians**. In: *Handbook of South American Indians*. Washington: United States Government Printing Office. 1950; **6**: 189–215.
Reference Source

Messineo C: **Vocabulario toba de Cerrito (Paraguay) [Toba vocabulary of Cerrito, Paraguay]**. In: Braunstein, José and Messineo, Cristina (eds.), *Hacia una nueva carta étnica del Gran Chaco*. Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 2009; **VIII**: 253–269.

Messineo C, Scarpa G, Tola F: **Léxico y categorización etnobiológica en grupos indígenas del Gran Chaco [Ethnobiological vocabulary and categorization among indigenous groups of the Gran Chaco]**. Santa Rosa: Universidad Nacional de La Pampa. 2010.
Reference Source

Moran S, Cysouw M: **The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles**. Berlin: Language Science Press. 2018.
Reference Source

Najlis EL: **Lengua abipona.** Archivo de lenguas precolombinas Buenos Aires 1.1-2. 1966.
Reference Source

Rojas A, Curtis T: **Diccionario Enxet Sur [Enxet Sur dictionary]**. Río Verde: Equipo de Traducción de Enxet Sur. 2017.

Rosso C: **Compilación y análisis preliminar de la fitonimia de la flora leñosa de comunidades mocovíes del sudoeste chaqueño [Compilation and preliminary analysis of woody flora phytonymy in Mocovi communities of Southwestern Chaco].** In: Messineo, C., Scarpa, G y Tola, F. (comps.), *Léxico y categorización etnobiológica en grupos indígenas del Gran Chaco.* Santa Rosa: Universidad Nacional de La Pampa. 2010; 251–272.

Rzymski C, Tresoldi T, Greenhill S, *et al.*: **The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies.** In: *Sci Data.* 2020; **7**(13): 13.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Sandalo MF: **A Grammar of Kadiweu.** Doctoral dissertation, University of Pittsburgh. 1995.
Reference Source

Scarpa G: **Hacia una etnotaxonomía vegetal chorote II: Clasificación de las plantas entre las parcialidades iyojwa'ja y iyowujwa del Chaco argentino [Towards a Chorote vegetal ethnotaxonomy II: plant classification among the Iyojwa'ja and Iyowujwa groups of the Argentine Chaco].** In: Messineo, C., Scarpa, G y Tola, F. (comps.), *Léxico y categorización etnobiológica en grupos indígenas del Gran Chaco.* Santa Rosa: Universidad Nacional de La Pampa. 2010; 157–198.
Reference Source

Schmeda Hirschmann G: **Etnobotánica Ayoreo. Contribución al estudio de la flora y vegetación del Chaco. XI. [Ayoreo ethnobotanics. Contribution to the study of the Chaco flora and vegetation. XI].** *Candollea.* 1998; **53**(1): 1–50.
Reference Source

Schweikhard NE, List JM: **Developing an annotation framework for word formation processes in comparative linguistics.** In: *SKASE Journal of Theoretical Linguistics.* 2020; **17**(1): 2–26.
Reference Source

Seelwische J: **Diccionario Nivaclé-Castellano [Nivacle – Spanish dictionary].** Asunción: CEADUC. 1980.
Reference Source

Suárez ME: **Fitonimia wichí de especies arbóreas y arbustivas del Chaco Semiárido salteño [Wichí phytonymy of trees and bushes of the semi-arid Chaco Salteño].** In: Messineo, C., Scarpa, G y Tola, F. (comps.), *Léxico y categorización etnobiológica en grupos indígenas del Gran Chaco.* Santa Rosa: Universidad Nacional de La Pampa. 2010; 199–224.
Reference Source

Suárez ME: **Etnobotánica wichí del bosque xerófito en el Chaco Semiárido salteño [Wichi ethnobotanics of the xerophyte woods of the semi-arid Chaco Salteño].** Don Torcuato: Autores de Argentina. 2014.
Reference Source

Tebboth T: **Diccionario toba [Toba dictionary].** In: *Revista del Instituto de Antropología de Tucumán.* Tucumán: Universidad Nacional de Tucumán. 1943; **3**(2): 33–221.
Reference Source

Ulrich M, Ulrich R: **Diccionario Ɨshiro (Chamacoco) – Español / Español – Ɨshiro (Chamacoco) [Spanish – Chamacoco dictionary].** Asunción: New Tribes Mission. 2000.

