Preprint, not peer-reviewed, under peer review. Passau: MCL Chair at the University of Passau.

Integration of Linguistic Legacy Data Collections through Digital Scholarly Editions: A Case Study on Vanuatu Languages

ARNE RUBEHN[1], TIHOMIR RANGELOV[2], LUCA CIUCCI[1], JOHN BURGESS[3], RICCARDO ROST[1], JOHANN-MATTIS LIST[1]
[1]CHAIR OF MULTILINGUAL COMPUTATIONAL LINGUISTICS, UNIVERSITY OF PASSAU, GERMANY
[2]MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY, LEIPZIG, GERMANY
[3]UNIVERSITY OF WESTERN AUSTRALIA, PERTH, AUSTRALIA

Abstract

The past two decades have witnessed a substantial increase in computational methods for investigating language diversity and history. The amount of digital data in comparative linguistics, however, is still lagging behind, with existing digitization efforts still mostly relying on the cumbersome labor of typing off data manually. Available Optical Character Recognition tools for automating this task have received relatively little attention in digitizing legacy data in linguistics, even though they are routinely used in other disciplines. At the same time, the editorial work must go beyond plain digitalization to add various layers of analysis and standardization, while recording the full provenance of each data point. We present an efficient and transparent workflow for digitizing legacy data in comparative linguistics and integrating it with larger data collections. As a result, we present a digital scholarly edition of lexical data for Vanuatu languages published by Darrell T. Tryon in 1976.

Keywords: legacy materials; FAIR principles; Vanuatu languages; comparative wordlists; historical linguistics; language typology; digital scholarly editing.

1. Introduction

Although the amount of digitally available cross-linguistic datasets has been constantly increasing during the past two decades, the amount of data that has *not* yet been digitized still largely exceeds the amount of digital data. The past years have seen several efforts to digitize and aggregate cross-linguistic data, both in the domains of grammar (Auderset 2020; Skirgård et al. 2023) and lexicon (Dellert et al. 2020; Blum et al. 2025). The improvements made in this regard, witnessed both by the increase in freely available data collections and articles devoted to their exploration, reflect a *quantitative turn* (Geisler & List 2022, see also Levinson & Evans 2010) that has greatly changed how cross-linguistic studies in historical linguistics and linguistic typology are carried out today.

What we have *not* seen so far are attempts to improve the way in which cross-linguistic data are harvested from books, articles, fieldwork notes, or other forms of gray literature. In the field of evolutionary biology, the quantitative turn dates back to the 1970s (Hull 1988: 121). However, it has not been until recently, when new techniques for the sequencing of genome data — known as "next-generation sequencing" (Liu et al. 2012) — were widely introduced, that we have seen an explosion of genetic data (see, e.g., Morel et al. 2021 for coronavirus sequences). Considering that large cross-linguistic datasets published in the past two decades have almost exclusively been built on manual labor, we think that an improved framework for the digitization of documents providing cross-linguistic information could lead to a similar explosion of data in historical linguistics and linguistic typology.

Legacy materials have traditionally received little attention, although they represent the only available data sources for many linguistic varieties. For this reason, they may play an essential role in language reclamation and revitalization, offer otherwise lost cultural information, or even be used for juridical purposes such as land claims. Even when a language still has speakers, the historical documentation may inform fieldwork, complement present-day data, and contribute to diachronic analysis. However, historical documents typically present interpretative challenges per se, since we only have a limited knowledge of the context in which they were produced (on legacy materials, see Austin 2017; Ciucci 2021; Dobrin & Schwartz 2021; Lahaussois 2025). The attempt to recover or reconstruct the missing context is therefore the task of a scholarly digital edition in order to "explain what is not evident to the present-day reader" and "bridge a distance in time" (Sahle 2016: 26), also with documents that, like Tryon's (1976) wordlists on Vanuatu languages, are relatively close to us but originated in a different linguistic framework and, for this reason, are legacy materials.

In this study, we will outline some core principles of improving digitization techniques in comparative linguistics. We present a digital workflow implementing those principles and illustrate its benefits in practice. Using this workflow, we present a *digital scholarly edition* of the lexical data presented in Tryon's (1976) *New Hebrides languages* survey. A digital scholarly edition[1] is "the critical representation of historical documents" (Sahle 2016: 23) and differs from a digitized edition in that it is "guided by a digital paradigm in […] theory, method and practice" (Sahle 2016: 28) so that it is "not printable without a major loss of content and functionality" (Sahle 2016: 27). In Sahle's foundational definition, the adjective "critical" refers to necessary editorial work, i.e. textual criticism, whereby all textual interventions are explicitly documented and transparently presented, but also to the addition of layers of interpretation that vary according to the questions and goals of the editor's research

[1] The notions of *scholarly digital edition* and *digital scholarly edition* are used mostly synonymously in the literature. Throughout this study, we use the latter term, without intending a distinction from the former term.

program and discipline. The effective implementation of FAIR principles, in the sense of Wilkinson et al. (2016), is a crucial issue in digital scholarly editions (Gengnagel et al. 2023; see also Rosselli del Turco 2023). While the discussion on digital scholarly editions usually concerns documents of literary or historical interest, the workflow that we propose here can be seen as an attempt to apply this concept to the context of comparative linguistics, taking inspiration from a theoretical reflection that developed in the fields of digital humanities, philology, literary studies and history.

