Lexibank: A public repository of standardized wordlists with computed phonological and lexical features

Johann-Mattis List (mattis_list@eva.mpg.de)
Max Planck Institute for Evolutionary Anthropology https://orcid.org/0000-0003-2133-8919
Robert Forkel (robert_forkel@eva.mpg.de)
Max Planck Institute for Evolutionary Anthropology https://orcid.org/0000-0003-1081-086X
Simon J. Greenhill
Max Planck Institute for Evolutionary Anthropology https://orcid.org/0000-0001-7832-6156
Christoph Rzymski
Max Planck Institute for Evolutionary Anthropology
Johannes Englisch
Max Planck Institute for Evolutionary Anthropology
Russell D. Gray
Max Planck Institute for Evolutionary Anthropology https://orcid.org/0000-0002-9858-0191

Data Note

Keywords: historical linguistics, computational linguistics, cross-linguistic resource, lexical data, cross-linguistic data formats

Posted Date: September 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-870835/v1

Abstract

The past decades have seen substantial growth in digital data on the world's languages. At the same time, the demand for cross-linguistic datasets has been increasing, as witnessed by numerous studies devoted to diverse questions on human prehistory, cultural evolution, and human cognition. Unfortunately, the majority of published datasets lack standardization which makes their comparison difficult. Here, we present the first step to increase the comparability of cross-linguistic lexical data. We have designed workflows for the computer-assisted lifting of datasets to Cross-Linguistic Data Formats, a collection of standards that increase the FAIRness of linguistic data. We test the Lexibank workflow on a collection of 100 lexical datasets from which we derive an aggregated database of wordlists in unified phonetic transcriptions covering more than 2000 language varieties. We illustrate the benefits of our approach by showing how phonological and lexical features can be automatically inferred, complementing and expanding existing cross-linguistic datasets.

Background & Summary

Comparing the languages of the world opens new windows on human prehistory, culture, and cognition. By comparing languages historically, we can trace their evolution back in time and compare it with findings from archaeology and genetics (Gray et al., 2009; Sagart et al., 2019). By comparing languages typologically, we can learn about universal tendencies and cultural variation underlying the distribution of linguistic traits (Blasi et al. 2016; Jackson et al. 2019) and investigate to which degree linguistic trends are shaped by external factors (Everett et al. 2015; Blasi et al. 2019). By comparing linguistic inferences across many languages with findings in cognitive science and psychology, we can foster a broader understanding of human cognition and behaviour (Majid et al. 2018; Thompson et al. 2020; Croijmans et al. 2021).

In order to compare the languages in the world, linguistic data must be assembled in a way that maximizes the comparability of individual data points across resources and language families. Although the amount of digitally available data for the world’s languages has been drastically increasing in the past decades (Dediu 2016), the amount of comparable data is still relatively low. This problem is exacerbated because more extensive collections of data compiled in the past have often not been long-term-archived (Donohue et al. 2013) or are published along with licenses that restrict their scientific reuse (Heggarty et al. 2019).

Inspired by the GenBank database, where scholars can deposit nucleotide sequences publicly (Benson et al. 2013), we have created Lexibank, a collection of cross-linguistic datasets in standardized formats (Forkel et al. 2018a), which offers access to word forms, sound inventories, and lexical features for more than 2000 language varieties derived from 100 individual high-quality datasets.

The Lexibank wordlist collection is a first attempt towards the integration of the wealth of language data which has been assembled during the past centuries. Although far away from being complete, we are convinced that the collection will provide a rich source for future investigations into the history, the diversity, and the psychology of the world’s languages.

There are numerous ways in which the Lexibank collection can be analyzed and utilized. For the purpose of historical language comparison, the resource offers the so far largest assembly of expert judgments on historically related (cognate) words. Given that computational methods for the detection of cognates are still not able to compete with experts (List et al. 2017), our collection thus offers rich material to test and train new methods in the future. In a similar way — given that the collection unifies data on a global basis — scholars can use the data collection to test new methods for the automated identification of borrowings (Zhang et al. 2019; List and Forkel 2021b), or to expand upon previous approaches to the automated detection of contact areas (Gast and Koptjevskaja-Tamm 2018; Matsumae et al. 2021; Ranacher et al. 2021). In addition, we illustrate how the data can be used to automatically extract various phonological and lexical features for individual language varieties.

By providing a detailed, replicable workflow through which lexical datasets in various formats are unified and lifted to common standards, the Lexibank collection also contributes to increasing the "FAIRness" of cross-linguistic datasets, by making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. et al. 2016), fulfilling the initial goal of the Cross-Linguistic Data Formats initiative (https://cldf.clld.org, CLDF, Forkel et al. 2018).

Given the success of open standardized data in evolutionary biology and genetics (Yeston 2021), there is hope that increased future collaborative efforts in data standardization and curation could instigate a similar boom of new methods and insights in the language sciences. Our plan for the future is not only to expand this data collection further by contributing new datasets ourselves, but to encourage colleagues all over the world who collect cross-linguistic data to contribute to this ongoing endeavor and to share their data in an open, standardized form.

Methods

Background on Cross-Linguistic Lexical Datasets

One key type of data used in cross-linguistic studies is structural datasets, such as the World Atlas of Language Structures (Dryer and Haspelmath 2013, https://wold.info). Structural datasets assemble linguistic data in the form of features that answer concrete questions on the phonology (*Does the language have labiodental sounds?*, cf. Blasi et al. 2019), the syntax (*What is the language's basic word order?*, cf. Dunn et al. 2011; Jäger and Wahle 2021) or the lexicon of a language (*Does the language use the same word to express "fear" and"surprise?"*, cf. Jackson et al. 2019). The advantage of structural datasets is that individual features can be compared directly across languages and that the answers — which tend to be in numerical or categorical form — are usually straightforward to interpret. The disadvantage of structural datasets is that they are difficult to assemble — since linguists usually have to assemble them from dictionaries and reference grammars — and that their extraction is error-prone since it depends directly on human interpretation and analysis (Hammarström 2021).

Alternative forms of data applicable for cross-linguistic studies are *multilingual wordlists* and *parallel texts*. Wordlists offer translations for collections of concepts (typically reflecting vocabulary of everyday use) into various target languages. Parallel text collections provide translations of the same base texts into several languages. Both parallel texts and wordlists have been collected for a long time, reaching back at least to the late 18th century (Leibniz 1768; Adelung 1815). However, since automated text and sequence comparison methods require digital data, it has not been until recently that scholars started to employ them for large-scale cross-linguistic studies (Holman et al., 2011; Bentz et al., 2015; Östling 2016).

Different attempts to assemble cross-linguistic wordlists have been made in the past. The ASJP database (https://asjp.clld.org) is the most extensive wordlist collection in terms of cross-linguistic coverage, offering wordlists of about 40 items for more than 5000 language varieties in a unified phonetic transcription system (Wichmann et al. 2013). The drawback of the database is that the coverage of concepts is meagre, and even the goal of providing translations for a small list of 40 concepts is not met in many cases. Additionally, the transcription system merges many distinctions provided in the traditional International Phonetic Alphabet and therefore only offers limited possibilities for cross-linguistic studies on phonological variation.

The *Intercontinental Dictionary Series* (https://ids.clld.org, IDS) is lower in cross-linguistic coverage but has a much more extensive concept list. It contains more than 1400 concepts into more than 300 language varieties (Key and Comrie 2016a). The major problem of the IDS is not its lack of cross-linguistic coverage but the fact that linguistic forms are not provided in unified phonetic transcriptions. As a result, the data can only be used for language-internal comparisons, such as the cross-linguistic investigation of colexification patterns. Using the same term for both 'hand' and 'arm' is a colexification pattern that is found in many of the world's languages (List et al. 2013). — e.g. Russian рука (*ruka*) typically refers to both 'hand' and 'arm'.

In addition to global wordlist collections, there are also extensive wordlist collections targeting specific linguistic macro-areas. The *NorthEuralex* database (http://northeuralex.org) offers standardized wordlists for more than 1000 concepts translated into more than 100 Eurasian languages (Dellert et al. 2020). The *Hunter-Gatherer* database (https://huntergatherer.la.utexas.edu/) contains wordlists of varying size and structural features for more than 400 language varieties (Bowern et al. 2021). Table 1 provides an overview of major lexical databases which have been published in the past.

So far, the basic strategy of large-scale wordlists collections has been to assemble data language by language. Following language or area specific documentation standards on concepts and orthographies,

scholars seek to assemble as many wordlists for as many languages as possible, eventually reaching a point where it becomes more and more challenging to add more data or where a region has been sufficiently covered. Since collections will inevitably exploit existing datasets, the process of data collection involves a considerable amount of reformatting, adjusting, and modifying of independently published datasets. This process bears the danger of introducing errors into the derived data, especially when a source is interpreted and converted to adjust to the new resources. Another problem arises from the lack of flexibility of closed data collections with a fixed number of concepts and a fixed phonetic transcription system. Since decisions to ignore or recode parts of the original data during data collection cannot be easily reverted, data collections often ignore more significant parts of the original information from which they are drawn.

Dataset	Source	Target Area	Concepts	Languages	Transcriptions
ABVD	Greenhill et al. 2009	Austronesian languages	210	> 1000	—
ASJP	Wichmann et al. 2013	Global	40	> 5000	custom
Chirila	Bowern 2016	Australia	~ 300	> 200	—
DIACL	Carling et al. 2018	Global	> 400	> 300	—
GLD	Starostin and Krylof 2011	Global	110	> 300	custom
HunterGatherer	Bowern et al. 2021	Australia and South America	> 700	> 400	—
IDS	Key and Comrie 2016	Global	1310	> 300	—
NorthEuralex	Dellert et al. 2020	North Eurasia	1005	> 100	IPA
Reflex	Ségerer and Flavier 2015	African languages	from < 100 to > 1000	> 300 (?)	—
STEDT	Matisoff 2015	Sino-Tibetan languages	from < 100 to > 1000	> 400 (?)	—
TransNewGuinea.org	Greenhill 2015	New Guinea languages	from 40 to > 700	> 1000	—

Table 1: Comparing lexical wordlist collections which have been published in the past decades.

