Check for updates

DATA NOTE

REVISED Lexibank 2: pre-computed features for large-scale lexical data

[version 2; peer review: 3 approved]

Frederic Blum[1,2], Carlos Barrientos[1], Johannes Englisch[1], Robert Forkel[1], Simon J. Greenhill[1,3], Christoph Rzymski[1], Johann-Mattis List[1,2]

[1]Department of Linguistic and Cultural Evolution, Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Saxony, 04103, Germany
[2]Chair for Multilingual Computational Linguistics, Universitat Passau, Passau, Bavaria, Germany
[3]School of Biological Sciences, The University of Auckland, Auckland, Auckland, New Zealand

v2 **First published:** 09 May 2025, **5**:126
https://doi.org/10.12688/openreseurope.20216.1
Latest published: 23 Jun 2025, **5**:126
https://doi.org/10.12688/openreseurope.20216.2

Abstract

Large-scale lexical and grammatical datasets nowadays play an important role in comparative linguistics. However, the lack of standardization remains a challenge exacerbating extension and reuse of published data. We present an updated version of Lexibank, a large-scale lexical dataset, expanding on previous efforts to standardize and unify cross-linguistic data. This new version includes over 3,100 languages and more than one-and-a-half million word forms, substantially broadening the scope and utility of the previous resource. Our dataset has been systematically curated using a dedicated computer-assisted workflow designed specifically for the lifting of published wordlist data to the standards recommended by the Cross-Linguistic Data Formats initiative. The expanded dataset features standardized references to language varieties, standardized semantic glosses that reference the concepts expressed by individual word forms, and standardized phonetic transcriptions for all word forms that our repository contains. Based on those standardizations we pre-compute semantic and phonological features, which can be used to carry out extensive automated analyses. We illustrate this potential by providing dedicated database queries to (1) infer words that are similar in pronunciation and meaning, (2) identify concepts that are colexified across languages in our sample, and (3) assess the semantic diversity of etymologically related words. These queries are not only fast to execute but also global in their scope, due to the largescale coverage provided by Lexibank 2. The queries are also easy to extend, thus having the potential to contribute to various studies in historical linguistics, linguistic typology, and related disciplines. The

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 23 Jun 2025	✓ view	✓ view	✓ view
version 1 09 May 2025	✓ view	✓ view	✓ view

1. **Dan Dediu**, University of Barcelona, Barcelona, Spain
2. **Chundra Cathcart**, University of Zurich, Zürich, Switzerland
3. **César Espin-Riofrio**, Universidad de Guayaquil, Guayaquil, Ecuador

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

updated dataset is a substantial step forward in the effort to create comprehensive, standardized, and accessible linguistic resources.

Keywords
cross-linguistic data, historical linguistics, computational linguistics, lexibank, comparative wordlists



This article is included in the Horizon 2020 gateway.



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

Corresponding authors: Frederic Blum (frederic_blum@eva.mpg.de), Johann-Mattis List (mattis.list@uni-passau.de)

Author roles: Blum F: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Barrientos C**: Data Curation, Writing – Review & Editing; **Englisch J**: Data Curation, Software, Writing – Review & Editing; **Forkel R**: Data Curation, Methodology, Software, Validation, Writing – Review & Editing; **Greenhill SJ**: Data Curation, Writing – Review & Editing; **Rzymski C**: Conceptualization, Data Curation, Methodology, Software, Supervision, Writing – Review & Editing; **List JM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, 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 Europe research and innovation programme (Grant agreement No. 101044282).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Blum F *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Blum F, Barrientos C, Englisch J *et al.* **Lexibank 2: pre-computed features for large-scale lexical data [version 2; peer review: 3 approved]** Open Research Europe 2025, **5**:126 https://doi.org/10.12688/openreseurope.20216.2

First published: 09 May 2025, **5**:126 https://doi.org/10.12688/openreseurope.20216.1

REVISED **Amendments from Version 1**

We have fixed small typos and added a section on the limitation of the queries. We have also re-made our Figure 1 and provided a more reader-friendly map of the language varieties included in the database.

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

Introduction

Large-scale lexical and grammatical datasets have become indispensable tools in comparative linguistics, offering researchers the ability to analyze and compare languages in new ways. Grammatical databases have been used to address questions on linguistic complexity (Shcherbakova *et al.*, 2023; Skirgård *et al.*, 2023), collections of corpus data have been used to study lengthening effects in speech across languages (Blum *et al.*, 2024c), and lexical wordlists have helped to shed light on cross-linguistics aspects of lexical semantics and human cognition (Jackson *et al.*, 2019; Tjuka, 2024). As the field of linguistics continues to make use of digital resources, the demand for standardized and accessible datasets has grown substantially. However, the diversity of data formats and the lack of consistent methodologies across various linguistic datasets remain significant challenges, limiting the full potential of these resources. At the same time, more and more data became available, opening up further challenges with respect to re-use and validation of existing datasets.

One of the primary obstacles for the creation of large-scale linguistic datasets is the lack of standardization in individual resources. Language names used to identify language varieties are often ambiguous and may easily change over time. Semantic glosses often differ greatly across datasets, even if they are intended to elicit the same concepts denoted by individual word forms. Phonetic transcriptions that would describe the pronunciations of word forms are often either lacking completely or they follow idiosyncratic systems that cannot be directly interpreted without further information. When datasets lack uniformity in format, transcription, and structure, it becomes difficult to conduct reliable cross-linguistic studies, drastically restricting the integration and re-use of data from different sources. As a result, the full potential of linguistic data can be rarely employed in cross-linguistic studies, despite the fact that the raw amount of such data has been drastically increasing over the past decades.

The major idea of the Lexibank repository (https://lexibank.clld.org, List *et al.*, 2022) was to address this problem by consistently standardizing as many lexical datasets as possible and providing access to the datasets within a unique repository. When Lexibank was launched for the first time in 2022, the repository consisted of 100 individual datasets from which phonological and lexical features were automatically computed (List *et al.*, 2022). The data in this version of Lexibank (v0.2) thus consisted mainly of *structural features* for a larger number of language varieties that were derived from the underlying data using an automated workflow. Later in 2023, the Lexibank repository was adjusted by providing not only the computed features, but also a selection of the largest wordlists available for each language (List *et al.*, 2023). In this version (v1.0), the wordlists of 74 datasets were provided using Cross-Linguistic Data Formats (CLDF, https://cldf.clld.org, Forkel *et al.*, 2018) to standardize major aspects of lexical data, most importantly the representation of language, meaning, and form. Thanks to the standardization, it was possible to supplement the improved Lexibank version with a CLLD web-application (Forkel, 2014) that offers users convenient ways to browse the data aggregated by Lexibank interactively (https://lexibank.clld.org).

Lexibank builds on CLDF to unify cross-linguistic lexical data along the three basic dimensions of *language*, *concept*, and *form* (for details on these three dimensions in historical linguistics, see List, 2016). In order to ensure that lexical data can be easily compared along these dimensions, Lexibank follows CLDF in using *reference catalogs*. Reference catalogs are large collections of metadata referencing major scientific constructs. While conventions to refer to scientific constructs, such as names for languages, semantic glosses for concepts, or transcriptions for speech sounds, often vary across datasets, traditions, and times, reference catalogs provide a unique anchor point with stable identifiers that allows scholars to reference the constructs consistently. Lexibank uses Glottolog (https://glottolog.org, Hammarström *et al.*, 2024) and its Glottocodes (Forkel & Hammarström, 2022) as a basic reference for languages. For the handling of semantic glosses that refer to individual concepts, the Concepticon catalog is used (https://concepticon.clld.org, List *et al.*, 2025), linking semantic glosses employed in several hundred published concept lists to unique *Concept Sets* (List *et al.*, 2016a). To unify phonetic transcriptions across individual cross-linguistic lexical datasets, the standard transcription defined by the Cross-Linguistic Transcription Systems initiative (CLTS, https://clts.clld.org, List *et al.*, 2024) is used. CLTS links several thousand speech sounds that are represented by a fixed set of feature values across different transcription systems, thereby offering a refined version of the International Phonetic Alphabet (IPA, IPA, 1999) that resolves several ambiguities (Anderson *et al.*, 2018).

Lexibank aggregates wordlist data from a collection of individually compiled CLDF datasets that were all assembled with the help of the *Lexibank workflow*. This workflow builds on CLDFBench (Forkel & List, 2020), a Python package that facilitates the conversion of individual data to CLDF. The Lexibank workflow uses the PyLexibank package (Forkel *et al.*, 2025a), built on top of CLDFBench, to convert lexical data (that may be provided in various idiosyncratic formats) into standardized CLDF datasets. Those datasets link languages to Glottolog, map concepts to Concepticon, and represent phonetic transcriptions in accordance with CLTS (List *et al.*, 2022).

Once the cross-linguistic data has been assembled in standardized form, it is straightforward to extract several features that provide information on certain properties on individual language varieties. Here, Lexibank uses CL Toolkit (https://pypi.org/project/cltoolkit), a software package that extracts phoneme inventory information as well as dedicated phonological and lexical features (syllable complexity, full and partial colexifications, see List, 2023a) from wordlists accessible in CLDF.

All in all, this results in a collection of 30 phonological features, 30 lexical features, and detailed information on phoneme inventory sizes for several thousand languages.

In this study, we present Lexibank 2, a new version of Lexibank that comes with several improvements in comparison to earlier versions. Lexibank 2 now includes data from over 3,100 languages with more than one-and-a-half million different word forms, adding data from more than 1,000 new languages and over 1,000,000 new lexical forms to the collection. The dataset has been systematically expanded and enhanced using an improved data curation workflow that addresses several shortcomings of the previous Lexibank version. In addition, we now enrich the dataset systematically by integrating it with pre-computed features that allow for an improved handling of phonetic and semantic similarities between words across and within the same language. We illustrate the usefulness of these new features by providing fast and efficient queries that allow us to explore the data in various ways, opening up new possibilities for linguistic research. By providing a comprehensive and comparable resource, it enables automated inferences of phonological and lexical features across languages, facilitating large-scale comparative studies. As a freely available resource, Lexibank aims to support and stimulate further studies into the diversity of human language through the combination of digital tools and collaborative efforts.

Materials and methods

Datasets

Lexibank 2 consists of 134 individual CLDF datasets, covering 3,107 languages from 207 language families with a total of 1,734,794 word forms. All datasets have been standardized by linking languages to Glottolog (Hammarström *et al.*, 2024, Version 5.1), mapping concepts to Concepticon (List *et al.*, 2025, Version 3.4.0), and using CLTS (List *et al.*, 2024, Version 2.3) as the basis for phonetic transcriptions. Similar to previous versions of Lexibank, the preparation of the data featured in Lexibank 2 was based on collaborative efforts. Individual teams coordinated the different tasks involved in the Lexibank workflow for CLDF conversion, making sure that concept lists were properly linked to Concepticon by extending the Concepticon reference catalog (Tjuka *et al.*, 2023), double-checking the assignment of languages to Glottocodes, and discussing open questions in the creation of orthography profiles (Moran & Cysouw, 2018), which are needed to convert original transcriptions to standardized CLTS representations. Additional corrections included the removal of duplicate entries that resulted from individual conversion errors in individual datasets, the adjustment of the citation information, and the correction of unnatural phonetic transcriptions that we detected by investigating consonant clusters in the data.

The full table of datasets is provided in Table 7, Appendix A. When preparing the data to be included in Lexibank 2, all datasets were thoroughly checked by us in teams, paying particular attention to the phonetic transcriptions, the mapping to Glottolog, and the linking of the concepts to Concepticon. After all these checks had been carried out and new versions of our three reference catalogs (Glottolog, Concepticon, and CLTS) had been integrated, new releases of the data were carried out and archived with Zenodo. From these releases – which we store along with additional bibliographic information in a separate table – the Lexibank 2 data was then aggregated automatically. The updated version of Lexibank improves on the previous release by further increasing data quality. For example, all datasets now have their concept list mapped to Concepticon (List *et al.*, 2025). This makes it easier to compare the data across individual datasets, with benefits that we will highlight in the next section. All new datasets have been compiled with individual orthography profiles (Moran & Cysouw, 2018), mapping the graphemes from the original source to the base transcription systems underlying CLTS. Datasets for which this was not feasible have been excluded from the release. Individual datasets, like the Intercontinental Dictionary Series, received major updates for subsets of the data (Miller & List, 2024), which have been incorporated into Lexibank as well. The improved data quality criteria ensure that Lexibank maintains high quality standards, while making data from different studies reusable.