2. Background

2.1. OCR-Assisted Digitization of Legacy Data

Optical Character Recognition (OCR) is a technique that automatically extracts text from image data and converts it into a machine-readable format. Over the past decades, the number of historical documents (in the widest sense, every document that is not born-digital) that have been binarized (e.g. via scans) has exploded. Consequently, OCR tools have received ever increasing attention in order to extract the text from these documents, resulting in large collections of digitized and transcribed historical documents, like the Google Ngram Corpus (Michel et al. 2011). Blanke et al. (2012: 2) have foreshadowed that this development would lead to a "fundamental shift for humanities research" with an increasing focus on digitizing historical documents, both for preservation and computational analyses.

With the substantial advancements over the past years in the field of machine learning, particularly deep learning, it should not come as a surprise that available tools for OCR have drastically increased in both quantity and quality (for a technical survey, see Memon et al. 2020). Researchers of widely different disciplines have since used OCR tools for streamlining the digitization of legacy data, ranging from biodiversity records (Rehbein et al. 2025) over historical climate protocols (Vercruysse et al. 2025) to clinical patient reports (Paulsen et al. 2020). Yet, this avenue has barely been systematically explored for recovering linguistic data. Comparative wordlists used in historical linguistics and language documentation are ideal application cases, since they already follow a relatively clear structure. At the same time, they present a data type that is often available not only for well-documented, but specifically also for under-resourced language varieties. Therefore, OCR-assisted workflows offer invaluable potential in digitizing large amounts of linguistic data, especially those describing understudied Indigenous and minority languages.

2.2. Digital Scholarly Editions of Linguistic Legacy Data

The successful application of OCR tools, as discussed in the previous section, results in an exact, but machine-readable copy of the original document. We call this representation the *digital transcript*. While this is already useful in its own right, we argue that it is necessary to go beyond this step in comparative linguistics. Every dataset has its own idiosyncrasies, such as different transcription systems, language names or glosses. In order to integrate a dataset and make it comparable with data from other sources, an additional layer of interpretation is needed. This is enabled in the creation of a digital scholarly edition that clearly separates the transcription from the interpretation.

A digital scholarly edition requires attention to the philological issues of the document. While philology can have different meanings, here it is understood stricto sensu as textual criticism, i.e. as the discipline that has at its core the scientific study of the transmission of textual information aimed at reconstructing its correct version according to the will of the author in a critical edition. While this is the ultimate goal of philology, not all scholarly digital editions are necessarily critical editions (Michelone 2021: 26-29) although resorting to the philologist's methodological toolkit is usually necessary. In the specific case of comparative linguistics, historical documents containing data differ in their typology (wordlists, grammars, dictionaries, narratives, etc.) and in their textual issues.

Textual criticism originated to establish the text of works whose original version is lost and has to be reconstructed (the so-called "philology of the copy"; on this philological method and its historical development, see Maas 1960; Reynolds & Wilson 1968; Timpanaro 2005, among others). This is typically the case for Classical and medieval texts, with which philology has traditionally been associated, but it occurs less frequently for documents concerning Indigenous and minoritized languages, which are often preserved in their original manuscripts. Their textual issues are the object of "authorial philology" (Italia & Raboni 2021).

The differences between these two subfields of philology can be illustrated with the documentation of Old Zamuco (Zamucoan; Bolivia, Paraguay) by the Jesuit Ignace Chomé (1696-1768). He wrote an Old Zamuco grammar, of which only a copy made by other missionaries is available (Ciucci 2018: 441). The content of the manuscript was published by Lussagnet (Chomé 1958) without any change, i.e. in a "diplomatic edition". However, in a critical edition a philologist has to amend the mistakes that normally occur when somebody is writing, plus those added by copyists (on such mistakes and how to detect them, see, e.g. Stussi 2002: 100-105), which requires familiarity with the languages of the document (in this case, Old Zamuco and Spanish) and the author's way of writing (technically, *usus scribendi*). Chomé also wrote an Old Zamuco dictionary whose original manuscript was recently rediscovered (Ciucci 2018). A critical edition of this document (Ciucci in prep.) requires particular care because the manuscript does not contain a final, polished version of the text. Chomé continuously checked his data, made corrections and cuts, and added new entries and examples. Following the methodology of authorial philology, the editor here, apart from amending mistakes, also has to identify and document the many authorial variants in order to illustrate the evolution of the text while reconstructing its "correct" version. While textual criticism is obviously necessary to ensure the quality of data that are analyzed in comparative linguistics, the interpretation of the authorial variants may also provide valuable information on the language.

In all cases, the textual issues and editorial criteria must be explained in a “note to the text”, while the relevant interventions on each page are illustrated in a “critical apparatus” (on this concept, see, e.g., Stussi 2002: 154-157; on its adaptation to the digital paradigm, see Apollon et al. 2014: 81-113 and Fischer 2019). The amount of philological work varies according to the type of document. For instance, the textual issues of printed books, which are the object of “textual bibliography” (see Villari 2014), are typically less than those of manuscripts. However, for centuries the hand-operated printing press allowed authors and typographers to alter the text during the printing process, which resulted in different versions of the same work. Applied to historical linguistic documents, textual bibliography could reveal numerous variants in early printed linguistic works produced by missionaries.

While textual criticism plays a limited role in the edition of Tryon’s (1976) wordlists, it is nevertheless necessary to ensure the correctness of the data. The book is the photo-offset of a typewritten manuscript, so, apart from issues that may be present in any textual document, there are also those related to the use of the typewriter or the printing process. For instance, a typist may confuse similar phonetic characters, some characters may be worn or their edges may be blurred, which may create ambiguities. The philological work applied to Tryon’s (1976) wordlists is illustrated by several examples in § 4, which also demonstrate other types of editorial intervention.