An alternative to assembling data language by language consists of *lifting* individual datasets to common standards from which custom data collections can be aggregated later. For this strategy, the availability of *reference catalogues* (which describe basic linguistic constructs, such as language varieties, concepts, speech sounds) and standard formats for data exchange (table structures, metadata) is crucial. Initial ideas to address the problems resulting from the lack of standards and exchange formats for cross-linguistic data were presented in the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al., 2018, https://cldf.clld.org). This initiative outlined a proposal to standardise cross-linguistic lexical and

structural data accompanied by software packages that help validate if data conforms to the newly proposed standards. Building on CLDF, we have developed improved ways to convert cross-linguistic lexical data into the new standards. We have tested these workflows by lifting various datasets published during the past decades and entertaining collaborations with active data collectors. In sum, this collection, which we call Lexibank, consists of 100 individual CLDF datasets covering more than 4000 wordlists from more than 2400 language varieties. In order to illustrate the interoperation and reuse potential of this data collection, we complement the collection with a new suite of software tools that allow us to extract various phonological and lexical features from the data automatically.

Cross-Linguistic Data Formats

The Cross-Linguistic Data Formats initiative was originally launched in 2014 by a group of researchers from different institutions in order to propagate the unification of cross-linguistic datasets by proposing relatively straightforward tabular formats for the representation of lexical, structural, and parallel text data (Forkel et al. 2018). From 2018 on, we further refined the original specifications by expanding the specification to account more properly for phonetic transcriptions in the form of sound sequences which may also contain information on the further sub-segmentation of words into morphemes, by integrating the CLTS reference catalog for phonetic transcriptions into the CLDF specification, and by drastically enhancing all three major reference catalogs (Glottolog, Concepticon, CLTS) in order to allow for a more detailed integration of cross-linguistic data. Details of this process can be found on the project website of the CLDF initiative (https://cldf.clld.org). For future refinements of the CLDF specification, we have adopted the practice to present them in dedicated studies first along with examples and then discuss whether to integrate them in subsequent new releases of the CLDF specification (see for example, Schweikhard and List 2020 on the representation of related word forms).

(Retro-)Standardization of Lexical Datasets

Standardization of lexical datasets with the help of CLDF comes in two forms. On the one hand, CLDF can be used to increase the comparability of existing datasets in the form of retro-standardization. On the other hand, CLDF can be used during the process of data collection and curation, where the formats can also serve as a very basic consistency check of linguistic data. In order to enhance both forms of standardization, we created the PyLexibank Python package (Forkel et al. 2021) on top of the generic CLDFBench package (Forkel and List 2020). CLDFBench allows users to convert their data with a few lines of code to CLDF formats, but lacks specific solutions that are important for the creation of lexical data. PyLexibank builds on CLDFBench to allow for a facilitated and more targeted curation of lexical data by providing integrated support of the Concepticon and the CLTS reference catalogs. The major service [illegible] explicit integration of the reference catalogs which are important

to make lexical data comparable, namely Concepticon, for the standardization of concept identifiers, derived from elicitation glosses in lexical wordlists (List et al. 2016), and CLTS for the standardization of phonetic transcriptions.

Procedure	Reference Catalog	Software	Description
link languages	Glottolog	pyglottolog	Link the language names to the identifiers provided by the Glottolog reference catalog. Currently, this is done manually in most parts.
map concepts	Concepticon	pyconcepticon	Map elicitation glosses in the original wordlist data to the concept identifiers provided by the Concepticon reference catalog. Software for semi-automated concept mapping is used for this task and then manually refined.
unify transcriptions	CLTS	pylexibank lingpy segments pyclts	Unify transcription systems by converting the transcriptions to the standards provided by the CLTS reference catalog. This procedure is by far the most complex one, which involves the cleaning of lexical forms, using dedicated routines in the *py* ≤ *ξbank* package, the creation of a draft profile with the help of the LingPy package, the manual refinement of the profile and its application with the help of the *segments* package, and finally its verification with the help of the *pyc* < *s* package.

Table 2: Basic operations involving the lifting of data to the CLDF standards with the help of the PyLexibank package.

While the linking of lexical data to the Concepticon project is organized in a dedicated workflow maintained by the editorial team of the Concepticon project, which has been described in detail in previous studies (List et al. 2018; Rzymski et al. 2020; Tjuka et al. 2021), the conversion of phonetic transcriptions to the standards provided by the CLTS project are organized with the help of *orthography profiles* (Moran and Cysouw 2018). Orthography profiles are straightforward lookup-tables which define individual graphemes in a given orthography (a grapheme being a unit consisting of one or more characters) along with their target value in the standardized transcription system. The creation and curation of orthography profiles is facilitated by the PyLexibank package by allowing users to create a draft profile from their raw data, in which a method for the automatic segmentation of phonetic transcriptions originally designed for the LingPy software package (List and Forkel 2021c) is used to provide a first “draft profile” which users can then refine systematically. The PyLexibank package offers additional routines to pre-process lexical forms with general cleaning routines (stripping off brackets, splitting entries, etc.). Having refined the profile, the data can be segmented with the Segments package (Forkel et al. 2019) and verified with the PyCLTS package (List et al. 2020). Details of the process of orthography profile creation have been discussed in previous studies (Geisler et al. forthcoming; Wu et al. 2020), Table 2 summarizes the basic operations.

Automatic Feature Extraction

Although language features are often defined in different ways, basic feature types can easily be identified and often computed in a common fashion. As an example, consider the feature “Consonant Size” which comprises the number of consonants in a given language. Once data are provided in a wordlist in phonetic transcription and segmented in such a way that unique sounds can be identified, a lower bound for the number of consonants in a given language can be approximated by counting the distinct sounds in the wordlist sample. Although this approach may fail to elicit all consonants, since there is no guarantee that a smaller collection of words will contain all sounds in a language (Dockum and Bowern 2018), it approximates the real number of sounds fairly well. Since all data in the LexiCore subset of our Lexibank collection are linked to the sound identifiers provided by the CLTS project (List et al. 2021a), which in turn define each sound by a bundle of distinctive features, we can easily extract additional subsets of sounds depending on their distinctive features. In this way, our code for feature extraction, which is implemented as part of a dedicated software package (CL Toolkit, https://github.com/cldf/cltoolkit, List and Forkel 2021a), defines various features by means of straightforward software operations which check if subsets of sounds in the sound inventory of a given language have a certain feature or a certain combination of features.

Some phonological features, like the features on prosody or sound symbolism, require additional data or functions. Prosodic features computed by CL Toolkit, for example, make use of a automatic syllabification procedure based on the sonority of individual sounds (List 2014) as implemented by the LingPy software package (List and Forkel 2021c). Features on sound symbolism, which are determined by checking of a word expressing a certain concept has certain phonetic properties, additionally need to take information from the Concepticon reference catalog into account (List et al. 2021b), which standardizes concepts in the Lexibank collection.

The extraction of lexical features checks for the full or partial identity of the word forms expressing dedicated concepts. Thus, in order to check whether “arm” and “hand” are colexified in a given language, the method first looks up the Concepticon Concept Sets “ARM” 1637, “HAND” 1277, and “ARM OR HAND” 2121 and then checks whether word forms for “ARM” and “HAND” are present and if so, if they are identical. If they are identical, it identifies a colexification, if not, it checks if a word form for “ARM OR HAND” is present, which would entail the colexification, identifying a colexification if this is the case or otherwise yielding a negative result. In a similar way, the method checks for the existence of common substrings or for affix colexifications.

The code for the automatic extraction of phonological and lexical features is written in such a way that it can be easily expanded by users in the future. Since the entities from which the features are extracted are standardized descriptors for sounds or concepts, extensions of our current code base can be easily written and integrated, or applied by creating light-weight plugins to our current solutions provided in the CL Toolkit package.

Data Records

Lexibank Wordlist Collection

By now, Lexibank assembles 4069 wordlists covering 2456 language varieties. Wordlists in the Lexibank collection show different degrees of standardization representing the level to which they can be lifted (see Table 3 for details). For 3320 wordlists taken from 94 datasets, fully standardized phonetic transcriptions are provided for at least 80 word forms. We call this dataset the LexiCore subset of Lexibank (see Chin 2015; dataset *ChinGelong* for an example). For 1806 wordlists from 52 datasets, large wordlists of at least 250 standardized concepts can be provided, but individual wordlists do not necessarily all offer fully standardized phonetic transcriptions. We call this dataset the ClicsCore subset of Lexibank (see Carling et al. 2018; dataset *DIACL* for an example). 1441 wordlists from 49 datasets are available in standardized phonetic transcriptions and offer information on etymologically related words (*cognate sets*) provided by experts. We call this dataset the CogCore subset of Lexibank (see Liú et al. 2007; dataset *LiuSinitic* for an example). A small subset of 18 wordlists from 4 datasets even offers proto-forms — forms inferred for unattested ancestral languages, using the traditional techiques of the comparative method (Weiss2015) — in standardized phonetic transcriptions. This dataset is called the ProtoCoree subset of Lexibank (see Davletshin (2012); dataset *DavletshinAztecan* for an example).

Figure 1 shows the distribution of the data for the LexiCore (wordlists with standardized transcriptions) and the ClicsCore (wordlists with extensive coverage in terms of concepts) wordlists in our collection. While we can clearly see that some regions of the world are less well covered than others, we can also see that the current collection has already reached a considerable worldwide coverage. Table 3 provides general statistics on the datasets assembled as part of the Lexibank collection. In total, the collection assembles lexical data from 100 different datasets, which together offer wordlists for 4069 language varieties, corresponding to 2456 distinct languages and dialects (as identified by Glottolog, Hammarström et al. 2021), and providing information for a total of 3110 lexical concepts, with a total of 1,912,952 words.