Data collections

The datasets and language varieties in Lexibank 2 are assigned to one or more *collections*, depending on the basic information they convey. A collection is a subset of wordlists in the repository that conforms to a specific criterion. One and the same wordlist for one and the same language variety can be assigned to different collections at the same time. Figure 2 shows the criteria to assign individual datasets to collections. In order to be included into the Lexibank repository, a dataset must be available in the form of a CLDF wordlist in which concepts have been mapped to Concepticon, languages are linked to Glottolog, and phonetic transcriptions follow the CLTS transcription system. The LexiCore collection (constituting the basis of all data in Lexibank) consists of all language varieties that have at least 80 different concepts mapped to Concepticon. This assures a minimum amount of coverage for comparative studies and provides the basis for the automated computation of sound inventories and phonological features that was introduced in Lexibank 0.2 (List *et al.*, 2022). Similarly, ClicsCore consists of wordlists for individual languages with at least 180 different concepts mapped to Concepticon. This collection is used for the automated computation of lexical features, including individual full colexifications (Rzymski *et al.*, 2020) and partial colexifications (List, 2023a). Datasets and language varieties assigned to the CogCore collection include manually identified cognate sets that show which words are etymologically related (for an extended reuse of this subset, see Häuser & List, 2025). The ProtoCore collection builds upon CogCore and includes only those cases where additionally proto-forms have been reconstructed.

The assignment of languages to collections is done at the level of the datasets but adds – where possible – a computational component. All datasets that were selected for the inclusion in Lexibank 2 were assigned to one or more subsets before aggregation. When aggregating the data, however, those properties that can be inferred automatically (e.g. the number of mappings to distinct Concept Sets in Concepticon), may override the decisions provided prior to aggregation at the level of the datasets. Figure 1 presents a map with all languages in the data with information about the membership of each language in the basic collections.

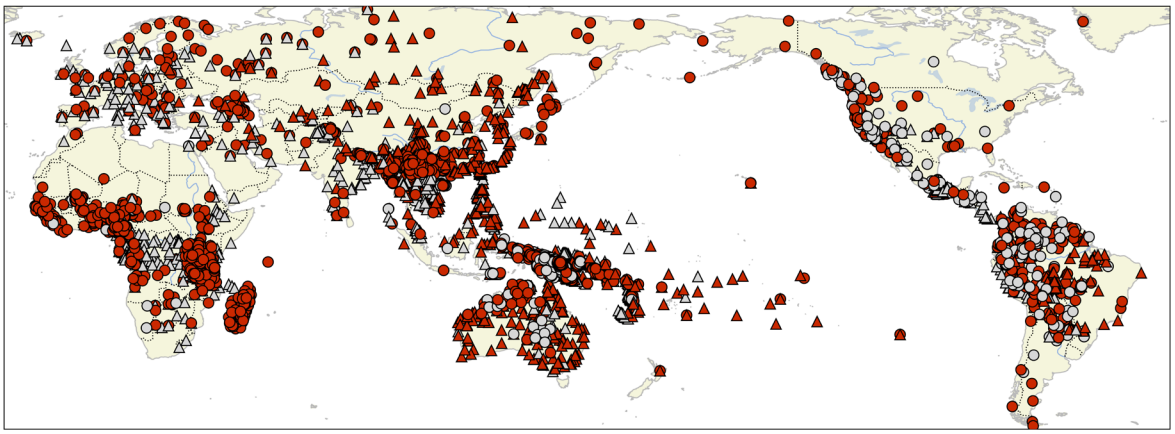


Figure 1. Map showing the global distribution of languages attested in Lexibank 2.

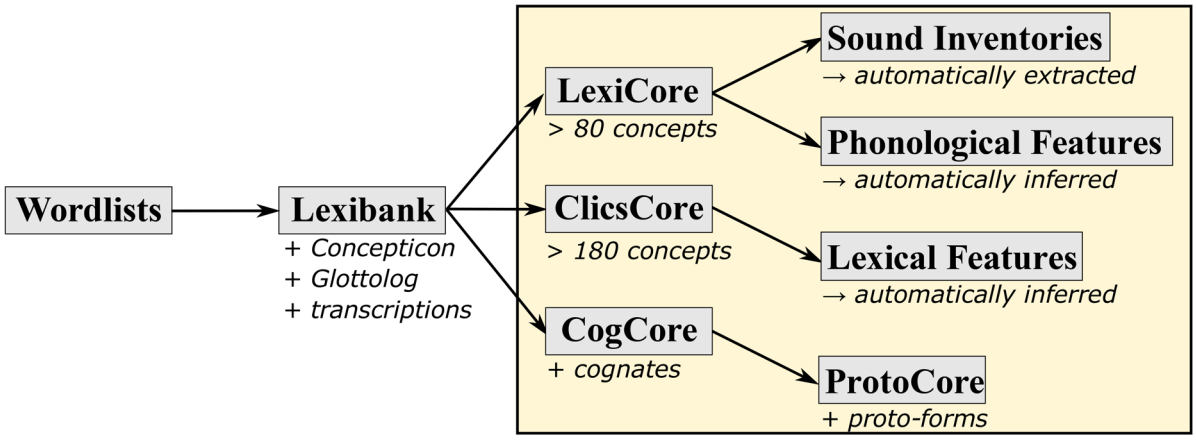


Figure 2. Criteria for the assignment of individual wordlists for individual language varieties to the basic collections of Lexibank 2.

Workflow for data curation

With Lexibank 2, we have now not only increased the number of datasets from 76 to 134, we have also improved the selection of datasets and the criteria to select datasets for the inclusion in Lexibank. In addition, we have further enhanced the aggregation workflow, modifying the practice of selecting only data from specific language varieties. Instead, we now aggregate the data from all datasets while at the same time providing interested users with the possibility to select only a balanced subset of the data in which every language variety is only represented by a single wordlist taken from one dataset. In the following, we will briefly discuss these three major improvements made to the new version of the repository.

Phonetic transcription as the norm. As a first improvement, Lexibank no longer offers datasets that do *not* provide standardized phonetic transcriptions. All 134 datasets are available in phonetically transcribed form, with the transcriptions being individually produced with the help of the workflows that were explained in detail in List *et al.* (2022). In order to enhance the previous Lexibank collection, which had still included several collections of wordlists lacking phonetic transcriptions entirely, we followed a two-fold strategy, according to which we deliberately removed certain datasets that lack phonetic transcriptions from the Lexibank collection, while at the same time investing additional time to compute phonetic transcriptions for several datasets that would only provide orthographic forms before (Hantgan & List, 2023; List, 2023a; Miller & List, 2024). In order to ease the curation of individual datasets, we added new routines dedicated to checking the sanity of the transcriptions in the Lexibank datasets by extending the `PyLexibank` package (Forkel *et al.*, 2025a). These include detailed commands allowing us to check the conversion process of graphemes to CLTS-conform segments, as well as a specific command that infers large consonant clusters from the transcription data in order to point to potential problems in the transcription process (see List, 2018 for an early example on consonant cluster inference).

Improved workflow for data aggregation. While individual word forms from selected languages were already introduced with Lexibank 1, this version did not provide access to all

datasets upon which the repository was originally built. Since Lexibank aggregates data from different sources, there are many cases in which information on the same language can be obtained from multiple datasets that differ regarding the size of the data they cover and in the individual concepts for which they provide translations. When aggregating data for Lexibank 1, the aggregation workflow would select only one dataset per language variety, using the number of word forms as the sole selection criterion. All alternative sources would be disregarded and excluded from both the comparative word-list and the web application. The aggregation workflow in Lexibank 2 introduces an alternative aggregation technique. Instead of deliberately excluding datasets from the repository, we now retain all data, while at the same time flagging those datasets that provide the largest number of words for a given language variety. This allows users to work both with *all* data available in Lexibank, while at the same time being able to select only one representative per language – where needed. We consider the variety of each glottocode as the most representative which has the most distinct concept mapped to Concepticon. Those varieties are tagged as part of the *Selexion* collection. As an additional improvement, the new workflow for data aggregation now assigns sources at the level of all individual word forms, thus allowing users to trace each datapoint to the original source from which it was taken.

Pre-computing phonetic and semantic features. As a third improvement, we expand the data with pre-computed features that systematically integrate information from the reference catalogs that are linked in the individual source datasets. In contrast to the previous abstract features included in Lexibank, the phonetic and semantic features allow us to compare and analyze the individual word forms. With respect to phonetic features, we systematically convert the segmental representation of word forms in Lexibank to two different sound class representations. Sound classes have been originally proposed by Dolgopolsky (1964), who lumped all consonants into ten major classes, assuming that correspondences within classes are more likely to occur inside than across classes (see List, 2014b for further details). In several computational approaches to cognate detection, scholars use Dolgopolsky's consonant classes to identify words that are potentially related by assigning all words that match in their first two consonant classes to the same cognate set (Turchin *et al.*, 2010). This approach has also successfully been used for large-scale language classification Blum *et al.* (2025b). In addition to representing all phonetic transcriptions by their Dolgopolsky sound classes, Lexibank 2 also provides the sound class model proposed by List (2014b). Both sound class representations in Lexibank 2 are retrieved from the CLTS reference catalog (List *et al.*, 2024).

With respect to semantic features, we employ the colexification network by Rzymski *et al.* (2020) to assign all concepts that gloss the word forms in Lexibank 2 to their *central concept*. The central concept here refers to the concept that contains most links to other concepts within the community to which it was assigned during the community analysis carried out in the creation of the CLICS[3] database. The central concepts allow us to determine in a straightforward way whether two words recur in the same community in colexification analyses and thus provides a very rudimentary measure for semantic similarity. In addition, all concepts are tagged with respect to the most common basic vocabulary lists in which they occur. This allows us to filter individual analyses in such a way that only basis concepts are taken into account. Information on central concepts and on the basic concept lists in which concepts occur are retrieved from the Concepticon reference catalog (List *et al.*, 2025). Since the usage of pre-computed features is still rather experimental at this stage, we restrict the number of features provided to the ones mentioned here. In the future, however, the amount of pre-computed features can be easily extended further.

Queries for data exploration

In 2020, Bernard Comrie approached us asking if we could help to compare individual word forms in one language against the whole of the data in Lexibank (which had by then not yet been officially published). Inspired by this idea, we began to test different search methods over the past years, trying to offer some kind of a service that would make it easy for scholars interested in particular aspects of Lexibank data to obtain quick results in order to start with a more thorough data exploration afterwards. These queries were never meant to be the endpoint of analysis, but rather as a convenient starting point for *exploratory data analysis* (Morrison, 2014), helping scholars working with cross-linguistic data to search for interesting patterns before initiating a more targeted and more thorough investigation. Later, this developed into the idea to extend queries to a set of *services*, offering a functionality similar to the BLAST tool in evolutionary biology (Altschul *et al.*, 1990). While BLAST itself uses sophisticated methods for automated sequence comparison to assemble its matches, the basic fuel that powers our query engine are the pre-computed phonetic and semantic features, as they allow us to identify phonetically and semantically similar words in a very fast way. Still far away from the power and possibilities of BLAST – also resulting from the fact that linguistic data and processes differ from data in bioinformatics (Blum & List, 2023; List *et al.*, 2016b; List, 2023b) – we consider the queries that we discuss in detail in the following as a first and promising attempt that could turn into a dedicated *Basic Linguistic Search Service* (BLISS) that runs on the top of Lexibank 2.