As pointed out by Villavicencio (2009: 279-281), there is a lack of philological tradition for the publication of documents in Indigenous languages, and most of the published works are diplomatic editions. While there are some critical or scholarly editions (see for instance Thun et al. 2015 and Ringmacher 2022 for Old Guaraní, Tupian), they are usually not digital-born editions, and their circulation is limited to specialists of a given language family or geographical area. A comparative reflection on the textual criticism involved in the publication of historical documents in Indigenous and minoritized languages is still missing. At the same time, the catalogue of digital editions by Franzini (2012) and that of scholarly digital editions by Sahle (2020) report a very low number of editions in non-Western languages. While this is partly due to a data bias pointed out by these authors (Kurzmeier et al. 2024), it also reflects the rarity of scholarly digital editions of documents in Indigenous and minoritized languages, for which our workflow was conceived. An increase in findable and interoperable scholarly digital editions will not only contribute to linguistics but will also allow us to compare the philological practices attested across different Indigenous and minoritized languages.

2.3. Standardization and Integration of Linguistic Legacy Data

A scholarly edition contextualizes and annotates the information provided in documents in a transparent way to allow interested researchers to interact with that information for their own purposes. Since linguistic legacy data does not exclusively deal with texts but often involves much more structured information on languages, ranging from dictionaries over wordlists to grammatic descriptions, a further step of annotation can be applied to *integrate* legacy data with existing cross-linguistic resources by *standardizing* documents beyond the level of traditional scholarly editions.

In the context of cross-linguistic datasets, this would mean that major objects of linguistic inquiry, such as *languages*, *concepts*, and *transcriptions* in comparative wordlists, are represented in such a way that a comparison with other published resources is facilitated not only for humans, but also for computational applications.

With the Cross-Linguistic Data Formats initiative (CLDF, Forkel et al. 2018 Forkel et al. 2018, https://cldf.clld.org), first attempts have been made to establish standard formats that can be used to archive and exchange cross-linguistic data. In order to address the problem of representing major linguistic objects in a unified way, the standard formats proposed by the CLDF initiative recommend the use of dedicated reference catalogs that represent specific metadata collections in which typical linguistic constructs are defined. Languages, for example, can be identified with the help of the Glottolog reference catalog (Hammarström et al. 2025, https://glottolog.org). One can use the standardized concept sets defined by the Concepticon project (List et al. 2025, https://concepticon.clld.org) for the handling of concepts in comparative wordlists, while, for the handling of phonetic transcriptions, the Cross-Linguistic Transcription Systems project (CLTS, List et al. 2024b, https://clts.clld.org) offers a standardized transcription system that is fully compatible with the International Phonetic Alphabet but has the advantage of being fully computer-readable (Anderson et al. 2018) and also easily extendable with additional metadata (Rubehn et al. 2024).

A specific advantage of CLDF as standard representation of linguistic legacy data is that this format facilitates the creation of interactive CLLD databases. CLLD is a computational framework for cross-linguistic data (Forkel 2014, https://clld.org) underlying many cross-linguistic databases that have been published in the last decade. CLLD websites allow users to browse linguistic data in multiple ways, offering not only a common look-and-feel, but also a direct integration across individual databases through linked data technologies.

While CLDF as a data representation format and CLLD as a data presentation framework were predominantly designed to facilitate the handling of cross-linguistic data in contemporary data collections and documentation projects, they have turned out to be applicable to linguistic legacy data as well. Thus, Geisler et al. (2021) illustrate how CLDF and CLLD can be used to represent and integrate dialect data of the French-speaking region of Switzerland, collected in the early 20th century by Gauchat et al. (1925), with the help of a new digital dialect atlas (https://tppsr.clld.org). Forkel et al. (2024a) present two examples in which CLDF and CLLD were used for the retrostandardization of linguistic legacy data from Africa (*Polyglotta Africana* by Koelle 1854, https://polyglottaafricana.clld.org) and India (*Linguistic Survey of India* by Grierson 1928, https://lsi.clld.org). Pulini & List (2025) illustrate how CLDF can be used to create a digital scholarly edition of Ancient Chinese texts written on Bamboo slips (https://cmzz.digling.org). These digital editions of linguistic legacy data also illustrate the added value of standardization. Thus, while none of the works by Gauchat et al. (1925), Grierson (1928) and Koelle (1854) offers individual phoneme inventories for the dialect and language varieties they report, the standardization through CLDF makes it easy to generate this information directly from the data and present it interactively in the web interface.

2.4. Summary

The potential of OCR tools and digital scholarly editions has received little attention in the field of comparative linguistics so far. On the one hand, OCR-assisted digitization workflows offer a significant speedup in extracting texts from legacy documents. On the other hand, digital scholarly editions combine two main advantages that benefit from each other. Indeed, digital editions enable the scalability of large amounts of data, relationability, interoperability, multimediality and user interaction. These advantages set them apart from print editions (Buzzoni 2016:59-60) and ensure the creation of FAIR data (Wilkinson et al. 2016). Scholarly editions, in turn, allow scholars to critically discuss and edit a source document, adding historical context, interpretations and corrections. (Sahle 2016: 22) argues that “the notion of scholarly editing should not be restricted to literary texts but has to cover all cultural artefacts from the past that need critical examination”, thus including legacy data in comparative linguistics. This valuable combination of digital and scholarly editing has not yet been widely applied to legacy linguistic data. If best practices are followed, a digital scholarly edition “furnishes the reader with all the documentation necessary to evaluate it and to produce another, maybe different edition that is nevertheless based on the same material” (Buzzoni 2016: 60). It is therefore imperative that all steps involved in the creation of a digital scholarly edition are transparently documented and reproducible.

In the following, we illustrate these principles by discussing the digitization process of legacy data on Vanuatu languages by Tryon (1976). We present a digital scholarly edition of the wordlists in that document that is fully integrated with large collections of comparative lexical data.