ID	Name	Description	Datasets	Varieties	Glottocodes	Concepts	Forms
Lexibank	all wordlists in the Lexibank collection	Metacollection of wordlists belonging to either of the datasets.	100	4069	2456	3110	1,912,952
LexiCore	wordlists with phonetic transcriptions)	Wordlists with phonetic transcriptions in which sound segments can be readily described by the CLTS system.	94	3320	2208	3050	1,041,766
ClicsCore	large wordlists with at least 250 concepts	Wordlists with large form inventories in which at least 250 concepts can be linked to the Concepticon.	52	1806	1098	3043	1,496,855
CogCore	wordlists with phonetic transcriptions and cognate sets	Wordlists with phonetic transcriptions in which cognate sets have been annotated (a subset of LexiCore).	49	1441	1114	1670	275,249
ProtoCore	wordlists with phonetic transcriptions, cognate sets, and proto-languages	Wordlists with phonetic transcriptions in which cognate sets have been annotated and which contain one or more ancestral languages whose forms are proto-forms from which forms in the descendant languages can be derived (a subset of CogCore).	4	18	18	951	8,750

Table 3: Datasets in the sample for the various subsets of the Lexibank wordlist collection.

Integration with Cross-Linguistic Resources

The Lexibank data collection provides data in formats that facilitate the *aggregation* of lexical data from different sources and the *integration* of aggregated data with other kinds of linguistic and non-linguistic information. Integration is guaranteed via the standards enforced by the CLDF specification and by reference catalogues, which provide extensive collections of metadata for standard constructs in linguistic research, such as languages (Glottolog, Hammarström et al. 2021), concepts (Concepticon, List et al. 2021b), and speech sounds (Cross-Linguistic Transcription Systems, CLTS, List et al. 2021a).

Since all reference catalogues provide additional information on the linguistic constructs they define, linking data to reference catalogues substantially enriches existing datasets. Furthermore, since the object identifiers (for languages, concepts, speech sounds) provided by the reference catalogues can be integrated into any additional resource, there are numerous ways to integrate the data. For example, via the

language identifiers by Glottolog, cultural data from the D-PLACE (Kirby et al. 2016) database can be compared with lexical data in our Lexibank collection. Via the concept identifiers by Concepticon, various kinds of speech norms, ratings, and conceptual relations can be retrieved via the NoRaRe database (Tjuka et al., 2021). Via the sound identifiers by CLTS, information on sound inventories from numerous sound inventory databases can be retrieved and compared Maddieson et al. (2013). Figure 2 illustrates how data provided in CLDF formats can be integrated by expanding the primary data with the help of reference catalogues and by analyzing and visualizing the data with the help of dedicated software tools.

Technical Validation

Due to the high level of integration and standardization of wordlists, the Lexibank collection has a high reuse potential. The data can be used as the starting point for various phylogenetic studies of individual language families. Given the large number of datasets in which etymological word relations across languages have been annotated by experts, the data can also serve as a benchmark to advance the development of new methods for automatic word comparison (List et al. 2017), which drastically exceeds the size of previously published benchmark datasets (List and Prokić 2014). In addition, the data can be used to *compute* various kinds of phonological and lexical features for individual language varieties and thus actively contribute to future studies on linguistic diversity, human prehistory, and human cognition. In the following, we will concentrate on this last aspect and show how phonological and lexical features can be automatically computed from the Lexibank collection. In this way, we contribute to recent attempts to increase the transparency of cross-linguistic collections of structural data, and expect that the role which the formal extraction of discrete and continuous features from language data plays at the moment will gain much more importance in the future.

Inference of Phonological Features

In comparative linguistics, various kinds of phonological features have been used in the past in order to compare languages. Phonological features comprise various characteristics related to the sounds of spoken languages or their combination, ranging from discrete features such as the phoneme size, reflecting the number of distinct sounds in a given language (Atkinson 2011; Moran et al. 2020), via continuous features, such as the ratio of consonant and vowel size (Maddieson 2013), and categorical features, such as the presence and type of lexical tone in a language (Everett et al. 2015), up to binary features, such as the presence of labiodental sounds (Blasi et al. 2019; Everett and Chen 2021). They are typically collected by extracting the relevant information directly from the linguistic literature (reference grammars, phonological descriptions, grammar sketches).

Since the LexiCore collection of the Lexibank wordlist collection contains word forms in standardized

ological features can be automatically computed from the

data. This has three major advantages. First, it saves a lot of time and labor, since the feature extraction can be done automatically. Second, it increases the flexibility of feature annotation, since we are not bound to decide on one representation (categorical, continuous, etc.) of feature values, before starting to collect the data, but can experiment with different representations when designing methods for feature inference. Third, it is much more transparent, since inferred features can be directly validated by referring back to the original data.

Our workflow for the extraction of phonological features from the wordlist in our LexiCore collection of Lexibank currently allows us to compute 30 distinct phonological features. Some of the features are also offered by large structural datasets (Dryer and Haspelmath 2013) and can be directly compared with them, while other features have not been assembled in publicly available datasets so far and may therefore offer interesting insights to language typologists.

Table 4 shows the 30 phonological features which we automatically extracted from the data. As can be seen from the table, the features can be classified into four distinct groups. There are discrete features on sound inventory sizes (1-7, number of vowels, consonants, etc.), there are various features on special sound types or individual specific sounds (8-19), there are three prosodic features (20-22), and eight features pertaining to specific sound-meaning relations (aka sound symbolism, 23-30).

No.	Identifier	Name	Type
1	ConsonantQualitySize	consonant quality size	inventory size
2	VowelQualitySize	vowel quality size	
3	VowelSize	vowel size	
4	ConsonantSize	consonant size	
5	CVRatio	consonant and vowel ratio	
6	CVQualityRatio	consonant and vowel ratio (by quality)	
7	CVSoundRatio	consonant and vowel ratio (including diphthongs and clusters)	
8	HasNasalVowels	has nasal vowels or not	special sounds
9	HasRoundedVowels	has rounded vowels or not	
10	VelarNasal	has the velar nasal (engma)	
11	PlosiveVoicingGaps	voicing and gaps in plosives	
12	LacksCommonConsonants	gaps in plosives	
13	HasUncommonConsonants	has uncommon consonants	
14	PlosiveFricativeVoicing	voicing in plosives and fricatives	
15	UvularConsonants	presence of uvular consonants	
16	GlottalizedConsonants	presence of glottalized consonants	
17	HasLaterals	presence of lateral consonants	
18	HasLabiodentalFricatives	inventory has labio-dental fricatives or affricates	
19	HasPrenasalizedConsonants	inventory has pre-nasalized consonants	
20	SyllableStructure	complexity of the syllable structure	prosody
21	SyllableOnset	complexity of the syllable onset	
22	SyllableOffset	complexity of the syllable offset	
23	FirstPersonWithM	fist person starts with an m-sound	sound symbolism
24	FirstPersonWithN	fist person starts with an n-sound	
25	SecondPersonWithT	second person starts with a t-sound	
26	SecondPersonWithM	second person starts with an m-sound	
27	SecondPersonWithN	second person starts with an n-sound	
28	MotherWithM	mother starts with m-sound	
29	FatherWithP	father starts with p-sound	
		ith f-sound	

Table 4: Phonological features automatically extracted from the LexiCore data in Lexibank.

In order to evaluate the usefulness of our approach for automatic feature extraction from lexical datasets, we compare how well the inferred values for five selected features in LexiCore correlate with the features provided in the WALS database (Dryer and Haspelmath 2013) and the features inferred from Phoible (Moran and McCloy 2019). As can be seen from the results of this comparison in Table 5, our approach receives reasonably high correlations with both the features in WALS and those extracted from Phoible, although Phoible and WALS generally show a higher correlation with each other. This is, however, not surprising, given that both datasets are based on very similar sources by the same contributor.

Feature	WALS / LexiCore	WALS / Phoible	LexiCore / Phoible	Sample
ConsonantSize	0.66 / p<0.01	0.92 / p<0.01	0.70 / p<0.01	233
VowelQualitySize	0.51 / p<0.01	0.66 / p<0.01	0.68 / p<0.01	235
CVRatio	0.55 / p<0.01	0.76 / p<0.01	0.68 / p<0.01	235
PlosiveFricativeVoicing	0.54 / p<0.01	0.69 / p<0.01	0.59 / p<0.01	235
PlosiveVoicingGaps	0.40 / p<0.01	0.60 / p<0.01	0.56 / p<0.01	235

Table 5: Spearman rank correlation (ρ) coefficients of feature values in WALS, Phoible and LexiCore, for five selected features, calculated for those parts of the data where information in all three dataset could be obtained, matching languages by their common Glottocodes. When more than one language was available for the same Glottocode, the median value was taken.

Investigating the features inferred with our workflows requires tools for exploratory data analysis. One way to explore large feature collections for cross-linguistic data is to plot them on a geographic map in order to see whether specific areal patterns emerge. CLDF comes with a dedicated suite of software tools for data visualization which greatly facilitate this part (CLDFViz, Forkel 2021), allowing users to create high-quality static and interactive maps in which features can be combined ad libitum. An example for such a map is shown in Figure 3, where we have plotted the features 28 and 29 in our collection, which ask whether words for “mother” and “father” start with [m] and [p] respectively, reflecting a well-known trend that can be observed in the worlds languages and is usually attributed to the sounds which children learn during first-language acquisition (Jakobson 1960). As can be seen from the map, our data confirms the global trend, in so far as many unrelated languages in distant geographic areas have words for “mother” which start with [m] and words for “father” which start with [p] or similar sounds (including labiodental fricatives like [f]). More detailed investigations would require in-depth analyses by language typologists, for which our dataset provides a useful starting point.