When experimenting with different ways to query the data in Lexibank, we realized that the most efficient way to deal with large CLDF datasets is to convert the tabular files in CLDF to an SQLite database file (https://sqlite.org). SQLite is a database software that allows to conduct SQL queries on local files. It is listed among the *recommended formats for datasets* of the Library of Congress, which lists data formats that focus on preservation and long-term access (LoC, 2024). Given that CLDF makes use of the CSVW specification for tabular data on the web (https://csvw.org, Gower, 2021) and represents all data with the help of CSV tables with additional metadata in JSON, any valid CLDF dataset is

fundamentally equivalent to a relational database. Accordingly, the `PyCLDF` software package (https://pypi.org/project/pycldf, Forkel *et al.*, 2025b) allows to easily convert any CLDF dataset to SQLite. Instead of accessing individual tables from Lexibank in the form of the original CSV files, we can design SQL queries that solve particular problems and apply them to the SQLite representation of Lexibank data. The advantage of using SQL queries compared to designing search commands on CSV files is a great gain in execution speed. While it may take several minutes on ordinary computers to load the CSV files with Python, not to speak of loading the data with the help of the dedicated `PyCLDF` package, an SQLite query is usually finished after seconds. Another advantage of SQL queries is that the queries themselves can be easily shared and extended. All that users need to conduct them is a valid installation of SQLite – which is available on most machines.

In order to illustrate the potential of pre-computed features along with the SQLite representation of Lexibank 2, we provide a tutorial with a set of SQLite queries that can be easily used to explore the data in Lexibank in various ways. While these initial queries are only illustrative in nature, we expect that it will be easy to expand them in the future. The three illustrative queries that we offer focus on (1) the automated detection of words that are similar in pronunciation and similar in meaning, (2) the automated inference of cross-linguistic colexifications from the data, and (3) the automated inference of partial colexification patterns. The full tutorial on different queries is published on Codeberg (https://codeberg.org/lexibank/lexibank-queries/src/tag/v0.1).

Query Lexibank 2 for similar word forms. Our first set of queries retrieves all those word forms in Lexibank 2 that are similar in form and meaning. Similarity in form is measured with the help of *matching consonant classes*, as provided as part of the pre-computed features that extend the regular wordlist aggregated from the individual datasets. Similarity in meaning is measured in a binary fashion by treating all those concepts as similar that are linked to the same central concept and therefore occur in the same semantic community that was inferred for CLICS[3]. Our queries can be applied in two ways. First, users can define an initial sound class sequence and an initial concept and then search for all word forms in Lexibank 2 that match the first two sound classes and the central concepts. Second, we can compute entire cognate sets within and across language families.

Query Lexibank 2 for cross-linguistic colexifications. A second example for a fast access of the Lexibank data is the fast retrieval of colexifications. While the CLICS[3] database (Rzymski *et al.*, 2020) offers a computation of most of the colexifications found in the first version of Lexibank (List *et al.*, 2022), it does not cover all concepts present in the data due to a high threshold of concepts for inclusion in ClicsCore (>180, see Figure 2). Through a simple query, we can retrieve cases of colexifications in the dataset for two specific concepts. One can easily add different concepts to the query, with the only requirement being that they are covered by the Concepticon reference catalog and properly linked to it (List *et al.*, 2025). The output provided is a table of all languages in which the colexification in question can be found along with the word forms in the respective languages.

Query Lexibank 2 for partial colexifications. In contrast to full colexifications that refer to those cases where one word form expresses two or more concepts in the same language, partial colexifications refer to those cases where the word forms expressing two or more different concepts share a certain amount of common lexical material. Partial affix colexifications have so far only been calculated on a small subset of the Lexibank data (a segmented form of the Intercontinental Dictionary Series, https://ids.clld.org, Key & Comrie, 2016b, see List, 2023a). Treating word forms in phonetic transcriptions as sound sequences, List (2023a) – inspired by Urban (2011) – suggests to compute *affix colexifications* from multilingual wordlists by searching for those cases where one word recurs in another word form within the same language variety. The resulting relations (one word being part of another word) can be handled with the help of a directed graph, and statistics similar to the ones computed for full colexifications can be derived, counting, for example, the number of language families in which a certain partial colexification can be identified.

While List (2023a) presents a new method that can efficiently search for partial affix colexifications, we realized that the same result can likewise be achieved with the help of a dedicated query on the Lexibank 2 database. This query iterates over all word pairs that express different concepts within the same language variety and then checks for each word pair if one recurs at the beginning or the end of the other word pair. While List (2023a) correctly points out that this strategy takes a lot of time when running it in Python, it turned out that the computation takes much less time when using the SQLite representation of CLDF. In addition, we can extend the query and not only search for matching substrings in the phonetic entries provided by Lexibank 2, but also apply the same procedure to the sound classes that Lexibank 2 provides as pre-computed features.

Dataset validation

Benefits of standardization

The standardization provided by CLDF adds several quality checks to the data in Lexibank 2. Overall, the updated version offers more data in better quality, and thus offers new possibilities for cross-linguistic comparative research. In addition to the aforementioned increase in data points, we fixed errors in Glottocodes, duplicated data points, and incorrect naming of language families. Many of those cases have been brought to our attention by the community of Lexibank users, for which we are very grateful. Furthermore, we have implemented three explicit checks that highlight the benefits of standardization.

The first implemented check is to compare all converted s egments against CLTS (List *et al.*, 2024) to confirm that all segments are represented by standardized symbols. In this way, we avoid the typical pitfalls of Unicode and phonetic transcriptions from which many linguistic datasets suffer (Moran & Cysouw, 2018). In different linguistic traditions, identical phone qualities might be transcribed with many different graphemes. Similarly, the same grapheme might represent different phonetic qualities. Using the CLTS mapping through orthography profiles, we ensure the comparability that is needed for computer-assisted studies across large datasets such as Lexibank.

When preparing the data for Lexibank 2, we have detected an additional problem that had been overseen in previous Lexibank versions. In order to enhance the phonetic transcriptions, we now compute statistics on consonant clusters, based on the assumption that all clusters in an individual dataset should be roughly checked by experts, paying particular attention to clusters with more than three consonants. This second check allowed us to enhance the individual orthography profiles that we use to convert the original orthographies to phonetic transcriptions in many cases. This check has been run for all individual source datasets of Lexibank 2. Borderline cases, where up to ten consonants would show up in a sequence without a vowel often turned out to go back to the handling of the grapheme <y>, which is often used to represent the glide [j], following a specific transcription practice frequently used in African and South American linguistics, or <ɤ> mapped to [ɣ] (fricative) instead of [ɤ] (vowel). Having identified these problems, we went back to all initial datasets in order to correct these problems.

We applied a third quality check that was an active search for unwanted duplicates in datasets. Duplicates here refer to those cases where one language has two or more identical word forms for the same concept. While one might expect that these cases should be low in general, we detected several thousand duplicates across the 134 datasets and later corrected the errors on the basis of individual datasets as well as by modifying the aggregation procedure. When aggregating Lexibank 2 data, a list of duplicates entries is created automatically and written to a file that is part of the release of the data. Identical entries themselves are not included into the aggregated wordlist.

Comparing Lexibank 1 and Lexibank 2

We provide a quick comparison between Lexibank 2 and the comparative wordlist shared as part of the previous version of Lexibank in Table 1. From these major characteristics, including the number of language varieties (all doculects from which individual wordlists are populated, including both dialect varieties of the same language and multiple sources for the same language), the number of languages (represented by the number of distinct Glottocodes in both wordlists), and the number of word forms, it can be easily seen that Lexibank 2 drastically expands upon the scope of Lexibank 1.

Table 1. Comparison of Lexibank 1 and Lexibank 2.

Criterion	Lexibank 1	Lexibank 2
Datasets	76	134
Varieties	2 028	5 477
Languages	2 028	3 107
Families	192	207
Concepts	3 033	3 205
Words	709 467	1 734 794
Duplicates	12 153	34

With respect to the differences regarding the number of datasets (76 in Lexibank 1 vs. 134 in Lexibank 2), it has to be noted that these refer to the *active datasets*, i.e., datasets that contribute one or more individual wordlists to the comparative wordlist aggregated from the CLDF datasets. Here, we can see that although Lexibank 1 sources data from *more* than 76 CLDF datasets (the wordlist data tries to aggregate from 98 CLDF datasets), there are as many as 22 datasets that are considered but fail to meet the criteria for inclusion into the final wordlist collection. When aggregating data for Lexibank 2, we have made sure that we only include those CLDF datasets into our collection that provide individual wordlists that meet the basic criterion for inclusion, providing at least 80 distinct concepts with phonetic transcriptions. In the table, we have also listed the number of duplicates that we could infer from both datasets when querying the SQLite version of both comparative wordlists. This shows that our targeted efforts aiming at identifying and removing duplicates have been quite successful. We will likewise address the remaining 34 in future versions of Lexibank.

Lexibank 2 ist not only based on adding more data to the existing data in Lexibank. In addition, we have also removed eight datasets in the new version, given that they failed to meet our criterion of providing clear phonetic transcriptions for all entries. The tutorial accompanying this study shows how these datasets can be identified and provides some additional information regarding the reasons that led to their exclusion.

Queries on pre-computed features

Our queries on pre-computed features allow to search for similar word forms, colexifications, and partial affix colexifications in Lexibank 2. While these queries are useful to explore the data in various ways, they also offer us the possibility to validate the data by checking if the results of the queries correspond to our expectations.

Similar word forms. The first of our queries for similar word forms takes as input a sequence of sound classes and a concept and then searches through all word forms in Lexibank 2

in order to identify words that resemble the word phonetically and semantically. This procedure should not be confused with a valid method for the detection of cognate words (see List *et al.*, 2018 on the latter), but rather as a quick way to allow scholars to get an overview on potential candidates for etymologies, borrowings, or coincidentally similar words in the languages of the world. As an example, we ran a query that searches for (1) words that contain Dolgopolsky's sound classes T and R as the first two consonant classes, expressing (2) concepts that occur in the same community as WOOD in CLICS[3], and recurring in (3) languages that form part of the Indo-European family. The query yields a total of 162 matches, covering the 32 different concepts in Table 2. Of these entries (all provided in the accompanying online tutorial), not all words are truly cognate. What we can see instead is that the method picks up several true cognates, usually within the same concept, while cognates attested across two or more concepts are rather rare.

This shows that – as we have mentioned before – the query should not be confused with a simple cognate detection method, but rather taken as a way to broadly search a dataset for words that exhibit rough similarities in form and meaning. When restricting the number of concepts further, however, we can usually find clusters of words that mostly *are* cognate. As an example, consider the word forms in Table 3, which result from the same query but have been filtered for the concepts TREE and WOOD, respectively. Here, we find the typical reflexes of Proto-Indo-European **doru-* 'tree / wood' that are prominent both in Slavic and Germanic languages.

From this experimental query on similar words in Indo-European, we can learn two lessons. On the one hand, we find that the consonant class matching approach by Dolgopolsky (1964), although pretty simple in nature, works sufficiently well and also sufficiently fast to query large databases such as Lexibank 2 for similar word forms. On the other hand, however, we can see that relaxing semantics, searching for semantically similar rather than semantically identical words, yields large amounts of unrelated words. This shows that any long-range approaches, trying to prove deep genealogical relationships that predate the time depth of the families established by traditional methodology, should be taken with great care, specifically when they relax the role that semantics plays in determining etymologically related words (see Ratcliffe, 2012 for similar arguments on this matter, and Blum *et al.*, 2024a for verifying such cognate sets through colexifications).

Cross-linguistic colexifications. Lexibank 2 continues previous Lexibank versions in inferring several cross-linguistic colexifications as lexical language features and sharing them in the form of a dedicated CLDF dataset published as integral part of the repository. Given that cross-linguistic colexifications on a large scale can be easily inspected through the CLICS database, computing additional colexifications through Lexibank 2 may seem unnecessary, especially since the most recent version of CLICS – CLICS[4] – now also models colexifications as lexical language features (Tjuka *et al.*, 2025). However, since

Table 2. Concepts along with the occurrences in the sample query searching for words in which the first two consonant classes are T and R and the concepts occur in the community with WOOD as central concept in CLICS[3].

Concept	Occurrences
BEAM	5
BEESWAX	3
CANDLE	2
CHISEL	9
CLUB	1
CONCEPT	1
COURT	3
CUSTOM	1
DOORPOST	7
EARWAX	3
FIREWOOD	3
FOOTPRINT	5
FOREST	1
HELL	1
LAW	1
MANNER	6
MAST	2
MATCH	1
NEEDLE (FOR SEWING)	2
PATH	7
PATH OR ROAD	4
POST	4
RAFTER	1
RIDGEPOLE	1
ROAD	7
STICK	3
STREET	1
TORCH OR LAMP	3
TREE	38
TREE STUMP	2
TREE TRUNK	15
WOOD	18

Table 3. Restricting the broad query for words with initial consonant classes T and R to the concepts TREE and WOOD (showing only five matches for each concepts).