3. Materials and Methods

3.1. Tryon’s “New Hebrides Languages”

The Republic of Vanuatu, a South Pacific small-island nation, known as the New Hebrides before gaining independence in 1980, is the country with the most languages per capita in the world (Crowley 2000: 50). We know this to a large extent owing to Darrell Tryon’s seminal 1976 work *New Hebrides languages: An internal classification*, which, for the first time systematically catalogued the linguistic diversity of Vanuatu. For his study, Tryon and his collaborators collected data from many communities throughout the archipelago, including some of its most remote corners, and performed quantitative and qualitative analysis aiming at establishing the number of languages and forming subgrouping hypotheses. In the study, Tryon presented 179 lists with up to 292 basic vocabulary lexical items[2], which are presented as tables in an appendix, with each page featuring four concepts (columns) and up to 41 varieties (rows).

Based on these lexical data, Tryon performed a qualitative historical linguistic analysis to identify regular sound changes and a quantitative lexicostatistical analysis, proposing subgroupings for the surveyed varieties based on shared cognacy rates. As a result, Tryon's study established 105 distinct languages, demonstrating for the first time the incredible linguistic diversity of Vanuatu. Previously, no more than 37 languages were known to be spoken on the archipelago (cf. Capell 1962). Such is the significance of Tryon's (1976) study that even 50 years after its publication, for dozens of Vanuatu languages, these lists remain virtually the only primary data available (Rangelov et al. 2025).

Unsurprisingly, Tryon's (1976) data have been the basis of many studies on the languages of Vanuatu. However, to our knowledge, these data in their entirety have so far been available only in the form of hard copies and in PDF form. In this study, we present a digital scholarly edition of the lexical data, showcasing tools to create digital scholarly editions of linguistic data and their benefits.

3.2. Digitization Workflow

For the digitization of Tryon's (1976) data, we followed a workflow including multiple representation layers, which in turn are generated by separate operations using different tools and techniques. The general workflow is illustrated in Figure 1, outlining the representation layers, operations and techniques used in our digitization workflow.

The workflow departs from a scan of the original document. The first operation is the *transcription* of this original document, resulting in a *digital transcript* that contains the textual data from the original document in a machine-readable format. We use the OCR software Transkribus (Kahle et al. 2017) to efficiently extract the text from the document scans. The digital transcript is supposed to reflect all textual data exactly as they are found in the original document. This is a diplomatic transcription from the perspective of textual criticism since there is no editorial intervention at this stage.

[2] The original questionnaire spanned 309 concepts. Some of these items were removed since they turned out to be difficult to elicit reliably, or since they would lead to a duplication of certain etyma due to common polysemies.

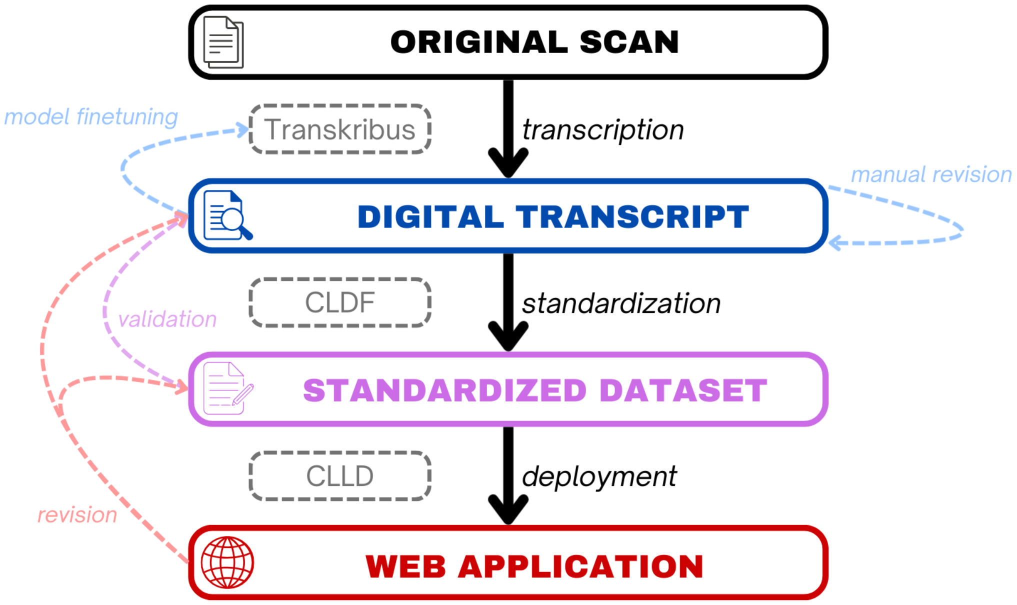


Figure 1: Illustration of the proposed digitization workflow.

The second operation is the *standardization* of the data, requiring a deeper *interpretation* of the original document and enabling the *integration* of the standardized data with other datasets from different sources. Through the standardization, we create a *standardized dataset* that ensures that the data is findable, accessible, interoperable and reusable (Wilkinson et al. 2016). The standardization process adheres to CLDF (Forkel et al. 2018) and is implemented with the help of the CLDFBench package (Forkel & List 2020) and the PyLexibank plugin (Forkel et al. 2024b). This step also serves as formal validation for the transcription, since unexpected characters or character sequences are automatically flagged, indicating potential problems in the transcription or the original document.

The third and final operation is the *deployment* of the digital scholarly edition as a *web application*. This is achieved using the CLLD framework (Forkel 2014) that offers convenient routines to turn CLDF datasets into interactive databases on the web. The great advantage of a web application is that data can be displayed interactively, allowing users to filter, sort, or search data. The web application also visualizes some additional information, such as the geographic distribution or the sound inventories of the included varieties. All previous layers of representation are displayed on the web app: for each form, the original scan, the digital transcription, and the standardized interpretation is provided. This ensures that the full provenance of all data is transparently recorded and displayed.