Inference Lexical Features

Languages differ in the way in which their lexicons are structured. One of the most prominent aspects in which languages differ is to which degree they use the same word forms to denote different concepts. Russian ruka, for example, can mean “arm” and “hand,” and German Decke can mean “ceiling” and “blanket.” This phenomenon, termed colexification in the recent linguistic literature (a cover term for polysemy on the one hand and homophony on the other hand, François 2008) has recently received broader attention among linguists (Schapper 2019), psychologists (Jackson et al. 2019), and computer scientists (Bao et al. 2021), and is most prominently represented in the Database of Cross-Linguistic Colexifications (CLICS, https://clics.clld.org, Rzymski et al. 2020) which aggregates colexifications from CLDF datasets for more than 2000 language varieties. While the original CLICS database was built from 30 datasets, the ClicsCore collection in Lexibank expands this collection by 20 additional datasets. Retaining only those languages which provide at least 250 concepts which can be linked to the Concepticon reference catalog, ClicsCore contains 1784 different language varieties corresponding to 1246 different languages (as reflected by unique Glottocodes in the Glottolog reference catalog).

While the original CLICS data identifies only those cases as colexifications where an identical word form denotes two different senses, we expand the notion of colexification in our feature extraction procedure by adding two more types of colexification which have so far only been sporadically discussed in the literature. First, we add a method for the identification of partial colexifications, defined as those cases in which two word forms expressing two different concepts are not identical, but share a common substring, and affix colexifications, where one word appears as a prefix or a suffix of another word (see Table 6 for examples and full definitions). Searching systematically for these colexifications in our data allows us to identify commonalities in the languages of the world and to investigate whether they are due to areal proximity, common descent, or rather general cognitive principles.

Type	Description	Examples
full colexification	Two different senses are expressed by the same word form.	Russian *ruka* “hand” vs. *ruka* “arm”. German *Decke* “blanket” vs. *Decke* “ceiling”.
partial colexification	Two word forms expressing two different senses are expressed by word forms which share a common substring	German *be*-**antwort**-en “answer” vs. *ver*-**antwort**-*en* “be responsible”.
affix colexification	Of two word forms expressing two different senses, one word form is identical with the beginning of the end of the other word form.	German *Fingernagel* “fingernail” vs. *Nagel* “nail (tool)”. German *Ellenbogen* “elbow” vs. *Bogen* “bow (arc)”.

Table 6: Colexification patterns which can be computed from the ClicsCore subset of the Lexibank wordlist collection.

The 30 features which we compute from the ClicsCore subset of our wordlist collection are given in Table 7. While we could easily expand this collection further, we have limited the features to those cases which have been previously discussed in the literature and collected manually in structural datasets.

No.	Identifier	Name	Type
1	LegAndFoot	has the same word form for foot and leg	colexification
2	ArmAndHand	arm and hand distinguished or not	
3	BarkAndSkin	bark and skin distinguished or not	
4	FingerAndHand	finger and hand distinguished or not	
5	GreenAndBlue	green and blue colexified or not	
6	RedAndYellow	red and yellow colexified or not	
7	ToeAndFoot	toe and foot colexified or not	
8	SeeAndKnow	see and know colexified or not	
9	SeeAndUnderstand	see and understand colexified or not	
10	ElbowAndKnee	elbow and knee colexified or not	
11	FearAndSurprise	fear and surprise colexified or not	
12	CommonSubstringInElbowAndKnee	elbow and knee are partially colexified or not	partial colexification
13	CommonSubstringInManAndWoman	man and woman are partially colexified or not	
14	CommonSubstringInFearAndSurprise	fear and surprise are partially colexified or not	
15	CommonSubstringInBoyAndGirl	boy and girl are partially colexified or not	
16	EyeInTear	eye partially colexified in tear	affix colexification
17	BowInElbow	bow partially colexified in elbow	
18	CornerInElbow	corner partially colexified in elbow	
19	WaterInTear	water partially colexified in tear	
20	TreeInBark	tree partially colexified in bark	
21	SkinInBark	skin partially colexified in bark	
22	MouthInLip	mouth partially colexified in lip	
23	SkinInLip	skin partially colexified in lip	
24	HandInFinger	hand partially colexified in finger	
25	FootInToe	foot partially colexified in toe	
26	ThreeInEight	three partially colexified in eight	
27	ThreeInThirteen	three partially colexified in thirteen	
28	FingerAndToe	finger and toe colexified or not	
29	HairAndFeather	hair and feather colexified or not	
		ar and smell colexified or not	

Table 7: 30 lexical features which can be automatically extracted from the ClicsCores subset of Lexibank. Features can be divided into three major classes, depending on the type of colexification they reflect: (A) colexifications, referring to cases of polysemy in which one word form expresses two distinct senses, (B) partial colexification, referring to cases in which two word forms expressing distinct senses share a common substring, and (C) overlap colexificatio, referring to cases in which one word form starts or ends with another word form.

As a first example for the potential of large aggregated datasets, Figure 4 shows which languages in our collection colexify “arm” with “hand” and “leg” with “foot,” respectively. Previous studies have almost exclusively concentrated on the global distribution of languages colexifying “arm” and “hand,” assuming that there is a geographic tendency to colexify the terms more frequently, the closer one comes to the equator (Brown 2013). Contrasting the colexification pattern with its logical counterpart yields interesting patterns, in so far, as our analysis suggests a rather strong systemic tendency across languages from different language families and areas to either express both “arm/hand” and “foot/leg” by the one word each, or to distinguish them both. More research on this topic is needed. The data we have assembled here are a helpful starting point.

Figure 5 provides another example on features which partially occur in correlated form. This time, we compare whether languages denote “woman” and “man” by means of a partial colexification (compare □□ *nǔ-rén* “female person → woman” vs. □□ *nán-rén* “male person → man” in Mandarin Chinese) on the one hand, and “daughter” and “son” (compare□□ *nǔ-ěr* “female offspring → daughter” vs. □□ *ěrzǐ* “offspring-son → son”) on the other hand. The analysis suggests a large areal cluster in South-East Asia, where the tendency of languages to use compound words in a rather analytical manner is well known, as well as some languages in the North of South America, but the pattern shows a less global distribution than the one for “arm” vs. “leg” shown in Figure 4.

As a final example, Figure 6 compares affix colexifications in which words recur in the beginning of another word, indicating strong semantic relations. In the concrete example, we check to which degree the word for “tear” in the languages in our sample is composed of the word for “eye” and the word for “water” respectively. That “tears” are denoted as “eye-water” is a common pattern that can be found in quite a few South-East Asian languages (compare Younuo [ki□□mo³²-□ŋ⁴⁴] “eye-water,” Chén 2012; Wu et al. 2020), but also in a few languages in South America (compare Guaraní esa-□ “eye-water” Key and Comrie 2016a). As can be seen from the Figure, we find that South-East Asian languages indeed overwhelmingly express “tears” as “eye-water,” in so far as they show an affix colexification of “eye” and of “water” with “tear,” but apart from this, the feature only occurs sporadically.

Usage Notes

Distribution of Lexibank Datasets

For the distribution of CLDF datasets in general and Lexibank datasets in specific, we make use of existing long-term archiving solutions provided by Zenodo (https://zenodo.org). Once a Lexibank dataset has been created and the creators consider the data ready to be shared publicly, a new version of the data is created and archived with Zenodo, using the automated integration of Zenodo with GitHub. In addition, the new version is tagged as part of the Lexibank community on Zenodo (https://zenodo.org/communities/lexibank), which guarantees the findability of the resource.

Promotion of Lexibank

Lexibank and lexical data in CLDF formats have been promoted in several ways so far. First, we have conducted detailed studies in which CLDF formats are used along with CLDFBench and the pylexibank software package, illustrating how data aggregation can be successfully carried out (List et al. 2018; Rzymski et al. 2020), or showing how data can be supplemented in transparent CLDF formats (Wu et al. 2020; List and Forkel 2021b). Second, we have created certain flagship projects which showcase specific aspects of CLDF and the advantage of using integrated data (Geisler et al. 2020; Ferraz Gerardi et al. 2021). Third, we have conducted projects with students and young scholars, who were trained to use our new resources and encouraged to share their knowledge in the form of small blog posts (published at https://calc.hypotheses.org) along with new datasets which bachelor, doctoral, and master students lifted themselves assisted by our team (Tjuka 2020; Blum 2021; Grond and Tüfekci 2021).

Declarations

Supplementary Material

The supplementary material contains all datasets and software packages required to replicate the analyses showin in this study. The main software package is curated on GitHub (https://github.com/lexibank/lexibank-analysed/tree/v0.2) and archived with Zenodo (https://doi.org/10.5281/zenodo.5227817). Individual datasets belonging to the Lexibank wordlist collection are curated on individual repositories on GitHub (see our master list at https://github.com/lexibank/lexibank-analysed/blob/v0.2/etc/lexibank.csv) and are also all archived with Zenodo (see https://zenodo.org/communities/lexibank/).

Acknowledgments

Many people were involved in the preparation of individual datasets which have been integrated in the Lexibank collection. We are very grateful for their help in standardizing lexical datasets. These contributors are listed as authors, editors, or in specific roles along with the individual Lexibank datasets archived with Zenodo. We express particular thanks to Tiago Tresoldi, Mei-Shin Wu, Yunfan Lai, and Hans-Jörg Bibiko for providing help in preparting individual datasets using our workflows, to Quentin Atkinson in sharing ideas in initial discussions on the data collection, and to Abbie Hantgan, Alexander Savelyev, Cathryn Yang, Claire Bowern, Damian Satterthwaite-Phillips, Fabrício Ferraz Gerardi, Fēng Wáng, Frederic Blum, Gerhard Jäger, George S. Starostin, Guillaume Ségerer, Jessica K. Ivani, Johannes Dellert, Kaj Syrjänen, Magnus Pharao Hansen, Maria Koptjevskaja-Tamm, Mary Walworth, Maurizio Serva, Michael Dunn, Muhammad Zakaria, Natalia Morozova, Nathan W. Hill, Nathaniel A. Sims, Olof Lundgren, Paul Sidwell, Sean Lee, Thiago C. Chacon, Timotheus A. Bodt, and Volker Gast, for generously sharing data and providing help in the preparation of individual datasets. Special thanks also go to the different teams contributing to the maintenance and further development of our three major reference catalogs, Glottolog (Harald Hammarström, Martin Haspelmath, and Sebastian Bank), Concepticon (Nathanael Schweikhard, Annika Tjuka, Kristina Pianykh, Carolin Hundt, Mei-Shin Wu, Tiago Tresoldi), and CLTS (Cormac Anderson, Tiago Tresoldi).