Language	Concept	Form
Belarusian	TREE	d r ɛ v a
Bulgarian	TREE	d ə r v o
Danish	TREE	t ʁ ɛ: ʔ
English	TREE	t r i:
Macedonian	TREE	d ə r v ɔ
Old Church Slavonic	WOOD	d r u v a
Russian	WOOD	d rʲ ɪ vʲ ɪ sʲ i n ə
Slovak	WOOD	d r ɛ v ɔ
Swedish	WOOD	t r ɛ:
Ukrainian	WOOD	d r ɔ w ɑ

Table 4. Colexification of SUN and MOON in Lexibank inferred with the query for cross-linguistic colexifications.

Language	Family	Form
Giacone	Arawakan	k e: r i
Bora	Boran	n ɯ ʔ p a
Mirana	Boran	n ɯ ʔ p a
Muinane	Boran	n ɨ ʔ ɨ b a
Guahibo	Guahiboan	h u a m e t o
Guayabero	Guahiboan	h u i m t
Playero	Guahiboan	h u a m e t o
Nukak	Kakua-Nukak	w i d ʔ
Matses	Pano-Tacanan	u ʂ ɨ
Barasana	Tucanoan	b ũ h ĩ h ũ
Cubeo	Tucanoan	a w i a
Desano	Tucanoan	a b e
Piratapuyo	Tucanoan	a s ĩ
Siriano	Tucanoan	a b e

the major idea behind CLICS is to assemble large-scale colexification networks that can then be inspected manually and automatically, it may still be useful to be able to check for individual colexification patterns, if they *are* reflected in Lexibank 2, and if so, which word forms and which languages exhibit them. In our sample query that illustrates how individual colexifications can be queried, we search for potential colexifications of SUN and MOON. While these may seem to be extremely rare at first sight, it turns out that they are reflected in many different language varieties from several different language families in South America. Table 4 shows the results of the query (with most examples coming from the databases of Chacon, 2017 and Blum *et al.,* 2024b). As can be seen from the individual forms that exhibit the colexification across the six different language families, words inside the same language family all seem to go back to the same root, while no commonality between forms across language families can be identified. This shows that this colexification pattern, as distinct as it may seem, seems to have originated at an earlier point in time in most language families. One may further speculate that the areal frequency of the correlation points back to ancient contact situations.

Partial affix colexifications. Thanks to the SQLite representation of the data in Lexibank 2, it is straightforward to compute partial affix colexifications for a much larger dataset now. Running our dedicated query, we receive as many as 360,649 individual instances of partial colexifications from all datasets in the sample. As shown already in List (2023a), this exceeds the amount of full colexifications that one typically receives largely and also makes clear why it is so important to filter these kinds of colexifications rigorously, retaining only those that recur across several language families. That the method nevertheless retrieves valid and interesting results, can be seen when checking only the 15 most frequently recurring patterns of affix colexifications, which are shown in Table 5.

As a huge advantage of the pre-computed features in Lexibank 2, the investigation of partial colexifications in this form is not limited to the phonetic sequences, but can also be applied to the sound class strings. While this will increase the noise even further, it may help to capture certain cases of allophony in word families that one would not be able to observe when sticking to phonetic transcriptions. Running the modified query on Lexibank 2, using Dolgopolsky's consonant classes as the major model for sound classes results in a huge increase in observed affix colexification pairs, with 7,930,552 instances. However, when restricting the comparison only to the most frequently recurring 15 colexifications, as shown in Table 6, we find a picture much similar to the one that we have observed in Table 5.

All in all, this shows that it may be worthwhile to further investigate partial affix colexifications computed from sound class representations. Apart from comparing the resulting networks or the individually inferred colexifications, it may also be interesting to see to which degree the resulting concept embeddings computed from sound class colexifications differ from concept embeddings computed from phonetic transcriptions (Rubehn & List, 2025).

Table 5. Most frequent partial affix colexifications in Lexibank 2.

#	Source	Target	Fams.
1	TEN	ELEVEN	81
2	TEN	TWELVE	80
3	TWO	TWELVE	75
4	ONE	ELEVEN	74
5	FOOT	TOE	58
6	TREE	BARK	56
7	HAND	FINGER	49
8	TWO	SEVEN	48
9	FOUR	NINE	47
10	WOMAN	GIRL	47
11	BREAST	NIPPLE	45
12	MOUTH	LIP	45
13	SKIN	BARK	45
14	TWO	TWENTY	45
15	TEN	TWENTY	42

Table 6. Most frequent partial affix colexifications in Lexibank 2 when computing partial affix colexifications from sound class strings.

#	Source	Target	Fams.
1	TEN	ELEVEN	91
2	TEN	TWELVE	89
3	ONE	ELEVEN	86
4	TWO	TWELVE	81
5	FOOT	TOE	64
6	TREE	BARK	63
7	WOMAN	GIRL	55
8	HAND	FINGER	54
9	THREE	EIGHT	53
10	FOUR	NINE	50
11	SKIN	BARK	50
12	TWO	SEVEN	50
13	TWO	TWENTY	50
14	WHAT	WHY	50
15	MAN	BOY	49

Table 7. Table of all datasets, their number of concepts, the number of concepts mapped to Concepticon, as well as the source of the individual datasets.

Dataset	Concepts	Concepticon	Source
aaleykusunda	230	230	Aaley & Bodt (2019)
abrahammonpa	306	304	Abraham *et al.* (2005)
abvdoceanic	191	191	Greenhill *et al.* (2008)
abvdphilippines	210	210	Greenhill *et al.* (2008)
allenbai	499	499	Allen (2007)
backstromnorthernpakistan	1233	224	Backstrom & Radloff (1992)
baf2	347	335	Hantgan *et al.* (2022)
bantubvd	430	415	Greenhill & Gray (2015)
barlowkilliantomoip	996	623	Barlow & Killian (2023)
barlowlote	551	443	Barlow (2024)
beidasinitic	905	738	Dàxué (1964)
berrywestpapuan	195	195	Berry & Berry (1987); Kamholz *et al.* (2024)
birchallchapacuran	125	125	Birchall *et al.* (2016)
blumpanotacana	501	501	Blum *et al.* (2024b)

Dataset	Concepts	Concepticon	Source
blustaustronesian	210	210	Greenhill *et al*. (2008)
bodtkhobwa	662	542	Bodt & List (2022)
bowernpny	344	341	Bowern & Atkinson (2012)
bremerberta	200	183	Bremer (2016)
cals	184	184	Mennecier *et al*. (2016)
carvalhopurus	205	173	de Carvalho (2021)
castrosui	592	518	Castro (2011)
castroyi	540	532	Castro *et al*. (2010)
castrozhuang	510	490	Castro & Hansen (2010)
chacolanguages	323	223	Brid *et al*. (2022)
chaconarawakan	102	100	Chacon (2017)
chaconbaniwa	243	233	Chacon *et al*. (2019)
chaconcolumbian	128	125	Chacon (2017)
chaconnorthwestarawakan	94	94	Chacon (2022)
chacontukanoan	142	127	Chacon (2014)
chenhmongmien	883	799	Chén (2012)
chindialectsurvey	460	451	Language and Social Development Organization (2019)
chingelong	99	98	Chin (2015)
clarkkimmun	757	636	Clark (2008)
constenlachibchan	110	110	Constenla Umaña (2005)
crossandean	150	150	Blum *et al*. (2023)
davletshinaztecan	100	100	Davletshin (2012)
deepadungpalaung	100	100	Deepadung *et al*. (2015)
dhakalsouthwesttibetic	243	240	Dhakal *et al*. (2024)
dravlex	100	100	Kolipakam *et al*. (2018)
dunnaslian	146	146	Dunn *et al*. (2013)
dunnielex	207	207	Dunn (2012)
duonglachi	100	100	Duong *et al*. (2020)
felekesemitic	150	139	Feleke (2021)
galuciotupi	100	100	Galucio *et al*. (2015)
gaotb	100	100	Gao (2020)
gerarditupi	244	242	Gerardi *et al*. (2021)
gravinachadic	717	716	Gravina (2014)
grollemundbantu	100	100	Grollemund *et al*. (2015)
halenepal	997	732	Hale (1973)
hantganbangime	300	300	Hantgan & List (2022)
hattorijaponic	200	200	Hattori (1973)

Dataset	Concepts	Concepticon	Source
heathdogon	944	880	Moran *et al*. (2016)
houchinese	140	139	Hóu (2004)
hubercolumbian	366	347	Huber & Reed (1992)
huntergatherer	744	408	Bowern *et al*. (2021)
idssegmented	1310	1308	Key & Comrie (2016a)
iecor	170	170	Heggarty *et al*. (2024)
ivanisuansu	250	250	Ivani (2019)
johanssonsoundsymbolic	344	284	Johansson *et al*. (2020)
joophonosemantic	100	100	Joo (2019)
kesslersignificance	200	200	Kessler (2001)
keypano	1310	1308	Miller & List (2024)
kleinewillinghoeferbikwinjen	119	117	Kleinewillinghöfer (2015)
kochtukanoan	804	445	Koch-Grünberg (1914)
kraftchadic	433	429	Kraft (1981)
lairgyalrong	291	291	Lai & List (2023)
lamanisoic	301	283	Lama (2012)
leeainu	199	199	Lee & Hasegawa (2013)
leecaijia	234	234	Lee (2022)
leejaponic	210	210	Lee & Hasegawa (2011)
leekoreanic	246	246	Lee (2015)
lieberherrkhobwa	100	100	Lieberherr & Bodt (2017)
lindseyende	1129	1128	Lindsey (2019)
lionnetyotonahua	364	292	Lionnet (1985)
listsamplesize	550	549	List (2014a)
liunewari	180	180	Liu *et al*. (2023)
liusinitic	203	202	Liú *et al*. (2007)
luangthongkumkaren	341	249	Luangthongkum (2019)
mannburmish	391	388	Mann (1998)
marrisonnaga	884	827	Marrison (1967)
mcelhanonhuon	140	140	McElhanon (1967)
mitterhoferbena	355	336	Mitterhofer (2013)
mixtecansubgrouping	240	204	Auderset *et al*. (2023)
naganorgyalrongic	1256	900	Nagano & Prins (2013)
nagarajakhasian	200	200	Nagaraja *et al*. (2013)
ndanda	200	200	Douanla Taffre (2021)
northeuralex	1016	954	Dellert *et al*. (2020)
northperulex	200	200	Barrientos *et al*. (Forthcoming)
oskolskayatungusic	254	254	Oskolskaya *et al*. (2022)

Dataset	Concepts	Concepticon	Source
othanieljen	300	285	Othaniel (2017)
papuanvoices	215	215	Sawaki *et al*. (2020)
peirosaustroasiatic	100	100	Peiros (1998)
pharaocoracholaztecan	100	100	Hansen (2020)
polyglottaafricana	319	263	Koelle (1854)
ratcliffearabic	100	100	Ratcliffe (2021)
robbeetstriangulation	253	253	Robbeets *et al*. (2021)
robinsonap	398	392	Robinson & Holton (2012)
saenkoromance	110	110	Saenko (2015)
sagartst	250	250	Sagart *et al*. (2019)
savelyevturkic	254	254	Savelyev & Robbeets (2020)
sawkatokaleya	201	199	Sawka *et al*. (2019)
seifartecheverriboran	408	351	Seifart & Echeverri (2015)
servamalagasy	207	207	Serva & Pasquini (2020)
sidwellbahnaric	200	200	Sidwell (2015)
sidwellvietic	116	116	Sidwell & Alves (2021)
simsrma	233	190	Sims (2020)
sohartmannchin	280	279	So-Hartmann (1988)
spagnolmaltese	1277	1275	Spagnol *et al*. (2024)
starostinkaren	110	110	Starostin & Krylov (2011)
starostinpie	110	110	Starostin (2005)
starostintujia	109	109	Starostin (2013)
suntb	1004	931	Sūn (1991)
syrjaenenuralic	173	173	Syrjänen *et al*. (2013)
tls	1052	650	Nurse & Phillipson (1975)
tolmiebritishcolumbia	210	200	Tolmie & Dawson (1884)
transnewguineaorg	1176	888	Greenhill (2015)
tryonsolomon	323	313	Tryon & Hackman (1983)
tuled	447	439	Ferraz Gerardi *et al*. (2022)
utoaztecan	121	121	Greenhill *et al*. (2023)
vanuatuvoices	426	426	Takau *et al*. (2024)
visserkalamang	855	854	Visser (2021)
walkerarawakan	100	100	Walker & Ribeiro (2011)
walworthpolynesian	210	210	Walworth (2018)
wangbai	458	404	Wang (2004)
wheelerutoaztecan	102	101	Wheeler & Whiteley (2015)
wichmannmixezoquean	110	110	Wichmann *et al.* (2011)
wold	1814	1458	Haspelmath & Tadmor (2009)