Unruh E, Kalisch H: **Moya'ansaeclha'nengelpayvaam nengeltomha enlhet.** Comunidad Enhlet. 1997.
Reference Source

Vidal A: **Diccionario Trilingüe Pilagá-Español-Inglés Interactivo [Interactive trilingual dictionary Pilaga – Spanish – English].** Formosa: EDUNAF. 2010.

Vidal A: **Enseñanza de la lengua pilagá [Pilaga language teaching].** Formosa: EDUNAF. 2013.
Reference Source

Viegas Barros P: **¿Existe una relación genética entre las lenguas mataguayas y guaycurúes? [Is there a genetic relationship between Mataguayan and Guaicuruan languages?].** In: Braunstein, José and Cristina Messineo (eds.), *Hacia una nueva carta étnica del Gran Chaco V.* Las Lomitas, Formosa: Centro del Hombre Antiguo Chaqueño. 1993; 193–213.
Reference Source

Viegas Barros P: **Evidencias de la relación genética lule-vilela [Evidence for the genetic relationship between Lule and Vilela].** *LIAMES: Línguas Indígenas Americanas.* Campinas: UNICAMP. 2001; **1**(1): 107–126.
Reference Source

Viegas Barros P: **La hipótesis de parentesco Guaicurú-Mataguayo: estado actual de la cuestión [The Mataguayo-Guaicuruan relatedness hypothesis: current state of affairs].** *Revista brasileira de linguística antropológica.* Brasilia: Universidade de Brasília. 2013a; **5**(2): 293–333.
Publisher Full Text

Viegas Barros P: **Proto-Guaicurú: Una reconstrucción fonológica, léxica y morfológica [Proto-Guaicuruan: a phonological, lexical, and morphological reconstruction].** Munich: Lincom Europa. 2013b.
Reference Source

Zamponi R: **Sulla fonologia e la rappresentazione ortografica del lule.** In: *Introducción de Riccardo Badini y Raoul Zamponi a Maccioni Antonio (2008 [1732]) Arte y Vocabulario de la Lengua Lule y Tonocoté, edición al cuidado de Riccardo Badini, Tiziana Deonette, Stefania Pineider, XXI-LVIII.* Cagliari: Centro di Studi Filologici Sardi. 2008.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 18 January 2023

https://doi.org/10.21956/openreseurope.16643.r30464

✔

Rik van Gijn
Leiden University Centre for Linguistics, Leiden, The Netherlands

I am happy to change the status of the paper to Approved

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 10 October 2022

https://doi.org/10.21956/openreseurope.16126.r29747

?

Rik van Gijn
[1] Leiden University Centre for Linguistics, Leiden, The Netherlands
[2] Leiden University Centre for Linguistics, Leiden, The Netherlands

I think this is an interesting initiative, which fills a gap in the comparative linguistic research domain: assessing the similarities between semantic structure of lexemes across languages. I do have a few remarks, questions, and things that were not quite clear to me for consideration.

1. The word list

The word list consists of basic vocabulary (the largest part) and another part which is ethnobotanical vocabulary. The paper mentioned that the authors realized that these two parts were best combined into one. Can you say more about the reason behind this, why is it better to combine them? And did or do you have any expectations wrt borrowability of semantic structure of basic vocabulary versus more peripheral (ethnobotanical) vocabulary. When it comes to form borrowing, the received wisdom is that basic vocabulary is more resistant to borrowing than flora and fauna vocabulary, but is there an equivalent expectation for structural borrowing?

2. Semantic coding
I understand that, in the example given, the words for lip and beak are both assigned to the concept LIP. The important pattern here seems to be that words for beak and lip are connected to one and the same concept, and it wasn't entirely clear to me what happens if two languages have one and the same underlying concept for both lip and beak, but in slightly different ways. I'll sketch two hypothetical scenarios. One (admittedly unlikely) scenario is that there might be independent evidence that in fact BEAK is the original meaning and that it is more truthful to connect both the words for lip and beak to the concept BEAK. Would that count as a full mismatch with the languages that connect both words to the concept of LIP? Another scenario would be that the word for lip is in fact connected to the concept MOUTH, e,g, the word for lip might semantically be something like OUTER MOUTH, this in turn may be extended to the word for beak, but in this case both words are connected to a third concept MOUTH. Is that also a full mismatch with a language that extends LIP to the word for beak?

3. Types of semantic structural isomorphisms
Related to the previous point, do I understand correctly that the obligatory presence of a possessive prefix with words for lip and beak in the example count as much for a match on structural borrowing as the fact that both words are connected to the concept LIP? It seems to me that one match is more indicative about past contact than the other 9in general: the more specific and unusual, the more informative). Or do you have ways to differentiate between different types of structural matches?