Funding Information

As part of the CLLD project (cf. https://clld.org) and the Glottobank project (cf. https://glottobank.org), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History, and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121). JML was funded by the ERC Starting Grant 715618 *Computer-Assisted Language Comparison* (cf. https://digling.org/calc/). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041).

Author Contributions

RDG initiated the Lexibank project as part of the larger Glottobank initiative of the Department of Linguistic and Cultural Evolution of the Max Planck Insitute for Evolutionary Anthropology in Leipzig (formerly Max Planck Institute for the Science of Human History in Jena) and provided financial, administrative, and conceptual support for the development of Lexibank. JML, RF, and SJG consecutively worked out the core aspects of the CLDF specification for the handling of multilingual wordlists, which was later expanded

consecutively by CR, JML, RF, and SJG. RF wrote the first version of the PyLexibank software package and CR, JML, and SJG contributed to its further development. JML and RF wrote the CL Toolkit package used to compute phonological and lexical features from wordlists in this study. JML and RF wrote the first version of the analyses provided as part of the lexibank-analysed repository, and CR and JE contributed to its further development. JML and RF made the graphics for this study. CR, JE, JML, RF, and SJG created, curated, tested, and archived the datasets published as part of the Lexibank collection. JML wrote the first draft, and JML, RDG, RF, and SJG expanded the first draft. All authors revised the second draft and agree with the final version.

Conflict of Interest

There are no conflicts of interest.

References

Adelung F von (1815) Catherinens der Grossen Verdienste um die vergleichende Sprachenkunde. Friedrich Drechsler, Sankt Petersburg.

Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332:346–349.

Bao H, Hauer B, Kondrak G (2021) On universal colexifications. In: Proceedings of the 11th Global Wordnet Conference. Global Wordnet Association, University of South Africa (UNISA), pp 1–7.

Benson DA, Cavanaugh M, Clark K, et al (2013) GenBank. Nucleic Acids Res 41:36–42.

Bentz C, Verkerk A, Kiela D, et al (2015) Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. PLOS ONE 10:e0128254. https://doi.org/10.1371/journal.pone.0128254

Blasi DE, Moran S, Moisik SR, et al (2019) Human sound systems are shaped by post-Neolithic changes in bite configuration. Science 363:1–10. https://doi.org/10.1126/science.aav3218

Blasi DE, Wichmann S, Hammarström H, et al (2016) Sound–meaning association biases evidenced across thousands of languages. Proceedings of the National Academy of Science of the United States of America 113:10818–10823

Blum F (2021) Data gathering in times of a pandemic: Upcycling Constenla Umaña's data on the Chibchan, Lencan and Misumalpam language families. Computer-Assisted Language Comparison in Practice

4:2751.

Bowern C (2016) Chirila: Contemporary and historical resources for the indigenous languages of Australia [dataset]. Language Documentation and Conservation.

Bowern C, Epps P, Hill J, McConvell P (2021) Languages of hunter-gatherers and their neighbors [dataset, version from 2021-04-27].

Brown CH (2013) Hand and arm. In: Dryer MS, Haspelmath M (eds) The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info

Carling G, Larsson F, Cathcart CA, et al (2018) Diachronic Atlas of Comparative Linguistics (DiACL). A database for ancient language typology. PLOS ONE 1–20

Chén Q. □□□ (2012) Miàoyáo yǔwén. Zhōngyāng Mínzú Dàxué □□□□□□ [Central Institute of Minorities], Běijīng

Chin AC (2015) The Gelong language in the multilingual hub of Hainan. Bulletin of Chinese Linguistics 8:140–156. https://doi.org/10.1163/2405478X-00801008

Croijmans I, Arshamian A, Speed LJ, Majid A (2021) Wine experts' recognition of wine odors is not verbally mediated. Journal of Experimental Psychology: General 150:545–559. https://doi.org/10.1037/xge0000949

Davletshin A (2012) Proto-Uto-Aztecans on their way to the Proto-Aztecan homeland: Linguistic evidence. Journal of Language Relationship 1:75–92.

Dediu D (2016) Typology for the masses. Linguistic Typology 20:579–581.

Dellert J, Daneyko T, Münch A, et al (2020) NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. Language Resources & Evaluation 54:273–301. https://doi.org/10.1007/s10579-019-09480-6

Dockum R, Bowern C (2018) Swadesh lists are not long enough: Drawing phonological generalizations from limited data. In: Austin PK (ed) Language documentation and description. EL Publishing, London, pp 35–54.

Donohue M, Hetherington R, McElvenny J, Dawson V (2013) World phonotactics database. Department of Linguistics. The Australian National University, Canberra.

Dryer MS, Haspelmath M (eds) (2013) WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info

Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. Nature 473:79–82.

Everett C, Blasi DE, Roberts SG (2015) Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. Proceedings of the National Academy of Sciences of the United States of America 112:1322–1327

Everett C, Chen S (2021) Speech adapts to differences in dentition within and across populations. Scientific Reports 11:1–10. https://doi.org/10.1038/s41598-020-80190-8

Ferraz Gerardi F, Reichert S, Aragon C, et al (2021) TuLeD: Tupían Lexical Database. Version 0.11. https://tular.clld.org/

Forkel R (2021) CLDFViz. A Python library providing tools to visualize data from CLDF datasets [software library, version 0.5.0]. https://doi.org/10.5281/zenodo.5162667

Forkel R, Bank S, Rzymski C, Bibiko H-J (2020) CLLD: A toolkit for cross-linguistic databases. [Software library, version 7.2.0]. Max Planck Institute for the Science of Human History, Jena. https://clld.org

Forkel R, Greenhill SJ, Bibiko H-J, et al (2021) PyLexibank. The python curation library for lexibank [software library, version 2.8.2]. Zenodo, Geneva. https://github.com/lexibank/pylexibank

Forkel R, List J-M (2020) CLDFBench. Give your cross-linguistic data a lift. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Luxembourg, pp 6997–7004. https://github.com/cldf/cldfbench

Forkel R, List J-M, Greenhill SJ, et al (2018) Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5:1–10. https://cldf.clld.org

Forkel R, Moran S, List J-M, et al (2019) Segments. Unicode standard tokenization routines and orthography profile segmentation [software library, version 2.1.3]. Zenodo, Geneva. https://github.com/cldf/segments

François A (2008) Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In: Vanhove M (ed) From polysemy to semantic change. Benjamins, Amsterdam, pp 163–215.

Gast V, Koptjevskaja-Tamm M (2018) The areal factor in lexical typology. Some evidence from lexical databases. In: Olmen D, Mortelmans T, Brisard F (eds) Aspects of linguistic variation. de Gruyter, Berlin; New York, pp 43–81.

Geisler H, Forkel R, List J-M (forthcoming) A digital, retro-standardized edition of the tableaux phonétiques des patois suisses romands (TPPSR). In: Avanzi M, LoVecchio N, Millour A, Thibault A (eds) Nouveaux regards sur la variation dialectale. Éditions de Linguistique et de Philologie, Strasbourg, pp 1–21.

Geisler H, Forkel R, List J-M (2020) The tableaux phonétiques des patois suisses romands online. Version 1.0. https://tppsr.clld.org

Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific Settlement. Science 323:479–483.

Greenhill, SJ (2015): TransNewGuina.org. An online database of New Guinea languages. PLOS One. 10.10: e0141563. https://doi.org/10.1371/journal.pone.0141563

Greenhill SJ, Blust R, Gray RD (2008) The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. Evolutionary Bioinformatics 4:271–283.

Grond FR, Tüfekci A (2021) Computer-assisted comparison fo Gelong and Hlai using Cross-Linguistic Data Formats. Computer-Assisted Language Comparison in Practice 4:2827. https://calc.digling.org/2827

Hammarström H (2021) Measuring prefixation and suffixation in the languages of the world. In: Proceedings of the third workshop on computational typology and multilingual NLP. Association for Computational Linguistics, Online, pp 81–89.

Hammarström H, Haspelmath M, Forkel R, Bank S (2021) Glottolog. Version 4.4. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://glottolog.org

Heggarty P, Shimelman A, Abete G, et al (2019) Sound comparisons: A new online database and resource for research in phonetic diversity. In: Calhoun S, Scudero P, Tabain M, Warren P (eds) Proceedings of the 19th International Congress of Phonetic Sciences. Asutralasian Speech Science; Technology Association, Canberra, pp 280–284.

Holman EricW, Brown CH, Wichmann S, et al (2011) Automated dating of the worldś language families based on lexical similarity. Current Anthropology 52:841–875.

Jackson JC, Watts J, Henry TR, et al (2019) Emotion semantics show both cultural variation and universal structure. Science 366:1517–1522.

Jäger G, Wahle J (2021) Phylogenetic typology. Frontiers in Psychology 12. https://doi.org/10.3389/fpsyg.2021.682132

Jakobson R (1960) Why ‘Mama´ and ‘Papa?. In: Kaplan B, Wapner S (eds) Perspectives in psychological theory: Essays in honor of Heinz Werner. International Universities Press, New York, pp 124–134.

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

Kirby LR, Gray RD, Greenhill, SJ, Jordan FM, Gomes-Ng S, Bibiko HJ, Blasi DE, Botero CA, Bowern C, Ember CR, Leehr D, Low BS, McCarter J, Divale W, Gavin, MC (2016) D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. PLOS ONE 11:1–14. https://doi.org/10.1371/journal.pone.0158391

Leibniz GW von (1768) Desiderata circa linguas populorum, ad Dn. Podesta. In: Dutens L (ed) Godefridi [illegible]n collecta, in classes distributa, praefationibus et indicibus

exornata. Fratres des Tournes, Geneva, pp 228–231.