Dataset	Concepts	Concepticon	Source
yanglalo	1000	877	Yang (2011)
yangyi	1001	877	Yang (2022)
yuchinese	1307	1305	Yu & Wang (2021)
zgraggenmadang	333	310	Z'graggen (1980)
zhaobai	201	200	Zhao (2006)
zhivlovobugrian	110	110	Zhivlov (2011)
zhoubizic	346	301	Zhou (2020)

Limitations of Lexibank queries

Querying the Lexibank collection for various purposes offers scholars the possibility to dive into the resource and explore cross-linguistic patterns in various ways. Given the sheer size of Lexibank as well as the fact that it is impossible to verify and double-check every single data point that the resource offers, the results of these queries should, however, be taken with a certain care. Past research has shown that even manually curated datasets often show remarkable differences, even when coding for the same language varieties (Anderson *et al.*, 2023) As a result, it seems even more important to carry out additional tests to verify that the data retrieved from Lexibank is useful for a given analysis. In this context, what is considered as useful and what not, depends -- obviously -- from the analysis at hand. For example, as List (2018) has shown, rudimentary patterns of consonant clusters in the world's languages can be easily inferred from Lexibank and similar data collections. Whether a concrete analysis of these results, as presented by Yin *et al.* (2023), however, is meaningful and stands rigorous tests by follow-up studies, is different question that can only be answered in future research. With Lexibank we hope to facilitate linguists' access to cross-linguistic data in the domain of the lexicon. This does not mean, of course, that the collection also facilitates the interpretation of the results obtained with the data. We hope that future engagement with Lexibank will help in finding out where the data may be useful and where its limitations lie.

Beyond Lexibank

Our work on Lexibank has shown the benefits of standardizing data in order to allow for large-scale data aggregation. However, given that the purposes of individual data collections may well differ in practice, the work on Lexibank has also shown that a single repository cannot accommodate all needs at once. What we often need instead are targeted collections that concentrate on particular aspects of cross-linguistic data, thus offering targeted solutions for detailed problems. Examples are the fourth installment of the *Database of Cross-Linguistic Colexifications* (Tjuka *et al.*, 2025), initial efforts to derive benchmark datasets from the data in Lexibank 2 (Häuser & List, 2025), new approaches to deriving correspondence patterns from datasets with alignments and cognate annotation (Blum & List, 2023), merging and analysis of family-specific datasets Blum *et al.* (2024a), as well as targeted web-based applications that allow for a targeted inspection of retro-standardized data (Forkel & List, 2024). In all these cases, the typical analyses that scholars desire to carry out largely exceed the needs of what a single data collection such as Lexibank 2 can offer. As a result, we do not only want to encourage colleagues interested in Lexibank 2 to test out different ways to query the data in its aggregated form, but also hope that new, particularly targeted collections might emerge from the individual datasets that we standardized for Lexibank 2.

Ethics and consent

Ethical approval and consent were not required.

Data and software availability

Table 7 presents an overview of the individual datasets. The data is curated on GitHub (https://github.com/lexibank/lexibank-analysed/tree/v2.1) and released on Zenodo (Blum *et al.* (2025a), https://doi.org/10.5281/zenodo.15194559).

The queries are published in a tutorial on Codeberg (https://codeberg.org/lexibank/lexibank-queries/src/tag/v0.1).

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Author contributions

F.B. and J.-M.L. initiated the study. F.B. wrote the initial draft of the article. J.-M.L. wrote the initial code for the compilation of the Lexibank 2 dataset. F.B. extended the project code. J.E. and R.F. reviewed the project code. F.B. and J.-M.L. designed the SQL queries. J.-M.L. and R.F. provided technical support for the computation of features and integration of SQLite. F.B., C.B., and C.R. prepared the updated releases and Zenodo integration for the individual datasets. J.-M.L. supervised the project. All authors provided editorial work in the inclusion and maintenance of individual datasets. All authors contributed to the manuscript.

Acknowledgements

We thank all communities and researchers which agree to make their data publicly available.

References

Aaley UR, Bodt TA: **New data on Kusunda.** 2019.
Publisher Full Text

Abraham B, Sako K, Kinny E, *et al.*: **Sociolinguistic research among selected groups in Western Arunachal Pradesh: highlighting monpa.** SIL International, Dallas, 2005.

Allen B: **Bai dialect survey.** SIL International, Dallas, 2007.
Reference Source

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | **Publisher Full Text**

Anderson C, Tresoldi T, Chacon T, *et al.*: **A cross-linguistic database of phonetic transcription systems.** *Yearbook of the Poznan Linguistic Meeting.* 2018; **4**(1): 21–53.
Reference Source

Anderson C, Tresoldi T, Greenhill SJ, *et al.*: **Variation in phoneme inventories: quantifying the problem and improving comparability.** *J Lang Evol.* 2023; **8**(2): 149–168.
Publisher Full Text

Auderset S, Greehnhill SJ, DiCanio CT, *et al.*: **Subgrouping in a 'dialect continuum': a Bayesian phylogenetic analysis of the mixtecan language family.** *J Lang Evol.* 2023; **8**(1): 33–63.
Publisher Full Text

Backstrom PC, Radloff CF: **Languages of Northern areas, volume 2 of sociolinguistic survey of Northern Pakistan.** National Institute of Pakistan Studies, Islamabad, 1992.
Reference Source

Barlow R: **Lote wordlist: based on elicitation sessions with Leo Lang from Uvol village, recorded in Kokopo, East New Britain, November 2022.** *Zenodo.* Geneva, 2024.
Publisher Full Text

Barlow R, Killian D: **Tomoip wordlist: based on elicitation sessions with Simon Mangil from Milim village, recorded in Kokopo, East New Britain, October to November 2022.** Max Planck Institute for Evolutionary Anthropology, Leipzig, 2023.
Publisher Full Text

Barrientos C, Blum F, Ingunza A, *et al.*: **NorthPeruLex - a lexical dataset of small language families and isolates from Northern Peru.** Forthcoming.

Berry C, Berry K: **A survey of some West Papuan phylum languages.** *Workpapers in Indonesian Languages and Cultures.* 1987; **4**: 25–80.
Reference Source

Birchall J, Dunn M, Greenhill SJ: **A combined comparative and phylogenetic analysis of the chapacuran language family.** *Int J Am Linguist.* 2016; **82**(3): 255–284.
Publisher Full Text

Blum F, List JM: **Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists.** In: *Proceedings of the 5th workshop on Research in Computational Linguistic Typology and Multilingual NLP.* Dubrovnik, Croatia, Association for Computational Linguistics, 2023; 52–64.
Publisher Full Text

Blum F, Barrientos C, Ingunza A, *et al.*: **A phylolinguistic classification of the Quechua language family.** *Indiana.* 2023; **1**(40): 29–54.
Publisher Full Text

Blum F, Barrientos C, Ingunza A, *et al.*: **Cognate reflex prediction as hypothesis test for a genealogical relation between the Panoan and Takanan language families.** *Sci Rep.* 2024a; **14**(1): 30636.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Blum F, Barrientos C, Zariquiey R, *et al.*: **A comparative wordlist for investigating distant relations among languages in Lowland South America.** *Sci Data.* 2024b; **11**(1): 92.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Blum F, Paschen L, Forkel R, *et al.*: **Consonant lengthening marks the beginning of words across a diverse sample of languages.** *Nat Hum Behav.* 2024c; **8**(11): 2127–2138.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Blum F, Barrientos C, Englisch J, *et al.*: **Lexibank analysed.** 2025a.

Blum F, Herbold S, List JM: **From isolates to families: using neural networks for automated language affiliation.** 2025b.
Publisher Full Text

Bodt TA, List JM: **Reflex prediction. A case study of Western Kho-Bwa.** *Diachronica.* 2022; **39**(1): 1–38.
Publisher Full Text

Bowern C, Atkinson QD: **Computational phylogenetics of the internal structure of Pama-Nguyan.** *Language.* 2012; **88**: 817–845.
Reference Source

Bowern C, Epps P, Hill J, *et al.*: **Languages of hunter-gatherers and their neighbors.** [dataset, version from 2021-04-27], 2021.
https://huntergatherer.la.utexas.edu/

Bremer ND: **A sociolinguistic survey of six Berta speech varieties in Ethiopia.** SIL International, Addis Ababa, 2016.
Reference Source

Brid N, List JM, Messineo MC: **Las lenguas del chaco desde la perspectiva de la semántica léxica. Análisis preliminar de patrones léxicos compartidos en el dominio etnobiológico.** *LIAMES.* 2022; **22**: 1–21.
Publisher Full Text

Castro A: **Southern Sui: a Fourth Sui dialect.** *Journal of the Southeast Asian Linguistics Society.* 2011; **4**(2): 1–31.
Reference Source

Castro A, Hansen B: **Hongshui He Zhuang dialect intelligibility survey.** SIL International, Dallas, 2010.
Reference Source

Castro A, Crook B, Flaming R: **A sociolinguistic survey of Kua-nsi and related Yi varieties in Heqing county, Yunnan province, China.** SIL International, Heqing, 2010.
Reference Source

Chacon T: **A revised proposal of Proto-Tukanoan consonants and Tukanoan family classification.** *Int J Am Linguist.* 2014; **80**(3): 275–322.
Publisher Full Text

Chacon TC: **Annotated Swadesh wordlists for Northwest Arawakan languages.** 2022.

Chacon TC: **Arawakan and Tukanoan contacts in Northwest Amazonia prehistory.** *PAPIA.* 2017; **27**(2): 237–265.
Reference Source

Chacon TC, Gonçalves AG, da Silva LF: **A diversidade linguística aruák no alto rio negro em gravações da década de 1950 [the diversity of arawakan languages from the upper rio negro in recordings from the 1950s].** *Forma y Función.* 2019; **32**(2): 42–67.
Publisher Full Text

Chin AC: **The Gelong language in the multilingual hub of Hainan.** *Bulletin of Chinese Linguistics.* 2015; **8**(1): 140–156.
Publisher Full Text

Chén Q: **Miáoyáo yǔwén. Zhōngyāng Mínzú Dáxué**. [Central Institute of Minorities], Běijīng, 2012.
Reference Source

Clark ER: **A phonological analysis and comparison of two Kim Mun varieties in Laos and Vietnam**. PhD thesis, Payap University, Chiang Mai, 2008.
Reference Source

Constenla Umaña A: **¿Existe relación genealógica entre las lenguas misumalpas y las chibchenses?** *Estudios de Lingüística Chibcha.* 2005.
Reference Source

Davletshin A: **Proto-Uto-Aztecans on their way to the Proto-Aztecan homeland: linguistic evidence.** *J Lang Relatsh.* 2012; **8**(1): 75–92.
Publisher Full Text

Dàxué B: **Hànyǔ fāngyán cíhuì**. 1964.

de Carvalho FO: **A comparative reconstruction of Proto-Purus (Arawakan) segmental phonology.** *Int J Am Linguist.* 2021; **87**(1): 49–108.
Publisher Full Text

Deepadung S, Buakaw S, Rattanapitak A: **A lexical comparison of the Palaung dialects spoken in China, Myanmar, and Thailand.** *Mon-Khmer.* 2015; **44**: 19–38.
Publisher Full Text

Dellert J, Daneyko T, Münch A, *et al.*: **NorthEuraLex: a wide-coverage lexical database of Northern Eurasia.** *Lang Resour Eval.* 2020; **54**(1): 273–301.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Dhakal DN, List JM, Roberts SG: **A phylogenetic study of South-Western Tibetic.** *J Lang Evol.* 2024; **9**(1–2): **14–28.**
Publisher Full Text

Dolgopolsky AB: **Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concering the oldest relationships among the language families of Northern Eurasia].** *Voprosy Jazykoznanija.* 1964; **2**: 53–63.