4. You mention a number of control languages. These are all spoken in the immediate vicinity of the Gran Chaco, and I don't think it can be excluded that there were contacts between speakers of the control languages and the target languages. So I wonder to what extent are these control languages. Are they meant to show a diminished number of commonalities, or are they meant to give a baseline of accidental commonalities? If the latter, I think the control language are not the best choices. In any case, it is good to make this clear.

5. I wonder if it is not a little too early to publish this paper. The concluding remarks suggest that the coding scheme can still change considerably. I cannot tell how much it may change and to what extent it would make the present publication obsolete, so I leave this for the consideration of the authors.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research interests include South American languages, reconstructing the social history in South America, language typology. I feel confident to assess the conceptual set up of the paper, but not to assess the technical details of the implementation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Oct 2022

Johann-Mattis List

We are very grateful for this thorough review and will try to address all points raised soon, hoping that our answers will be convincing and that our modified version of the article will properly address all points raised.

Competing Interests: No competing interests were disclosed.

Author Response 02 Dec 2022

Johann-Mattis List

We would like to express our deep gratitude to the reviewer for taking the time to check our study and providing very profound and important comments. We have tried to directly react to the major criticisms brought up by the reviewer in our revised version of the study. We list these in the following in the form of bullet points in which we partially quote the points brought up by the reviewer.

- "Can you say more about the reason behind this, why is it better to combine them? And did or do you have any expectations wrt borrowability of semantic structure of basic vocabulary versus more peripheral (ethnobotanical) vocabulary. When it comes to form borrowing, the received wisdom is that basic vocabulary is more resistant to borrowing than flora and fauna vocabulary, but is there an equivalent expectation for structural borrowing?"

While we were initially mostly interested in an investigation of Flora and Fauna vocabulary, since previous research indicated that the most interesting patterns of pattern borrowing could be found in this area of the lexicon, we later realized that the motivation of individual terms still depends to a large degree on words and morphemes that can primarily be found in the basic vocabulary. As a result, our combination of the two concept lists may not be perfect, but it provides an initial idea, how detailed studies on pattern borrowing that take lexical motivation patterns into account, can be carried out. We have tried to clarify this in

the introduction to the Materials section of the study, where we added more information on the advantages of the combined collection of basic vocabulary along with ethnobotanical terms.

- "One (admittedly unlikely) scenario is that there might be independent evidence that in fact BEAK is the original meaning and that it is more truthful to connect both the words for lip and beak to the concept BEAK. Would that count as a full mismatch with the languages that connect both words to the concept of LIP? Another scenario would be that the word for lip is in fact connected to the concept MOUTH, e,g, the word for lip might semantically be something like OUTER MOUTH, this in turn may be extended to the word for beak, but in this case both words are connected to a third concept MOUTH. Is that also a full mismatch with a language that extends LIP to the word for beak?"

This is a very good remark which emphasizes the importance of taking historical pathways of semantic change into account when trying to match patterns of lexical motivation. We agree that one could definitely argue that the underlying patterns have different origins, while our current annotation practice points to a very specific direction of change. In order to avoid this, however, we can also employ an annotation of the partial colexification patterns that does not make any decisions regarding the direction of semantic change and lexical motivation processes. In such an annotation, we would leave it open, which form (BEAK or LIP) we take as the primary one, and we would indicate this by using a gloss BEAK/LIP in both cases. While this is a very simple solution to account for the problems raised here, it is clear that it may not be satisfying. However, we assume that the reviewer will agree with us that it is in any case difficult to judge which direction of change would be more probable. We have added a statement in our example that emphasizes that there are different solutions than the ones we propose and which also points to the “neutral” solution of morpheme glossing as an alternative.

- "Related to the previous point, do I understand correctly that the obligatory presence of a possessive prefix with words for lip and beak in the example count as much for a match on structural borrowing as the fact that both words are connected to the concept LIP? It seems to me that one match is more indicative about past contact than the other (in general: the more specific and unusual, the more informative). Or do you have ways to differentiate between different types of structural matches?"