List J-M (2014) Sequence comparison in historical linguistics. Düsseldorf University Press, Düsseldorf.

List J-M, Anderson C, Tresoldi T, Forkel R (2021a) Cross-Linguistic Transcription Systems. Version 2.1.0. Max Planck Institute for the Science of Human History, Jena. https://clts.clld.org

List J-M, Anderson C, Tresoldi T, Forkel R (2020) PYCLTS. A Python library for the handling of phonetic transcription systems [Software Library, Version 3.0.0]. Zenodo, Geneva. https://github.com/cldf-clts/pyclts/

List J-M, Cysouw M, Forkel R (2016) Concepticon. A resource for the linking of concept lists. In: Chair) NC (Conference, Choukri K, Declerck T, et al. (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Luxembourg, pp 2393–2400.

List J-M, Forkel R (2021c) LingPy. A Python library for quantitative tasks in historical linguistics [software library, version 2.6.8]. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://github.com/lingpy/lingpy/

List J-M, Forkel R (2021b) Automated identification of borrowings in multilingual wordlists. Open Research Europe 1:1–11. https://doi.org/10.12688/openreseurope.13843.1

List J-M, Forkel R (2021a) CL toolkit. A Python library for the processing of cross-linguistic data [software library, version 0.1.1]. Zenodo, Geneva. https://github.com/cldf/cltoolkit

List J-M, Greenhill SJ, Anderson C, et al (2018) CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. Linguistic Typology 22:277–306.

List J-M, Greenhill SJ, Gray RD (2017) The potential of automatic word comparison for historical linguistics. PLOS ONE 12:1–18.

List JM, Rzymski C, Greenhill S, et al (2021b) Concepticon. A resource for the linking of concept lists. Version 2.5.0. Version 2.5.0. Max Planck Institute for the Science of Human History, Jena. https://concepticon.clld.org/

List J-M, Terhalle A, Urban M (2013) Using network approaches to enhance the analysis of cross-linguistic polysemies. In: Proceedings of the 10th International Conference on Computational Semantics – Short Papers. Association for Computational Linguistics, Stroudsburg, pp 347–353.

Liú Lìlǐ □□□, Wáng Hóngzhōng □□□, Bǎi Yíng □□ (2007) Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí □□□□□□□□□·□□□ [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects. Fènghuáng, Nánjīng.

Maddieson I (2013) Consonant-vowel ratio. In: Dryer MS, Haspelmath M (eds) The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info

Maddieson I, Flavier S, Marsico E, et al (2013) LAPSyD: Lyon-Albuquerque Phonological Systems Database. In: Proceedings of Interspeech.

Majid A, Roberts SG, Cilissen L, et al (2018) Differential coding of perception in the world's languages. Proceedings of the National Academy of Sciences United States of America 115:11369–11376. https://doi.org/10.1073/pnas.1720419115

Matisoff JA (2015) The Sino-Tibetan Etymological Dictionary and Thesaurus project. University of California, Berkeley

Matsumae H, Ranacher P, Savage PE, et al (2021) Exploring correlations in genetic and cultural variation across language families in northeast asia. Science Advances 7. https://doi.org/10.1126/sciadv.abd9223

Michael L, Stark T, Clem E, Will Chang and (2021) South American Phonological Inventory Database [dataset, version 1.1.5]. http://linguistics.berkeley.edu/~saphon/en/

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

Moran S, Grossman E, Verker A (2020) Investigating diachronic trends in phonological inventories using BDPROTO. Language Resources & Evaluation 0:1–25. https://doi.org/10.1007/s10579-019-09483-3

Moran S, McCloy D (eds) (2019) PHOIBLE 2.0. Max Planck Institute for the Science of Human History, Jena. https://phoible.org

Östling R (2016) Studying colexification through massively parallell corpora. In: Schapper A, Roque LS, Hendery R (eds) The lexical typology of semantic shifts. De Gruyter Mouton, Berlin; Boston, pp 157–176.

Ranacher P, Neureiter N, Gijn R van, et al (2021) Contact-tracing in cultural evolution: A bayesian mixture model to detect geographic areas of language contact. Journal of The Royal Society Interface 18:20201031. https://doi.org/10.1098/rsif.2020.1031

Rzymski C, Tresoldi T, Greenhill S, et al (2020) The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies. Scientific Data 7:1–12. https://clics.clld.org

Sagart L, Jacques G, Lai Y, et al (2019) Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Science of the United States of America 116:10317–10322.

Schapper A (2019) The ethno-linguistic relationship between smelling and kissing: A Southeast Asian case study. Oceanic Linguistics 58:92–109. https://doi.org/10.1353/ol.2019.0004

Schweikhard NE, List J-M (2020) Developing an annotation framework for word formation processes in comparative linguistics. SKASE Journal of Theoretical Linguistics 17:2–26.

Segerer G, Flavier S (2015) RefLex: Reference Lexicon of Africa. http://reflex.cnrs.fr

Starostin GS, Krylov P (eds) (2011) The Global Lexicostatistical Database: Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-form. https://starling.rinet.ru/new100/main.htm

Thompson B, Roberts SG, Lupyan G (2020) Cultural influences on word meanings revealed through large-scale semantic alignment. Nature Human Behaviour 4:1029–1038. https://doi.org/10.1038/s41562-020-0924-8

Tjuka A (2020) Adding concept lists to concepticon: A guide for beginners. Computer-Assisted Language Comparison in Practice 3. 1: 2225. https://calc.hypotheses.org/2225

Tjuka A, Forkel R, List J-M (2021) Linking norms, ratings, and relations of words and concepts across multiple language varieties. Behavior Research Methods. 1–21.

Weiss M (2015) The comparative method. In: Bowern C, Evans N (eds) The Routledge Handbook of Historical Linguistics. Routledge, New York, pp 127–145.

Wichmann S, Müller A, Wett A, et al (2013) The ASJP Database. Version 16. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://asjp.clld.org

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3.

Wu M-S, Schweikhard NE, Bodt TA, et al (2020) Computer-assisted language comparison. State of the art. Journal of Open Humanities Data 6:1–14.

Yeston JS (2021) Progress in data and code deposition. Science Editors' Blog. https://blogs.sciencemag.org/editors-blog/2021/07/15/progress-in-data-and-code-deposition/

Zhang L, Manni F, Fabri R, Nerbonne J (forthcoming) Detecting loan words computationally. In: Variation Rolls the Dice. A Worldwide Collage in Honour of Salikoko S. Mufwene. Amsterdam: Benjamins.

Figures

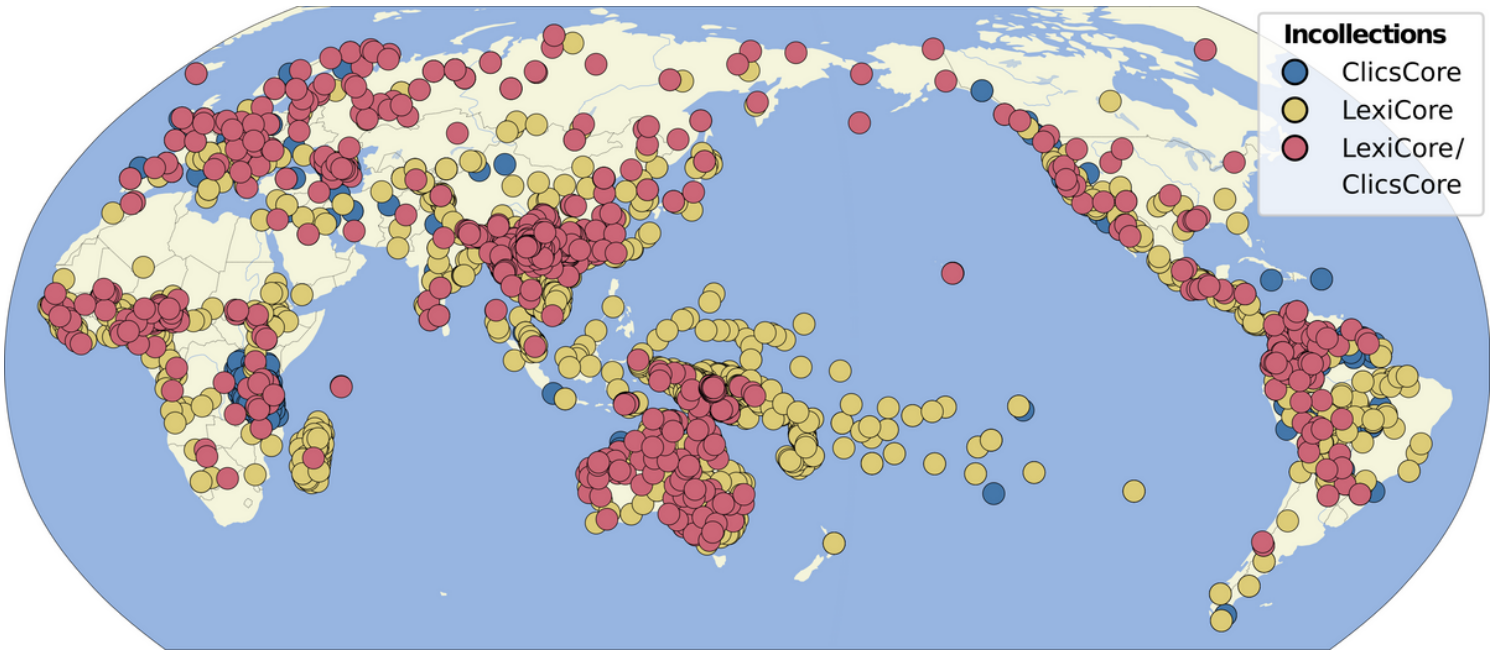

Figure 1

Distribution of lexical resources with phonetic transcriptions (LexiCore) and lexical resources with a larger number of lexical forms (ClicsCore) in the Lexibank wordlist collection.