Douanla Taffre AV: **Étude comparative du point de vue lexicostatistique et morphologique des dialectes du nda'nda'**. Ma thesis, Université de Yaounde, 2021.
Reference Source

Dunn M: **Indo-European lexical cognacy database (IELex)**. 2012.
Reference Source

Dunn M, Kruspe N, Burenhult N: **Time and place in the prehistory of the Aslian languages.** *Hum Biol.* 2013; **85**(1–3): 383–400.
PubMed Abstract | **Publisher Full Text**

Duong TH, Nguyen TQ, Nguyen VL: **The language of the La Chi people in Ban Diu Commune, Xin Man Distric, Ha Giang Province, Vietnam.** In: N. J. Enfield, Jack Sidnell, and Charles H. P. Zuckerman, editors, *Studies in the anthropology of language in mainland Southeast Asia.* University of Hawaii Press, Honolulu, 2020; 124–138.
Reference Source

Feleke TL: **Ethiosemitic languages: classifications and classification determinants.** *Ampersand.* 2021; **8**: 100074.
Publisher Full Text

Ferraz Gerardi F, Reichert S, Aragon C, *et al.*: **Tuled: tupían lexical database.** [dataset, version 0.12], 2022.
https://tular.clld.org/contributions/tuled

Forkel R: **The Cross-Linguistic Linked Data project**. In: Christian Chiarcos, John Philip McCrae, Petya Osenova, and Cristina Vertan, editors, *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing.* Reykjavik, ELRA, 2014; 61–66.
Reference Source

Forkel R, Bibiko HJ, Blum F, *et al.*: **PyLexibank. The Python Curation Libary for Lexibank [Software, Version 3.5.0]).** Max Planck Institute for Evolutionary Anthropology, Leipzig, 2025a.
Reference Source

Forkel R, Hammarström H: **Glottocodes: identifiers linking families, languages and dialects to comprehensive reference information.** *Semantic Web.* 2022; **13**(6): 917–924.
Publisher Full Text

Forkel R, List JM: **CLDFbench give your cross-linguistic data a lift**. In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation.* Luxembourg, European Language Resources Association (ELRA), 2020; 6997–7004.
Reference Source

Forkel R, List JM: **Cross-Linguistic Data Formats (CLDF): D'où venons nous? Que sommes nous? où allons nous?** In: *Digital humanities im deutschsprachigen Raum 2024*. *Zenodo*. 2024; 1–5.
Publisher Full Text

Forkel R, List JM, Greenhill SJ, *et al.*: **Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics.** *Sci Data.* 2018; **5**(1): 180205.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Forkel R, Rzymski C, Bank S: **PyCLDF [Software, Version 1.40.4])**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2025b.
Reference Source

Galucio AV, Meira S, Birchall J, *et al.*: **Genealogical relations and lexical distances within the Tupian linguistic family.** *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas.* 2015; **10**: 229–274.
Reference Source

Gao T: **Reconstruction and analysis of phylogenetic network on Tibeto-Burman languages in China.** *J Chin Linguist.* 2020; **48**(1): 257–293.
Reference Source

Gerardi FF, Reichert S, Aragon CC: **TuLeD (Tupían Lexical Database): introducing a database of a South American language family.** *Lang Resour Eval.* 2021; **55**(4)**: 997–1015.**
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gower R: **CSV on the Web**. Swirrl, Stirling, 2021.
Reference Source

Gravina R: **The phonology of Proto-Central Chadic. The reconstruction of the phonology and lexicon of Proto-Central Chadic, and the linguistic history of the Central Chadic languages.** PhD thesis, Leiden University, Leiden, 2014.
Reference Source

Greenhill SJ: **TransNewGuinea.org: an online database of new Guinea languages.** *PLoS One.* 2015; **10**(10): e0141563.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Greenhill SJ, Blust R, Gray RD: **The Austronesian basic vocabulary database: from bioinformatics to lexomics.** *Evol Bioinform Online.* 2008; **4**: 271–283.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Greenhill SJ, Gray RD: **Bantu basic vocabulary database**. 2015.
Reference Source

Greenhill SJ, Haynie HJ, Ross RM, *et al.*: **A recent northern origin for the Uto-Aztecan family.** *Language.* 2023; **99**(1): 81–107.
Publisher Full Text

Grollemund R, Branford S, Bostoen K, *et al.*: **Bantu expansion shows that habitat alters the route and pace of human dispersals.** *Proc Natl Acad Sci U S A.* 2015; **112**(43): 13296–13301.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hale A: **Clause, sentence, and discourse patterns in selected languages of Nepal IV: word lists**. Summer Institute of Linguistics and Tribhuvan University Press, Kathmandu, 1973.
Reference Source

Hammarström H, Forkel R, Haspelmath M, *et al.*: **Glottolog database**. Max-Planck Institute for Evolutionary Anthropology, Leipzig, 2024.
Publisher Full Text

Hansen MP: **¿Familia o vecinos? investigando la relación entre el proto-náhuatl y el proto-corachol [family or neighbors? investigating the relation between proto-náhuatl and proto-corachol]**. In: Rosa H. Yañez Rosales, editor, *Lenguas yutoaztecas: historia, estructuras y contacto lingüístico. Homenaje a Karen Dakin*. Universidad de Guadalajara, Guadelajara, 2020.
Reference Source

Hantgan A, Babiker H, List JM: **First steps towards the detection of contact layers in Bangime: a multi-disciplinary, computer-assisted approach [version 2; peer review: 2 approved].** *Open Res Eur.* 2022; **2**: 10.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hantgan A, List JM: **Bangime: secret language, Language isolate, Or language island? A computer-assisted case study.** *Papers in Historical Phonology.* 2022; **7**(1): 1–43.
Publisher Full Text

Hantgan A, List JM: **Creating a CLDF wordlist from heath *et al.*'s dogon comparative wordlist.** *Computer-Assisted Language Comparison in Practice.* 2023; **6**(2): 59–63.
Publisher Full Text

Haspelmath M, Tadmor U: **World Loanword Database**. 2009.
Reference Source

Hattori S: **Japanese dialects**. In: Henry M. Hoenigswald and Robert H. Langacre, editors, *Diachronic, Areal and typological linguistics*, number 11 in Current Trends in Linguistics. Mouton, The Hague and Paris, 1973; 368–400.

Heggarty P, Anderson C, Scarborough M: **Indo-European cognate relationships database (IE-CoR version 1.1)**. 2024.
Reference Source

Häuser L, List JM: **Lexibench: towards an improved collection of benchmark data for computational historical linguistics.** *Computer-Assisted Language Comparison in Practice.* 2025; **8**(1): 9–16.
Publisher Full Text

Hóu J: **Xiàndài hànyǔ fāngyán yīnkù**. CD-ROM, 2004.

Huber RQ, Reed RB: **Vocabulario comparativo: palabras selectas de lenguas indígenas de Colombia [Comparative vocabulary. Selected words from the indigeneous languages of Columbia]**. Asociatión Instituto Lingüístico de Verano, Santafé de Bogota, 1992.
Reference Source

International Phonetic Association IPA: **Handbook of the International Phonetic Association**. Cambridge University Press, Cambridge, 1999.
Reference Source

Ivani JK: **A first overview on Suansu, a Tibeto-Burman language from Northeastern India**. *Talk, Held at the 29th Conference of the Southeast Asian Linguistic Society*. 2019.

Jackson JC, Watts J, Henry TR, *et al.*: **Emotion semantics show both cultural variation and universal structure.** *Science.* 2019; **366**(6472): 1517–1522.
PubMed Abstract | **Publisher Full Text**

Johansson NE, Anikin A, Carling G, *et al.*: **The typology of sound symbolism: defining macro-concepts via their semantic and phonetic features.** *Linguist Typol.* 2020; **24**(2): 253–310.
Publisher Full Text

Joo I: **Phonosemantic biases found in Leipzig-Jakarta lists of 66 languages.** *Linguist Typol.* 2019; **24**(1): 1–12.
Publisher Full Text

Kamholz D, van Schie A, Verdizade A, *et al.*: **OUTOFPAPUA database: Lexicons of the West Papuan language area.** [Database, Version 1.0], 2024.
Reference Source

Kessler B: **The significance of word lists**. CSLI Publications, Stanford, 2001.
Reference Source

Key MR, Comrie B: **The intercontinental Dictionary Series**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2016a.
Reference Source

Key MR, Comrie B: **The Intercontinental dictionary series**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2016b.
Reference Source

Kleinewillinghöfer U: **Bikwin-Jen - comparative wordlist (Swadesh 100)**. 2015.
Reference Source

Koch-Grünberg T: **Betóya-Sprachen Nordwestbrasiliens und der angrenzenden Gebiete (fortsetzung).** *Anthropos.* 1914; **9**(5/6): 812–832.
Reference Source

Koelle SW: **Polyglotta Africana or comparative vocabulary of nearly three hundred words and phrases in more than one hundred distinct African languages**. Church Missionary House, London, 1854.
Reference Source

Kolipakam V, Jordan FM, Dunn M, *et al.*: **A Bayesian phylogenetic study of the Dravidian language family.** *R Soc Open Sci.* 2018; **5**(3): 171504.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kraft CH: **Chadic wordlists**. Dietrich Reimer, Berlin, 1981.
Reference Source

Lai Y, List JM: **Lexical data for the historical comparison of Rgyalrongic languages [version 2; peer review: 3 approved].** *Open Res Eur.* 2023; **3**: 99.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lama ZQF: **Subgrouping of Nisoic (Yi) languages: a study from the perspective of shared innovation and phylogenetic estimation**. PhD thesis, University of Texas at Arlington, 2012.
Reference Source

Language and Social Development Organization: **Chin dialect survey**. LSDO, 2019.

Lee S: **A sketch of language history in the Korean Peninsula.** *PLoS One.* 2015; **10**(5): e0128448.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lee MH: **Phonological features of Caijia that are notable from a diachronic**

perspective. *J Hist Linguis.* 2022; **13**(1): 82–114.
Publisher Full Text

Lee S, Hasegawa T: **Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages.** *Proc Biol Sci.* 2011; **278**(1725): 3662–3669.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lee S, Hasegawa T: **Evolution of the Ainu language in space and time.** *PLoS One.* 2013; **8**(4): e62243.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lieberherr I, Bodt TA: **Sub-grouping Kho-Bwa based on shared core vocabulary.** *Himalayan Linguistics.* 2017; **16**(2): 25–63.
Publisher Full Text

Lindsey KL: **Ende (Papua New Guinea) - IDS word list**. 2019.
Reference Source

Lionnet A: **Relaciones internas de la rama sonorense.** *Amerindia.* 1985; **10**: 1–30.
Reference Source

List JM: **Investigating the impact of sample size on cognate detection.** *Journal of Language Relationship.* 2014a; **11**: 91–101.
Publisher Full Text

List JM: **Sequence comparison in historical linguistics**. Düsseldorf University Press, Düsseldorf, 2014b.
Publisher Full Text

List JM: **Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction.** *J Lang Evol.* 2016; **1**(2): 119–136.
Publisher Full Text

List JM: **Inferring consonant clusters from CLICS data with LingPy.** *Computer-Assisted Language Comparison in Practice.* 2018; **1**(11): 2018.
Publisher Full Text

List JM: **Inference of partial colexifications from multilingual wordlists.** *Front Psychol.* 2023a; **14**: 1156540.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

List JM: **Evolutionary aspects of language change**. In: Agathe du Crest, Martina Valković, André Ariew, Hugh Desmond, Philippe Huneman, and Thomas Reydon, editors, *Evolutionary thinking across disciplines. Problems and perspectives in generalized Darwinism*. Springer, Berlin and New York, 2023b; 103–124.
Publisher Full Text

List JM, Cysouw M, Forkel R: **Concepticon - a resource for the linking of concept lists**. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Luxembourg, European Language Resources Association (ELRA), 2016a; 2393–2400.
Reference Source

List JM, Pathmanathan JS, Lopez P, *et al.*: **Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics.** *Biol Direct.* 2016b; **11**: 39.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

List JM, Walworth M, Greenhill SJ, *et al.*: **Sequence comparison in computational historical linguistics.** *J Lang Evol.* 2018; **3**(2): 130–144.
Publisher Full Text

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

List JM, Forkel R, Greenhill SJ, *et al.*: **Lexibank analysed.** *Zenodo*. Geneva, 2023.
Publisher Full Text

List JM, Anderson C, Tresoldi T, *et al.*: **CLTS. Cross-Linguistic Transcription Systems.** 2024.
Publisher Full Text

List JM, Tjuka A, Blum F, *et al.*: **CLLD concepticon 3.4.0.** Max Planck Institute for Evolutionary Anthropology, Leipzig, 2025.
Publisher Full Text

Liu Z, Jacques G, List JM: **Creating a standardized comparative wordlist of Newari varieties.** *Computer-Assisted Language Comparison in Practice.* 2023; **6**(2): 53–58.
Publisher Full Text

Library of Congress (LoC): **Library of congress recommended formats statement. 2024-2025.** Technical report, Library of Congress, Washington, 2024.
Reference Source

Liú L, Wáng H, Bǎi Y: **Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects].** Fènghuáng, Nánjīng, 2007.