As can be seen, this multi-layered workflow allows for a clear distinction between the *transcription* and the *interpretation* of the original documents, which are represented separately in the *digital transcript* and the *standardized dataset*, respectively. Each representation is included in the digital scholarly edition and allows for the revision of previous representation layers, as indicated by the backward arrows in Figure 1. Keeping humans in the loop (cf. Vercruysse et al. 2025) ensures a high quality of data across all representation levels. In the following, we discuss the details of every operation, its resulting representation, and the tools employed.

3.3. Digital Transcription with Transkribus

The first operation of our workflow, outlined above, is the transcription of the original document. By that, we mean extracting textual data from scans of the original document and representing these data in a machine-readable way. The resulting machine-readable representation of the original document is the digital transcript.

For the transcription, we used the web-based OCR software Transkribus (Kahle et al. 2017), which offers straightforward means to use and fine-tune pre-trained models, alongside other convenient functionalities for digitizing historical documents. In particular, we employed a table recognition module before running the actual transcription. This has two main advantages: First, it substantially improves the detection of relevant text regions; second, it allows us to record the provenance of each lexical form in maximal detail, as we can pinpoint exactly *where* a certain form was found on the respective page. This enables us to include the relevant snippets in the web application, as we show in § 3.5 and discuss in § 4.4.

For the text extraction, we used a custom OCR model derived from the “Transkribus Print M1” model. Fine-tuning the OCR model was necessary, since deep neural OCR models implicitly feature a language model. This means that any model used off-the-shelf would favor character sequences found in the training data, i.e. n-grams that recur in the languages the model was trained on. Since our data features lexical items from 179 different low-resource varieties (which surely were not included in the training data of any available OCR model), we needed to fine-tune the base model to make it more suitable for the data at hand. For this purpose, we manually transcribed the first 21 pages of the wordlist. 19 of these pages were used as training data for fine-tuning the model, with the remaining two pages serving as validation data. In this initial training run, we achieved a Character Error Rate (CER) of 1.08%, giving us a reasonably reliable OCR model as a strong base to proceed with transcribing the wordlist in a mostly automated fashion.

For the remainder of the wordlist, we proceeded in batches of constantly increasing size. Each batch was first transcribed automatically and then corrected by a human annotator. The corrected batch was added to the training data, on which the OCR model was fine-tuned further. This created a constant feedback loop that iteratively improved the model, which in turn decreased the required correction efforts for the annotators. The final model, trained on 205 pages and validated on 22 pages (of 365 pages in total), achieved a CER of 0.54%, leaving only very few errors that needed to be corrected manually. This loop between automated transcriptions and manual corrections warranted a significant speed-up compared to simply “typing it off” while ensuring a high transcription quality.

As a final step, the digital transcript was exported from Transkribus in the software's idiosyncratic XML format. To make the digital transcript more human-readable, the relevant information was extracted from the XML files and represented in a tabular format. Here, each entry in the original document is represented as a triplet of form, variety and concept, accompanied by metainformation indicating where this entry can be found in the original document. Following the core idea that the digital transcript should be a completely faithful, but digital representation of the original document, all data points in this digital transcript appear exactly as they appear in the source document.

3.4. Standardization and Integration with CLDF

The next operation departs from the digital transcript and creates a digital scholarly edition by employing computer-assisted workflows for standardization of cross-linguistic data. This standardization ensures that the resulting edition of the data conforms to the CLDF specifications (Forkel et al. 2018) which enable the aggregation of cross-linguistic data from multiple different sources, as showcased by the Lexibank collection (Blum et al. 2025). To achieve comparability of data between different sources, forms are transcribed following a strict implementation of the International Phonetic Alphabet (IPA 1999) as specified by the Cross-Linguistic Transcription Systems (List & Anderson & Tresoldi & Rzymski & et al. 2024). Concepts and varieties are linked to the respective reference catalogs Concepticon (List et al. 2025) and Glottolog (Hammarström et al. 2025), ensuring that they can be compared across sources even if different glosses or language names were used.

We used the Python library PyLexibank (Forkel et al. 2024b) to create a CLDF edition from the data in the digital transcript. Languages and concepts were both mapped to their respective reference catalogs in a computer-assisted fashion: For languages, an initial mapping was generated by referencing the Austronesian Basic Vocabulary Database (Greenhill et al. 2008) for which a CLDF version already exists; for concepts, an initial mapping was generated automatically based on glosses found in other datasets using the PyConcepticon library (Forkel et al. 2019). In both cases, those automatically generated mappings were manually refined and corrected. The subgrouping hypotheses proposed by (Tryon 1976: 87-93) were also included in the varieties' metadata alongside the information of the respective data collector.

Forms had to be normalized as well to ensure that the phonetic transcriptions in the digital edition conform to the IPA. While Tryon did use a mostly regular transcription system that closely resembles the IPA, there were some deviations — for instance, Tryon writes <y> to indicate the palatal glide [j]. PyLexibank supports orthography profiles (Moran & Cysouw 2018) that allow users to conveniently define mappings between graphemes and corresponding IPA symbols, which in turn are applied automatically in the CLDF conversion process.