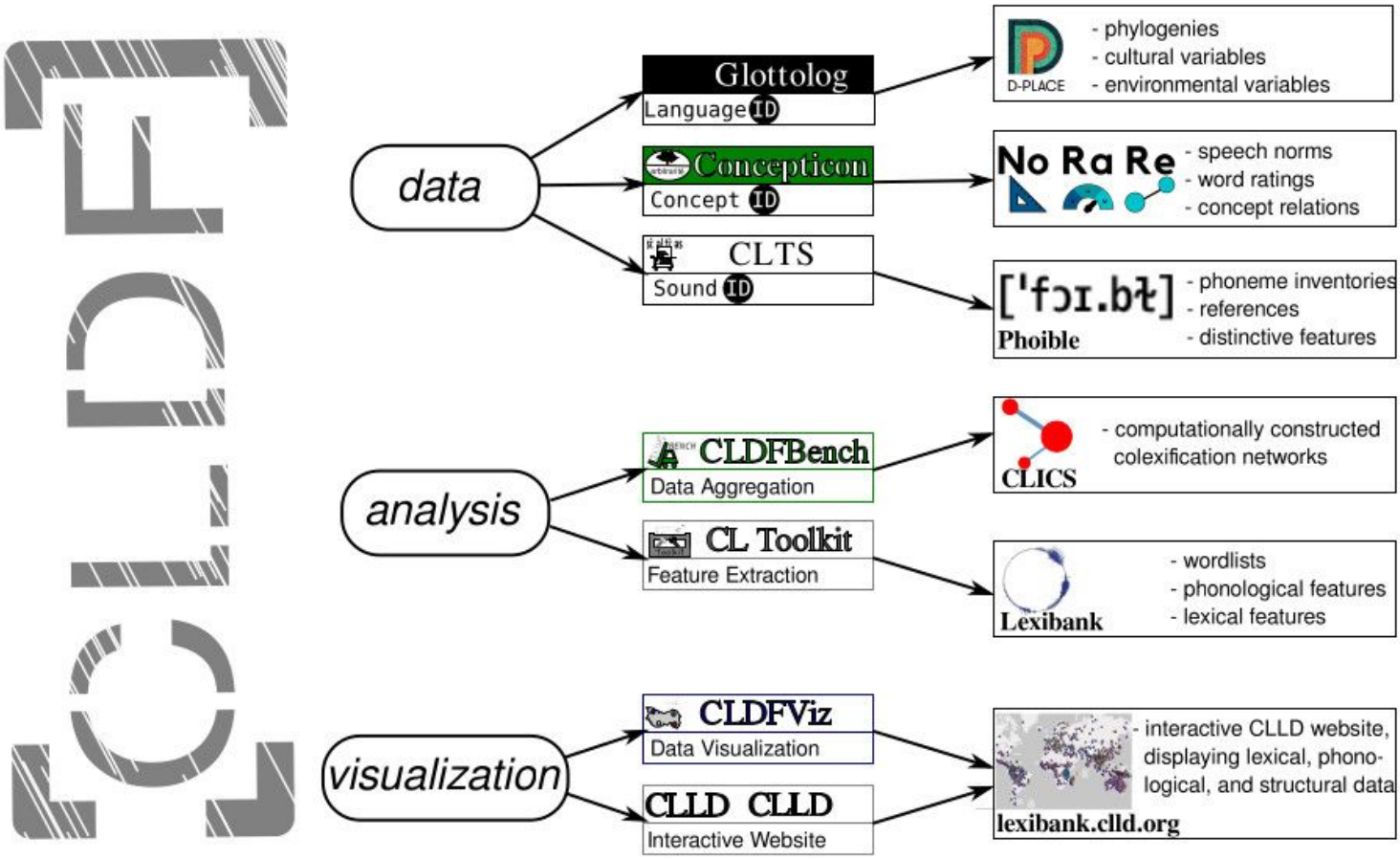

Figure 2

Reference catalogs, tools for analysis, and tools for visualization, integrated by CLDF datasets. By providing active links to the identifiers of Glottolog, Concepticon, and by converting phonetic transcriptions to the standard transcriptions provided by the CLTS catalog, CLDF datasets can be integrated with other existing datasets, such as D-Place (Kirby et al. 2016), NoRaRE (Tjuka et al. 2021), and Phoible (Moran and McCloy 2019). With the help of dedicated packages for the analysis of CLDF datasets, data can be easily aggregated with CLDFBench (Forkel and List 2020), and features can be automatically extracted with the help of CL Toolkit (List and Forkel 2021a). For the visualization of CLDF datasets, data can be plotted on geographic maps with the help of CLDFViz (Forkel 2021) and shared on interactive websites with the help of CLLD (Forkel et al. 2020).

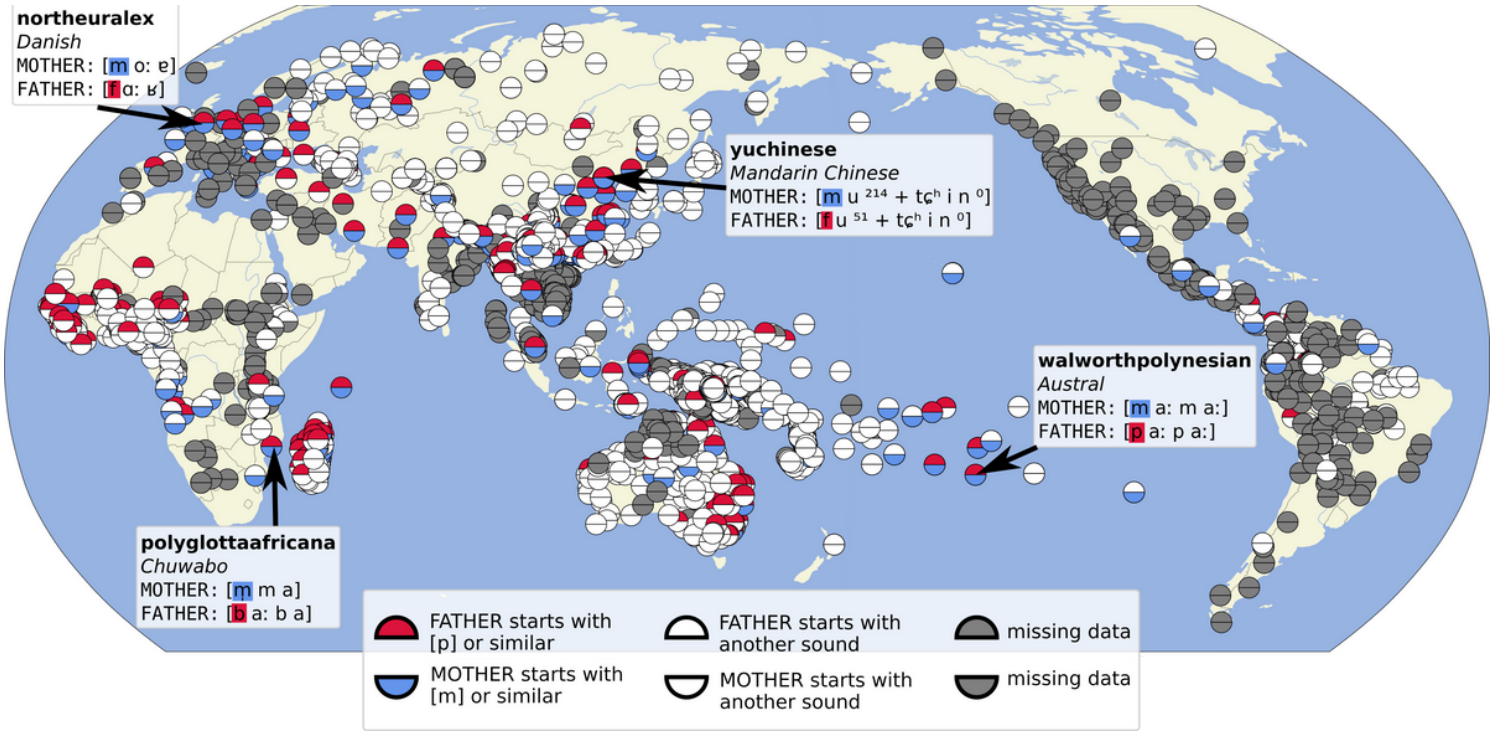

Figure 3

Comparing cross-linguistic patterns of sound symbolism involving words for “mother” and “father” in the world’s languages. The four datasets from which the four examples showing actual forms for individual language varieties are taken are indicated in the figure.