Luangthongkum T: **A view on Proto-Karen phonology and lexicon.** *Journal of the Southeast Asian Linguistics Society.* 2019; **12**(1): i–lii.
Reference Source

Mann NW: **A phonological reconstruction of Proto Northern Burmic.** PhD thesis, The University of Texas, Arlington, 1998.
Reference Source

Marrison GE: **The classification of the Naga languages of North-East India.** phdthesis, School of Oriental and African Sciences, London, 1967.
Reference Source

McElhanon KA: **Preliminary observations on Huon Peninsula languages.** *Oceanic Linguistics.* 1967; **6**(1): 1–45.
Publisher Full Text

Mennecier P, Nerbonne J, Heyer E, *et al.*: **A central Asian language survey.** *Language Dynamics and Change.* 2016; **6**(1): 57–98.
Reference Source

Miller J, List JM: **Adding standardized transcriptions to Panoan and Tacanan languages in the Intercontinental dictionary series.** *Computer-Assisted Language Comparison in Practice.* 2024; **7**(2): 69–77.
Publisher Full Text

Mitterhofer B: **Lessons from a dialect survey of Bena: analyzing wordlists.** SIL International, 2013.
Reference Source

Moran S, Cysouw M: **The unicode cookbook for linguists: managing writing systems using orthography profiles.** Language Science Press, Berlin, 2018.
Publisher Full Text

Moran S, Forkel R, Heath J: **Dogon and Bangime linguistics.** 2016.
Reference Source

Morrison DA: **Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis.** *WIREs Data Mining and Knowledge Discovery.* 2014; **4**(4): 296–312.
Publisher Full Text

Nagano Y, Prins M: **rGyalrongic languages database.** National Museum of Ethnology, Osaka, 2013.
Reference Source

Nagaraja KS, Sidwell P, Greenhill S: **A lexicostatistical study of the Khasian language.** *MonKhmer Stud.* 2013; **42**: 1–11.
Reference Source

Nurse D, Phillipson G: **Tanzania language survey.** Department of Foreign Languages and Linguistics, University of Dar es Salaam, Dar es Salaam, 1975.
Reference Source

Oskolskaya S, Koile E, Robbeets M: **A Bayesian approach to the classification of Tungusic languages.** *Diachronica.* 2022; **39**(1): 128–158.
Publisher Full Text

Othaniel NK: **A phonological comparative study of the Jen language cluster**. PhD thesis, Jos: Theological College of Northern Nigeria, 2017.
Reference Source

Peiros I: **Comparative linguistics in Southeast Asia (Pacific linguistics)**. Pacific Linguistics. Research School of Pacific and Asian Studies, Australian National University, ISBN 0858834898,9780858834897, 1998.
Reference Source

Ratcliffe RR: **On calculating the reliability of the comparative method at long and medium distances: Afroasiatic comparative lexica as a test case.** *Journal of Historical Linguistics.* 2012; **2**(2): 239–281.
Publisher Full Text

Ratcliffe RR: **The glottometrics of Arabic. Quantifying linguistic diversity and correlating it with diachronic change.** *Language Dynamics and Change.* 2021; **11**(1): 1–29.
Publisher Full Text

Robbeets M, Bouckaert R, Conte M, *et al.*: **Triangulation supports agricultural spread of the Transeurasian languages.** *Nature.* 2021; **599**(7886): 616–621.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Robinson LC, Holton G: **Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon.** *Language Dynamics and Change.* 2012; **2**(2): 123–149.
Publisher Full Text

Rubehn A, List JM: **Partial colexifications improve concept embeddings.** *arXiv.* 2025; **2502**(09743): 1–15.
Publisher Full Text

Rzymski C, Tresoldi T, Greenhill SJ, *et al.*: **The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies.** *Sci Data.* 2020; **7**(1): 13.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Saenko M: **Annotated Swadesh wordlists for the Romance group (Indo-European family)**. 2015.
Reference Source

Sagart L, Jacques G, Lai Y, *et al.*: **Dated language phylogenies shed light on the ancestry of Sino-Tibetan.** *Proc Natl Acad Sci U S A.* 2019; **116**(21): 10317–10322.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Savelyev A, Robbeets M: **Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family.** *J Lang Evol.* 2020; **5**(1): 39–53.
Publisher Full Text

Sawaki Y, Lekeneny J, Arilaha A, *et al.*: **Word list "Papuan Voices"**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2020.

Sawka KS, Mbewe C, Josephat D, *et al.*: **Toka-Leya of Zambia: Southern Province: Livingstone, Kazungula, Zimba, and Kalomo Districts**. SIL International, 2019.
Reference Source

Seifart F, Echeverri JÁ: **Proto Bora-Muinan.** *LIAMES: Línguas Indígenas Americanas.* 2015; **15**(2): 279–311.
Publisher Full Text

Serva M, Pasquini M: **Dialects of Madagascar.** *PLoS One.* 2020; **15**(10): e0240170.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Shcherbakova O, Michaelis SM, Haynie HJ, *et al.*: **Societies of strangers do not speak less complex languages.** *Sci Adv.* 2023; **9**(33): eadf7704.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Sidwell P: **Austroasiatic dataset for phylogenetic analysis: 2015 version.** *Mon - Khmer Studies: a journal of Austroasiatic philology.* 2015; **44**: lxviii–ccclvii.
Publisher Full Text

Sidwell P, Alves M: **Vietic 116 item phylogenetic lexicon.** *Zenodo.* Geneva, 2021. The dataset was created for a paper provisionally entitled "The Vietic Languages: A Phylogenetic Analysis". The paper is submitted for journal publication and a version submitted for presentation at ICAAL9, November 2021.
Publisher Full Text

Sims NA: **Reconsidering the diachrony of tone in Rma.** *Journal of the Southeast Asian Linguistic Soceity.* 2020; **13**(1): 53–85.
Reference Source

Skirgård H, Haynie HJ, Blasi DE, *et al.*: **Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss.** *Sci Adv.* 2023; **9**(16): eadg6175.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

So-Hartmann H: **Notes on the Southern Chin languages.** *Linguistics of the Tibet-Burman Area.* 1988; **11**(2): 98–119.
Reference Source

Spagnol M, Comrie B, Bibiko HJ: **Maltese IDS wordlist by Michael Spagnol and Bernard Comrie.** *Zenodo.* 2024.
Publisher Full Text

Starostin GS: **Annotated Swadesh wordlists for the Tujia group (Sino-Tibetan family).** 2013.
Publisher Full Text

Starostin GS, Krylov P: **The global lexicostatistical database.** 2011.
Reference Source

Starostin SA: **Indo-European files n dbf/var.** 2005.
Reference Source

Sūn H: **Zàngmiǎnyǔ yǔyīn hé cíhuì.** 1991.

Syrjänen K, Honkola T, Korhonen K, *et al.*: **Shedding more light on language classification using basic vocabularies and phylogenetic methods: a case study of Uralic.** *Diachronica.* 2013; **30**(3): 323–352.
Publisher Full Text

Takau L, Walworth M, Shimelman A, *et al.*: **Vanuatu voices.** *Zenodo.* Geneva, 2024.
Publisher Full Text

Tjuka A: **Objects as human bodies: cross-linguistic colexifications between words for body parts and objects.** *Linguist Typol.* 2024; **28**(3): 379–418.
Publisher Full Text

Tjuka A, Forkel R, List JM: **Curating and extending data for language comparison in concepticon and NoRaRe [version 3; peer review: 2 approved].** *Open Res Eur.* 2023; **2**: 141.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Tjuka A, Forkel R, Rzymski C, *et al.*: **Advancing the database of cross-linguistic colexifications with new workflows and data.** *arXiv.* 2025; **2503**(11377): 1–14.
Publisher Full Text

Tolmie WF, Dawson GM: **Comparative vocabularies of the Indian tribes of British Columbia, with a map illustrating distribution.** Dawson Brothers, Montreal, 1884.
Publisher Full Text

Tryon DT, Hackman BD: **Solomon Islands languages: an internal classification.** Pacific Linguistics, Canberra, 1983.
Publisher Full Text

Turchin P, Peiros I, Gell-Mann M: **Analyzing genetic connections between languages by matching consonant classes.** *Journal of Language Relationship.* 2010; **3**: 117–126.
Reference Source

Urban M: **Asymmetries in overt marking and directionality in semantic change.** *Journal of Historical Linguistics.* 2011; **1**(1): 3–47.
Publisher Full Text

Visser E: **Kalamang - IDS word list.** 2021.

Walker RS, Ribeiro LA: **Bayesian phylogeography of the Arawak expansion in lowland South America.** *Proc Biol Sci.* 2011; **278**(1718): 2562–2567.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Walworth M: **Polynesian segmented data (Version 1).** [Data set], Max Planck Institute for the Science of Human History, Jena, 2018.
http://www.doi.org/10.5281/zenodo.1689909

Wang F: **Language contact and language comparison. The case of Bai.** PhD thesis, City University of Hong Kong, Hong Kong, 2004.
Reference Source

Wheeler WC, Whiteley PM: **Historical linguistics as a sequence optimization problem: the evolution and biogeography of Uto-Aztecan languages.** *Cladistics.* 2015; **31**(2): 113–125.
PubMed Abstract | **Publisher Full Text**

Wichmann S, Holman EW, Rama T, *et al.*: **Correlates of reticulation in linguistic phylogenies.** *Language Dynamics and Change.* 2011; **1**(2): 205–240.
Publisher Full Text

Yang C: **Lalo regional varieties: phylogeny, dialectometry and sociolinguistics.** PhD thesis, La Trobe University, Bundoora, 2011.
Publisher Full Text

Yang C: **The phonetic tone change *high > rising: evidence from the Ngwi dialect laboratory.** *Diachronica.* 2022; **39**(2): 226–267.
Publisher Full Text

Yin R, van de Weijer J, Round ER: **Frequent violation of the sonority sequencing principle in hundreds of languages: how often and by which sequences?** *Linguist Typol.* 2023; **27**(2): 381–403.
Publisher Full Text

Yu HJ, Wang Y: **Mandarin Chinese - IDS word list.** 2021.

Zhao Y: **Zhaozhuang Baiyu miaoxie yanjiu (Investigations of Zhaozhuang Bai).** PhD thesis, Zhongyang Minzu Daxue, 2006.