The inclusion of such an orthography profile, however, has another great advantage apart from simply standardizing phonetic transcriptions. The PyLexibank implementation requires all segments to be explicitly defined, even seemingly obvious ones, like <n> corresponding to [n]. Those segments that are not defined are marked as *unknown* in the CLDF edition. While this might seem unnecessarily cumbersome at first glance, it actually serves to check whether the digital transcript only contains those graphemes that are expected. Indeed, unexpected graphemes point towards potentially problematic transcriptions that therefore can be explicitly double-checked. Based on whether the initial transcription is correct or not, either the transcription needs to be corrected, or the orthography profile needs to be updated to handle the highlighted grapheme accordingly. In practice, this resulted in an iterative process, where problematic transcriptions were highlighted based on their incompatibility with the current orthography profile. These transcriptions were manually checked and corrected if needed, and, based on the updated digital transcript, the orthography profile was adjusted. With this workflow, we were able to weed out quite some transcription errors that otherwise might have slipped through the cracks — including very subtle ones, like instances where the OCR model transcribed wrong characters that were visually very similar to their targets: e.g., <ß> instead of <β> or <æ> instead of <æ>.

In other instances, this validation technique pointed us towards potential errors or inconsistencies in the original document. Since the scholarly digital edition requires a certain degree of interpretation, we could also amend obvious issues with the original data at this stage. Examples of this are discussed in § 4.2. The respective forms were corrected on a case-by-case basis during the standardization workflow. The digital transcript remained untouched by such corrections, therefore still representing the original document faithfully.

By following the above steps, we obtained a scholarly digital edition of the source data that can be integrated with numerous other multilingual wordlists in the Lexibank collection. Following the principle that a correct normalization is an annotation and does not change the primary data (Hirschmann 2019: 29), this digital edition was derived from the digital transcript in a fully transparent and reproducible way. Retaining the reference to the original document and the digital transcript for each form allows users to clearly distinguish choices made in transcription from those made in interpretation.

3.5. Deployment of the Digital Scholarly Edition with CLLD

In the final step, all previous representation layers are presented together on an interactive and searchable web application. This web application was built using the CLLD framework (Forkel 2014), which features convenient bootstrapping methods for displaying CLDF datasets. The web application displays all forms, varieties and concepts in interactive tables that incorporate features like sorting, filtering or searching. The geographic location of varieties is displayed on an interactive map, and their respective sound inventories are inferred automatically using PyCLTS (List 2024a) and presented using the CLLD IPA chart plugin (Forkel 2024). All of these features highlight the benefit of digital scholarly editions, since these functionalities exceed the limitations of print documents.

The web application includes the relevant snippets from the original document for each form. It is therefore one of the first CLLD web applications to include image data from the original document directly in the data tables, following the technical implementation by Pulini & List (2025). This means that the full provenance of the data can be displayed transparently, from the original document over the digital transcript to the standardized and segmented IPA representations. We discuss the advantages of this in further detail in § 4.4.

4. Results and Examples

4.1. General Wordlist Statistics

Wordlist Summary	
Varieties	179
Concepts	292
Lexemes	45,364
Synonymy	1.01
Tokens (Segments)	254,074
Types (Segments)	133
Inventory Size (avg.)	31.06

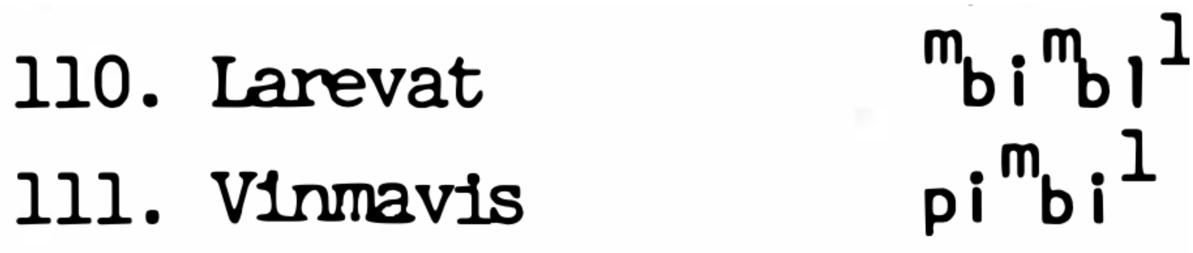
Table 1: Summarized statistics of the wordlist data.

Using the techniques outlined in § 3, we created a CLDF dataset derived from Tryon's (1976) data, whose summary statistics are presented in Table 1. The data includes lexical forms for 292 different concepts in 179 varieties, amounting to a total of 45,364 lexemes that are further segmented in a total of 254,074 phonetic units. Tryon usually presents only one form per concept and variety; only in rare cases, he offers phonetic or lexical variants. This results in a very low synonymy of 1.01, which reflects the average number of lexical items for a concept in a variety. Varieties feature a mean inventory size of 31.06 distinct segment types, which is very close to the average size of sound inventories reported in global samples (30.97 in Maddieson 1984; 31.69 in Moran 2012: 226 under genealogical resampling). The sizes of the sound inventories in our dataset range from 18 in Tolomako to 44 in Seke.

4.2. Challenges in Transcription

Since the original document was produced with a typewriter, and the scans were of high quality, the transcription was generally straightforward and unambiguous. There were, however, a few occasional uncertainties that we briefly address here.

The most prominent issue affecting the readability of the original document is that <l> (the lowercase <L>) and <i> may be hard to distinguish. In the typesetting of the original document, there is only little space between the dot and the body of the <i>, such that, owing to either smeared ink or issues with the scan, <i> and <l> may look identical. The similarity between <i> and <l> was already an issue during the preparation of the original document: We find instances where <i> and <l> are clearly distinguishable, but the two characters were confused. This is clearly illustrated by the forms <iiise> (Mpotovoro: 'far'; p. 527) and <lll> (Lorediakarkar 'fly (n)'; p. 256). In such cases, where the printed character is clearly identifiable, we initially provide a completely faithful transcription. These cases are corrected later in the standardized representation of the digital scholarly edition (see § 4.4), with editorial comments justifying each individual decision.