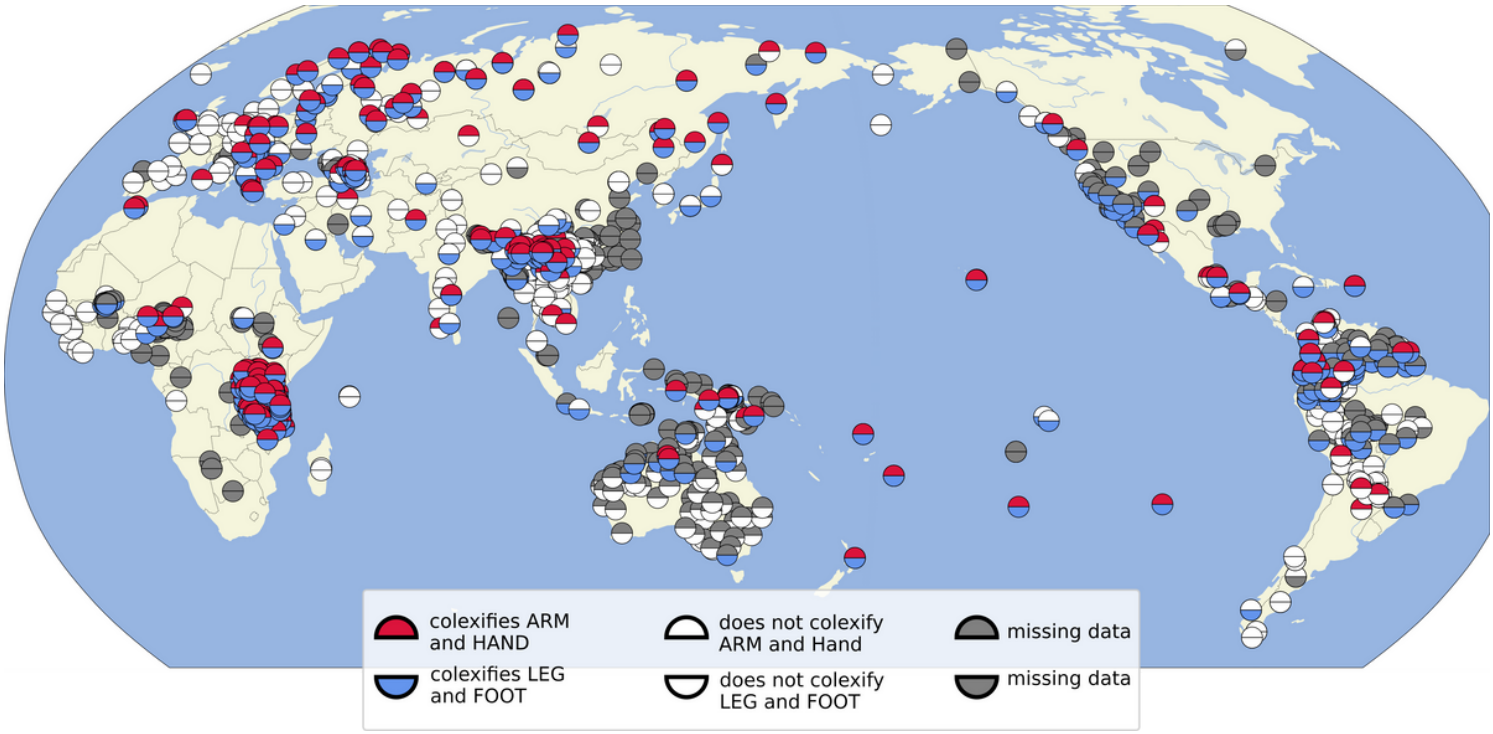

Figure 4

Global distribution of languages in the ClicsCore subset of Lexibank which colexify “arm” and “hand” and “leg” and “foot” respectively.

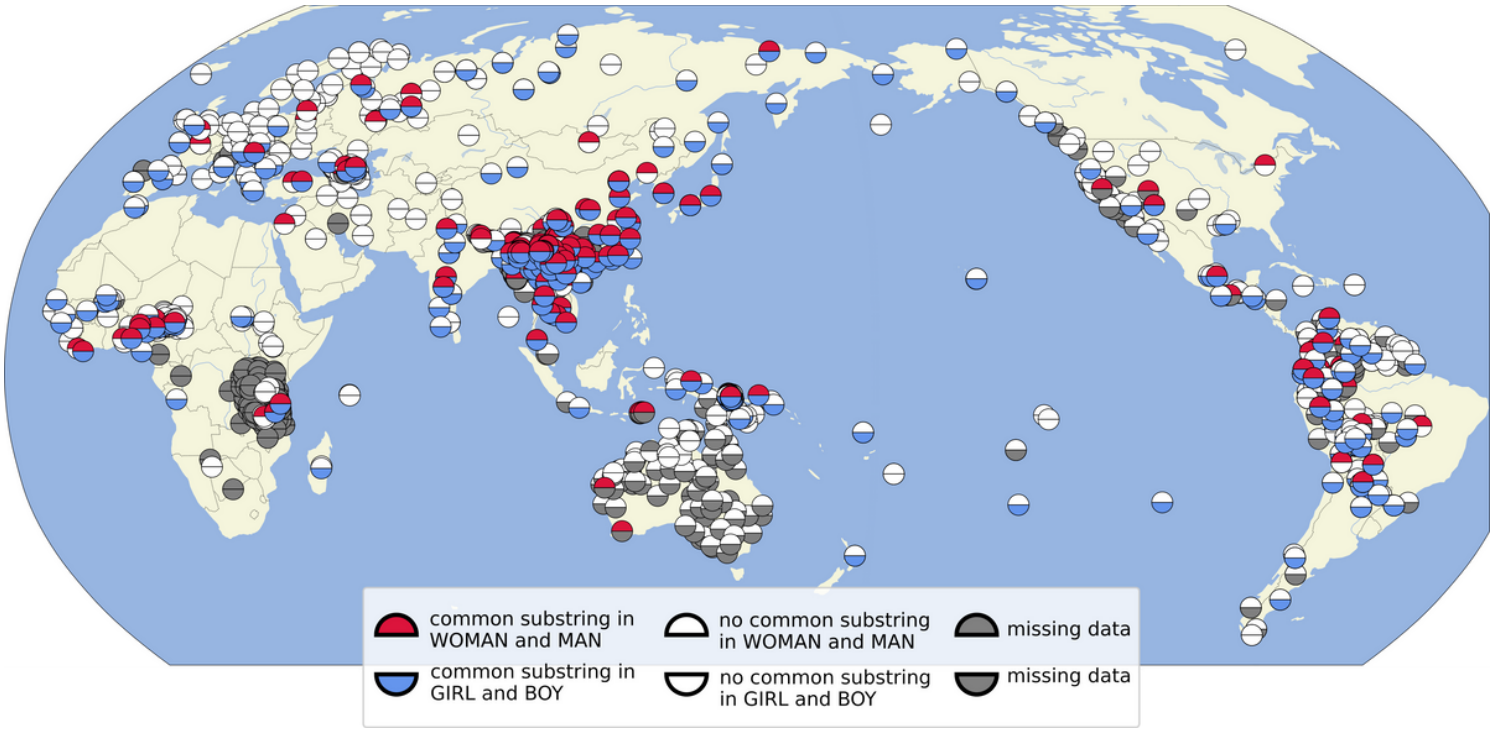

Figure 5

Partial colexifications between “woman” and “man” and between “daughter” and “son”.

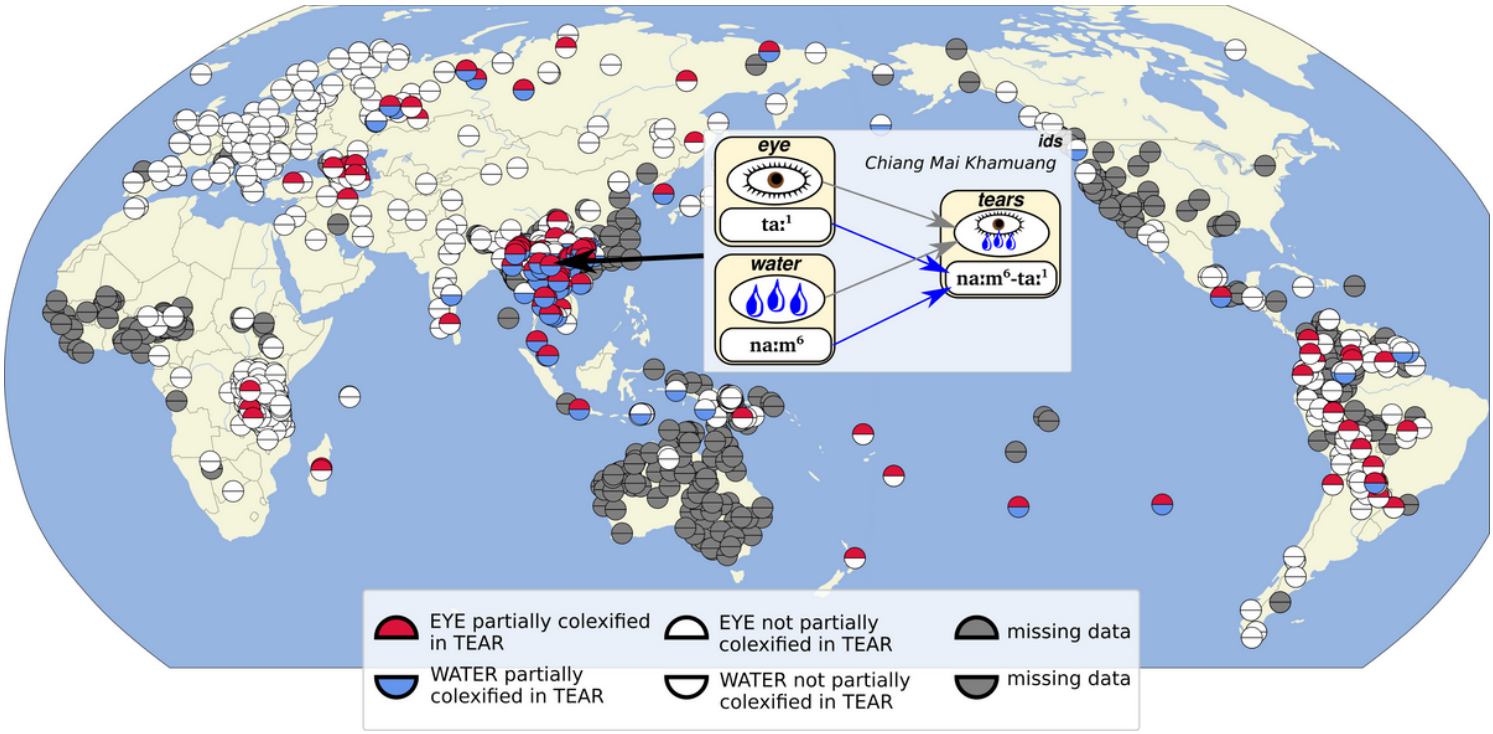

Figure 6

Comparing which languages express “tear” as “eye-water” in the ClicsCore sample of Lexibank.