Zhivlov M: **Annotated Swadesh wordlists for the Ob-Ugrian group (Uralic family).** 2011.
Publisher Full Text

Zhou Y: **Proto-Bizic. A study of Tujia historical phonology.** Bachelor thesis, Stanford University, Stanford, 2020.
Publisher Full Text

Z'graggen JA: **A comparative word list of the Mabuso languages, Madang province, Papua New Guinea.** Australian National University, Canberra, 1980.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 02 July 2025

https://doi.org/10.21956/openreseurope.22414.r55975

© **2025 Espin-Riofrio C.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **César Espin-Riofrio**
Universidad de Guayaquil, Guayaquil, Ecuador

No further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Natural Language Processing

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.

Reviewer Report 02 July 2025

https://doi.org/10.21956/openreseurope.22414.r55976

© **2025 Cathcart C.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Chundra Cathcart**
University of Zurich, Zürich, Switzerland

My main comments have been adequately addressed.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Historical linguistics, evolutionary 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.

Reviewer Report 25 June 2025

https://doi.org/10.21956/openreseurope.22414.r55977

© 2025 Dediu D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Dan Dediu
University of Barcelona, Barcelona, Spain

it seems ok to me, except that I think Figure 1 is lacking a full legend (e.g., there's no description of what the colors and shapes mean)

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity; computational methods; statistics

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 03 June 2025

https://doi.org/10.21956/openreseurope.21876.r54517

© 2025 Espin-Riofrio C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

César Espin-Riofrio
[1] Universidad de Guayaquil, Guayaquil, Ecuador
[2] Universidad de Guayaquil, Guayaquil, Ecuador

The article clearly describes the standardization challenges in comparative linguistics and how Lexibank 2 addresses them, thereby justifying the expansion of the previous dataset (Lexibank 1). The methodology is based on well-established tools such as CLDF, PyLexibank, and CLTS. The use of SQLite for large-scale querying reflects a technically rigorous approach.
The workflow for data conversion and validation is extensively documented, including the use of specific software packages, the integration with reference catalogs, and the criteria for data inclusion.

The datasets are publicly available on Zenodo and GitHub. Moreover, the data are provided in the CLDF format, a widely adopted and thoroughly documented standard. Ready-to-use SQL queries, tutorials, and complete source files are also made available.
Several specific adjustments are recommended. First, it would be advisable to include a more systematic empirical validation of the phonetic and semantic queries, as they are currently presented as exploratory tools without a formal assessment of their performance. Additionally, it is recommended to clarify in greater detail the criteria used to exclude certain datasets, which would help make the curation process more transparent.
It would also be useful to reinforce the cautions regarding the interpretation of the results, particularly colexifications, to avoid misleading inferences about linguistic relationships. Finally, the inclusion of a brief comparison between versions of Lexibank is suggested, to help users better understand the changes introduced and their impact on the resource's use.

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?
Yes

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

Competing Interests: human or generated text

Reviewer Expertise: Natural Language Processing

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 17 Jun 2025

Frederic Blum

Several specific adjustments are recommended. First, it would be advisable to include a more systematic empirical validation of the phonetic and semantic queries, as they are currently presented as exploratory tools without a formal assessment of their p erformance. They are supposed to be exploratory tools, so this is fine. For the utility of those features, you could check out Blum et al. (2025, https://doi.org/10.48550/arXiv.2502.11688) , or studies using the CLICS3 database.
Additionally, it is recommended to clarify in greater detail the criteria used to exclude certain datasets, which would help make the curation process more transparent. We enhanced our argument for the inclusion of datasets that either a) did not represent sufficient data for inclusion (less than 80 mapped entries in a language), or b) for which no phonemic transcriptions could be provided.
It would also be useful to reinforce the cautions regarding the interpretation of the

results, particularly colexifications, to avoid misleading inferences about linguistic relationships. We do not see how the colexifications could be used for infering linguistic relationships between individual languages, since we do not provide any specific comparisons of languages, only of concepts. **Finally, the inclusion of a brief comparison between versions of Lexibank is suggested, to help users better understand the changes introduced and their impact on the resource's use.** Such a comparison is already part of the initial submission, see page 7 (Section "Comparing Lexibank 1 and Lexibank 2") and Table 1.

Competing Interests: No competing interests were disclosed.

Reviewer Report 03 June 2025

https://doi.org/10.21956/openreseurope.21876.r54153

© 2025 Cathcart C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chundra Cathcart
[1] University of Zurich, Zürich, Switzerland
[2] University of Zurich, Zürich, Switzerland

I think this paper and the release of Lexibank 2 are very timely and I recommend acceptance. Below, I give a précis with some feedback woven in, along with stray comments. The paper proceeds as follows:

The introduction gives some context on the use of lexical and grammatical databases in historical work. In general there is a great deal more that can be cited in terms of work using cognacy databases to infer phylogenies or carry out phylogenetic comparative methods, but this is also understandably not the focus of the paper. Some details on the historical development of Lexibank are provided as well.

The section "Materials and methods" describes the structure of Lexibank 2, steps taken to process and normalize the data, and describes the possibility of exploiting the data's relational structure for SQL queries. Data sets in lexibank are present in different collections depending on what features are available. The subsection "Workflow for data curation" details steps involving the removal of certain data sets (specifically, ones that cannot be converted to phonemic transcription —— I would have been curious to know what specific traits characterize resources that cannot be converted to phonemic transcription; this was not entirely clear to me although the excluded resources are listed in the online documentation). This subsection also mentions alternative ways of dealing with languages that are duplicated across datasets, and lists some precomputed phonological and semantic features (e.g., conversion of segments to sound classes, assignment of central concepts, etc.). The subsection "Queries for data exploration" outlines some possible data retrievals that can be undertaken using SQLite.

The section "Dataset validation" is essentially a Results section describing the outcomes of certain processing steps and SQL queries. The subsection "Benefits of standardization" details some quality checks, including heuristics used to find incorrectly converted segments. The subsection "Queries on pre-computed features" gives the results of three queries (similar word forms, cross-linguistic colexifications, and partial colexifications, the latter of which is challenging given the string matching needed, and best carried out using SQL rather than Python). The first query presents a worked example that involves finding all forms with T and R as C1 and C2 with connection to WOOD, which returns a number of reflexes of PIE *doru in Germanic and Slavic (a subset of which are given in Table 3). It would have been nice to actually use gold standard cognacy judgments (which if I am not mistaken should be available for IE languages) to get a sense of not only the precision but also the recall of this strategy in finding cognates (e.g., do we get hits for the IA, Hellenic, etc., reflexes of this PIE etymon?). And while the two lessons contrasted at the end of this example (p. 9: "From this experimental query...") are well taken (semantic permissiveness coupled with Dolgopolsky's sound class method is often criticized, e.g., https://doi.org/10.1075/jhl.00018.ceo), the wording here is confusing: TREE and WOOD are semantically closer than e.g. EARWAX, but they are still not identical.

I have to admit that I found it a bit disorienting to have to refer back to the previous section when reading the results described, but this may just be my personal preference.

All in all, the steps taken to normalize the data are well described. However, one message that did not always come through clearly in the text pertains to what researchers should be using these precomputed features for —— and what they should *not* be using them for. It is mentioned in passing that we should not uncritically take queries involving Dolgopolsky's method to represent cognates. But I found myself wondering for instance if we could take syllable complexity at face value (it is very easy to automatically extract syllable templates, but different languages may syllabify CVCCV differently) for use in analyses. Or are these features more geared towards some downstream application (e.g., for use in cognate detection)? An analogy is the use of byte-pair encoding in NLP: as a linguistic representation, it is pretty meaningless, but it improves performance in language modeling when used as a tokenization step in preprocessing. Some clearer guidance along these lines would likely be helpful to Lexibank 2 users, and it would not hurt to repeat any cautionary statements at various places in the paper (e.g., the conclusion).

Other issues:

Figure 1: I am not a fan of using pie charts to represent multiple variables on a map, as visibility is challenging. I also don't understand why the redundant feature "LexiCore" is included, since all doculects have the value 1. Removing this variable would improve visibility (via e.g. a 50/50 pie chart or color/shape aesthetic parameterization).

Some URLs are broken: https://codeberg.org/lexibank/lexibank-queries/v0.1 was not found, though https://codeberg.org/lexibank/lexibank-queries worked fine.

Table 7: weird diacritic in "Zǵraggen (1980)"

p. 18: "which is often used to represent the glide [i]" -> "glide [j]"?

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?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Historical linguistics, evolutionary 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 04 Jun 2025

Frederic Blum

Dear Chundra Cathcart, thanks for the quick and detailed review. We will elaborate an updated version soon and let you know once we have uploaded the final version of the article. On behalf of the authors, Frederic

Competing Interests: No competing interests were disclosed.

Author Response 17 Jun 2025

Frederic Blum

All in all, the steps taken to normalize the data are well described. However, one message that did not always come through clearly in the text pertains to what researchers should be using these precomputed features for —— and what they should *not* be using them for. It is mentioned in passing that we should not uncritically take queries involving Dolgopolsky's method to represent cognates. But I found myself wondering for instance if we could take syllable complexity at face value (it is very easy to automatically extract syllable templates, but different languages may syllabify CVCCV differently) for use in analyses. Or are these features more geared towards some downstream application (e.g., for use in cognate detection)? An analogy is the use of byte-pair encoding in NLP: as a linguistic representation, it is pretty meaningless, but it improves performance in language modeling when used as a tokenization step in preprocessing. Some clearer guidance along these lines would likely be helpful to Lexibank 2 users, and it would not hurt to repeat any cautionary statements at various places in the paper (e.g., the conclusion).
We have added a further paragraph on the limitations of such queries

Figure 1: I am not a fan of using pie charts to represent multiple variables on a map, as visibility is challenging. I also don't understand why the redundant feature "LexiCore" is included, since all doculects have the value 1. Removing this variable wo uld improve visibility (via e.g. a 50/50 pie chart or color/shape aesthetic parameterization).
We agree that the original figure was very cluttered. In agreement with reviewer 1 and reviewer 2, we created a new figure that indicates pertinence to ClicsCore via color, and to CogCore via shape (triangle). The LexiCore membership is superfluous in the figure, since it is a base condition for all Lexicank datasets. We have also adapted Figure 2 accordingly.

> Some URLs are broken: https://codeberg.org/lexibank/lexibank-queries/v0.1 was not found, though https://codeberg.org/lexibank/lexibank-queries worked fine.
> Table 7: weird diacritic in "Zǵraggen (1980)"
> p. 18: "which is often used to represent the glide [i]" -> "glide [j]"?
We have fixed these errors.

Competing Interests: No competing interests were disclosed.

Reviewer Report 03 June 2025

https://doi.org/10.21956/openreseurope.21876.r54158

© **2025 Dediu D.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Dan Dediu
[1] University of Barcelona, Barcelona, Spain
[2] University of Barcelona, Barcelona, Spain

I found the article very well written and clear, the examples useful, and there is no question that this is a potentially very useful extension of the previous version of Lexibank.
So, overall, I have no real comments nor questions, with the exception of something that I personally found a bit unclear in the paper: so, could the differences and similarities between Lexibank and CLICS be made clearer? I think this could be quite important for this new version of Lexibank.
Minor issues:
p4: "were be" -> "were"?

Figure 1: the map is way too busy to make any sense of it, especially for people with abnormal colour perception: I suggest breaking it down in at least 3 different maps (one figure with 3 panels or 3 separate figures), if possible -- that would also get rid of the very busy and hard-to-see circle sectors as well

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?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linguistic diversity; computational methods; statistics

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 04 Jun 2025

Frederic Blum

Dear Dan Dediu, thank you for the quick and insightful review. We will update the article accordingly and let you know once we have submitted the final version. On behalf of the authors, Frederic.

Competing Interests: No competing interests were disclosed.

Author Response 17 Jun 2025

Frederic Blum

p4: "were be" -> "were"? We have fixed this. **Figure 1: the map is way too busy to make any sense of it, especially for people with abnormal colour perception: I suggest breaking it down in at least 3 different maps (one figure with 3 panels or 3 separate figures), if possible -- that would also get rid of the very busy and hard-to-see circle sectors as well** We agree that the original figure was very cluttered. In agreement with reviewer 1 and reviewer 2, we created a new figure that indicates pertinence to ClicsCore via color, and to CogCore via shape (triangle). The LexiCore membership is superfluous in the figure, since it is a base condition for all Lexibank datasets. We have also adapted Figure 2 accordingly.

Competing Interests: No competing interests were disclosed.