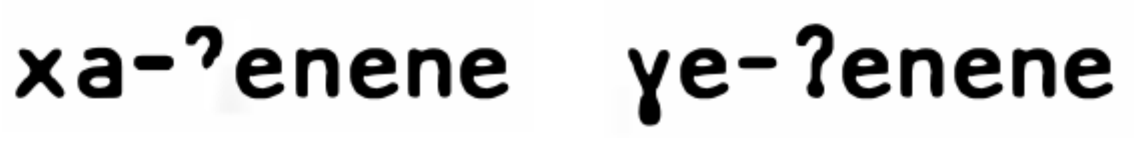


110. Larevat ^{m}bi^{m}bl^{1}

111. Vinmavis pi^{m}bi^{1}

Figure 2: The entries for 'uncle' in Larevat and Vinmavis (Tryon 1976: 242).

By contrast, when <i> and <l> cannot clearly be distinguished from each other, we transcribe the character that is more probable based on phonotactics and comparison with cognates in other varieties. Such a case is illustrated in Figure 2, where the Larevat form can be read as either <^{m}bi^{m}bl> or <^{m}bi^{m}bi>, but phonotactic considerations and the form <pi^{m}bi> in Vinmavis — a close relative, immediately below Larevat in the original document — makes the latter reading more likely.



xa-ˀenene ye-ʔenene

(a) (b)

Figure 3: The entries for 'straight' in Vatrara (a) and Sasar (b) (Tryon 1976: 400).

Further uncertainties originate in the representation of glottal stops. Figure 3 shows two word forms with a different graphic variant of the glottal stop letter. These two symbols are not standard typewriter characters, but arguably an ad hoc mechanical modification of the character for the question mark <?>. However, we find clearly distinguishable instances between a modification that completely removes the dot (Figure 3a) and one that retains it, connecting it to the body (Figure 3b). Since these character modifications are visibly distinct from each other, we choose to represent them as <ˀ> and <ʔ>, that is as two different characters in the initial digital transcript. Considering the visual information conveyed by the print characters in the original document, we argue that this is the most faithful representation possible.

Our interpretation, however, is that the difference between these two characters is merely due to technical reasons and does not reflect an intentional distinction made by Tryon. The data features a total of 862 instances of glottal stop characters, 820 (95%) of which we read as <ˀ>. Since <ˀ> is so much more common than <ʔ>, it is highly unlikely that this distinction is intentional. Following this, we interpret both <ˀ> and <ʔ> as instances of the glottal stop [ʔ] and map them accordingly in the standardized representation. However, we document the visual difference between the two characters in our initial digital transcript.

The original document also contained superscript characters that lack a Unicode equivalent. This concerned a very small number of word forms and represented an edge case, because such characters cannot be written in plain machine-readable text, so the only option was representing them with alternative graphemes. In the web application, the original superscripts can be restored and displayed with markup representations.

4.3. Challenges in Interpretation

Throughout his work, Tryon generally uses a fairly consistent transcription system that closely resembles the IPA. Therefore, the interpretation and conversion to IPA have usually been straightforward. However, there are a few instances where the symbols used by Tryon are not clearly interpretable. These cases highlight the need to separate the initial transcription from the subsequent interpretation of the document, while documenting both processes.

The most challenging interpretation concerns the representation of rhotics. Tryon uses the symbols <r>, <ř> and <r̃>. In most languages, only <r> occurs, but in a few languages, either <ř> or <r̃> appear alongside <r>.[3] Some Vanuatu languages spoken on Santo, Malekula and Efate also feature a complex potentially rhotic segment, the so-called “NDR sound”, which can be considered a “prenasalised coronal trill with a plosive-like release” (Rangelov 2023); Tryon normally transcribes this sound as <$^{\text{n}}\text{d}^{\text{r}}$>. In the next paragraphs, we discuss the uses of <ř> and <r̃> and how they can be interpreted.

Firstly, Tryon was probably aware that some Vanuatu languages may feature a phonemic contrast between an alveolar trill [r] and tap/flap [ɾ]. At the same time, we know that, cross-linguistically, for trill targets “the aperture size and air-flow must fall within critical limits for trilling to occur, and quite small deviations mean that it will fail” (Ladefoged & Maddieson 1996: 217), which results in trill phonemes often having tap/flap allophones in free distribution with the trill allophones. Since for most varieties Tryon did not have enough data to make conclusive phonemicity judgements (cf. Dockum & Bowern 2019), he likely used different symbols to denote a potential distinction in those few cases where he might have suspected one.

When we compare Tryon’s data with later phonological analyses for some of the relevant varieties, we find that there is a tendency for <ř> to represent a trill and

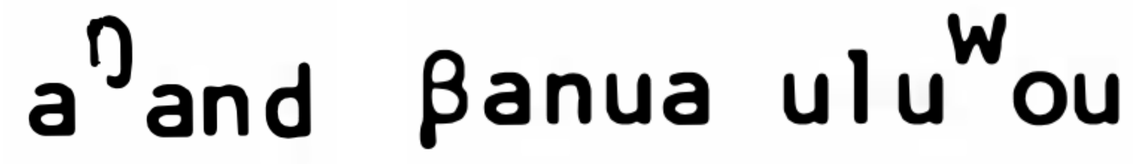
[3] Araki is the only language where all three occur, with only a single instance of <r̃>. This is most likely a typo, as we discuss in § 4.3.

<r̃> a flap/tap. However, this tendency does not always hold, and Tryon's choices are not always transparent, so a few elaborations are in order.