This is a very good point which we have not really thought through so far. It is clear that in theory, one should be able to rank the evidence. In this way, one could distinguish more surprising types of structural matches from less surprising ones and use this to indicate which one we consider as more likely than the others. However, at this stage in our analysis, where we are still trying to figure out the most transparent ways to analyze the data, we are not able to provide[4] We have, however, added a short paragraph in the final outlook of our study, where we indicate that it would be desirable to a) rank the evidence by strength, and to b) come up with some kind of a hierarchy on pattern borrowing that could guide the ranking process. any solutions for the ranking or for a systematic comparison of different types of commonalities and their respective force to provide strict evidence for pattern borrowing. [4] We have, however, added a short paragraph in the final outlook of our study, where we indicate that it would be desirable to a) rank the evidence by strength, and to b) come up with some kind of a hierarchy on pattern borrowing that could guide the ranking process.

- "You mention a number of control languages. These are all spoken in the immediate

vicinity of the Gran Chaco, and I don't think it can be excluded that there were contacts between speakers of the control languages and the target languages. So I wonder to what extent are these control languages. Are they meant to show a diminished number of commonalities, or are they meant to give a baseline of accidental commonalities? If the latter, I think the control language are not the best choices. In any case, it is good to make this clear."

We agree that the control languages are not a good choice to serve as a baseline for chance commonalities. Instead, the hope was to show that the closeness of the languages in the Chaco area leads to more commonalities between Chaco languages than with languages which are still spoken in South America but not in direct contact (thus corresponding to the first scenario mentioned). Adding control languages that might serve to illustrate accidental commonalities is an idea that we should discuss in the future. [5] For now, we have tried to clarify that the control languages in the current study were included as examples of South American languages that are not spoken in the Chaco, in order to see to what degree the possibility of a Sprachbund in the Chaco area might have eased the large amount of pattern borrowings that can be found there.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 September 2022

https://doi.org/10.21956/openreseurope.16126.r29839

Joshua Birchall
[1] Department of Linguistics, University of New Mexico, Albuquerque, NM, USA
[2] Department of Linguistics, University of New Mexico, Albuquerque, NM, USA

This data note describes a preliminary dataset of basic vocabulary and ethnobiological terms compiled for 23 languages from the Gran Chaco region of South America and neighboring areas. Few large-scale cross-linguistic lexical datasets are currently available on South American indigenous languages, and many primary sources can be difficult to access, making this work especially relevant and important. This rationale is clearly conveyed in the article.

In general, sufficient detail is provided in the text and in the references cited to replicate the workflow. However, one issue that is worth mentioning is that some forms are attributed to multiple sources, e.g. the entries for Ayoreo and Chorote. When combined with the absence of page numbers for where particular forms are located in their respective sources, and an absence of the original orthographic transcription of the form in its source, identifying the provenance of certain forms is somewhat more difficult than need be. The presence of the original transcription of the forms in the source material would further make the phonological retranscription procedure carried out by the authors more transparent and replicable. It may be advantageous to

consider each record in the dataset as a particular instance of a documented form attributed to a particular semantic concept in a particular source, a 'docunym'. While this may produce multiple forms attributed to the same concept for a single language, adding further complexity to computational work, this would address some of the issues discussed by the authors regarding having "to decide which word corresponds best to a given concept" across different sources and different language varieties.

The dataset as presented is useable and accessible to the target user, either as a reference for lexical information on the languages of the Gran Chaco or as a starting point for comparative analyses. Furthermore, the protocols adopted for the creation of this dataset make use of a suite of workflows, tools and reference catalogs that are not only appropriate, but help to define a standard for comparative lexical work within the field.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My areas of research are the documentation and description of South American indigenous languages, language typology, and historical linguistics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 20 Sep 2022

Johann-Mattis List

Thanks a lot for this very thoughtful review. We will respond to the critical points in a detailed reply, once we have received additional reviews for this study. We hope that we will be able to address problems mentioned by the reviewer in a revised version or at least to make the problematic points mentioned more transparent in our data description and avoid them in future work or design plans to make up for them in the future.

Competing Interests: No competing interests were disclosed.

Author Response 02 Dec 2022

Johann-Mattis List

We thank the reviewer for this very interesting and encouraging review. We deeply appreciate the time it took to check both paper and data. We agree with the reviewer that the mixing of multiple sources for the same language varieties constitutes a problem of the current database. While we cannot change this problem at the moment, we will try to avoid it in the future. We have added a short paragraph at the end of our study, where we emphasize this problem more transparently and recommend colleagues who plan similar data collections in the future, to make sure to avoid these problems.

We also agree with the reviewer that we missed a chance by not listing the original transcription and our modification in all cases. Again, we have added a short paragraph to the discussion of our study to make sure readers can take this as a recommendation to try and avoid these problems from the beginning when carrying out similar projects of data collection.

Competing Interests: No competing interests were disclosed.