In Tryon's data, <ř> appears in Araki, Polonombauk, Narango and Piamatsina. More recent Araki data (François 2002) confirm a phonemic distinction between flaps and trills, and that Tryon's <ř> generally denotes the trill. For Piamatsina, <ř> only occurs after <n> in Tryon's data, and comparison with recent Piamatsina data by Takau et al. (2025) strongly suggests that the two characters together (<nř>) refer to the NDR sound, whose trill phase is often well pronounced following the plosive-like release (Rangelov 2023).

On the other hand, <r̃> is used in six varieties in Tryon's data: Malo North, Malo South, Shark Bay I, Lorediakarkar, Wusi-Mana and Sowa. For Malo North, Jauncey (2011: 13, 24) finds only one phonemic rhotic and reports impressions that it tends to surface as a tap word-initially. In Tryon's data, both <r> and <r̃> are found both word-initially and intervocalically, probably also reflecting the cross-linguistic tendency for free allophony mentioned above. Similarly, recent data on Wusi (Takau et al. 2025) suggest that there is no phonemic distinction between a tap/flap and a trill in this variety. Given that in Tryon's Malo and Wusi varieties <r> is substantially more frequent, we assume that this character was used to represent the default trill realization (as in most other varieties), while <r̃> was reserved for tap/flap realizations.

The above discussion shows that Tryon's use of the three rhotic symbols is not systematic. However, we wanted to ensure a 1-to-1-mapping of Tryon's rhotic symbols to IPA, so we have made the following choices: Since there is evidence for <ř> representing a trilled articulation where a contrast has been suspected, we use [r] to represent this symbol. For Tryon's <r> we use [ɾ], which is likely a frequent realization anyway. Finally, we represent <r̃> as [ɾ̥] to keep Tryon's distinctions.


aᵑand
βanua uluʷou

(a) Timbembe 'ant (red)' (p. 272)

(b) Vunapu 'man's house' (p. 241)

Figure 4: Entries with characters misrepresented as diacritics.

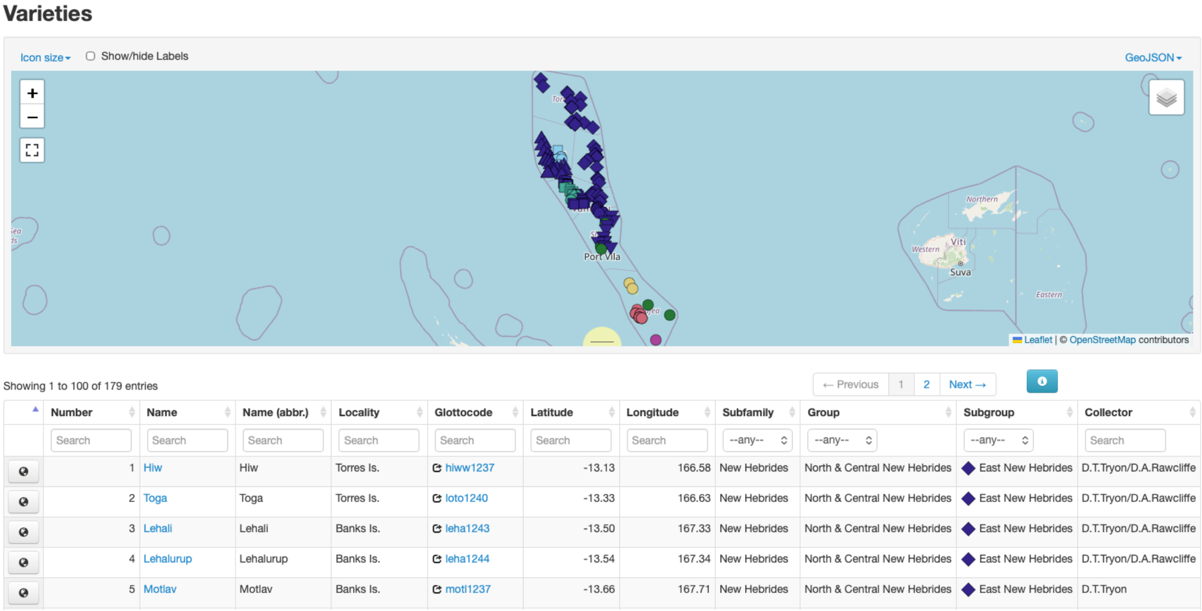
In comparison, other instances of characters that are not readily interpretable in IPA were more straightforward to handle. In a few forms, the diacritics <ᵑ> and <ʷ> occur in isolation without a valid sound to modify, as illustrated in Figure 4. In these cases, we assume that this is simply a typo in the original document, and we represent these characters as full segments [ŋ] and [w]. Finally, there are a few instances where vowels are followed by the interpunct <·>. We interpret this symbol as indication for a slight, but not full lengthening, therefore transcribing the modified vowel as semi-long [Vˑ]. We find evidence for longer vowels in the respective forms in independent sources, for example in Schütz (1969) for Pwele.

The following section showcases how the web application transparently indicates these kinds of editorial choices in the interpretive layers of our edition and clearly separates them from the digital transcript.

4.4. CLLD Application as Scholarly Edition

A digital scholarly edition of a historical document is a critical, often multi-layered representation of the data therein that cannot be printed without losing substantial information (Sahle 2016: 27). Digitality is therefore not only a convenience, but a core functional feature. In our digital edition of Tryon's data, this is best illustrated by the web application showcasing all representation layers of the data while transparently indicating all editorial choices made in transcription and interpretation.

At the core of the web application are three tabs that allow the user to browse varieties, concepts, and forms, respectively. Varieties and concepts are enriched by metadata and external references to Glottolog (Hammarström et al. 2025) and Concepticon (List et al. 2025), which are directly linked.

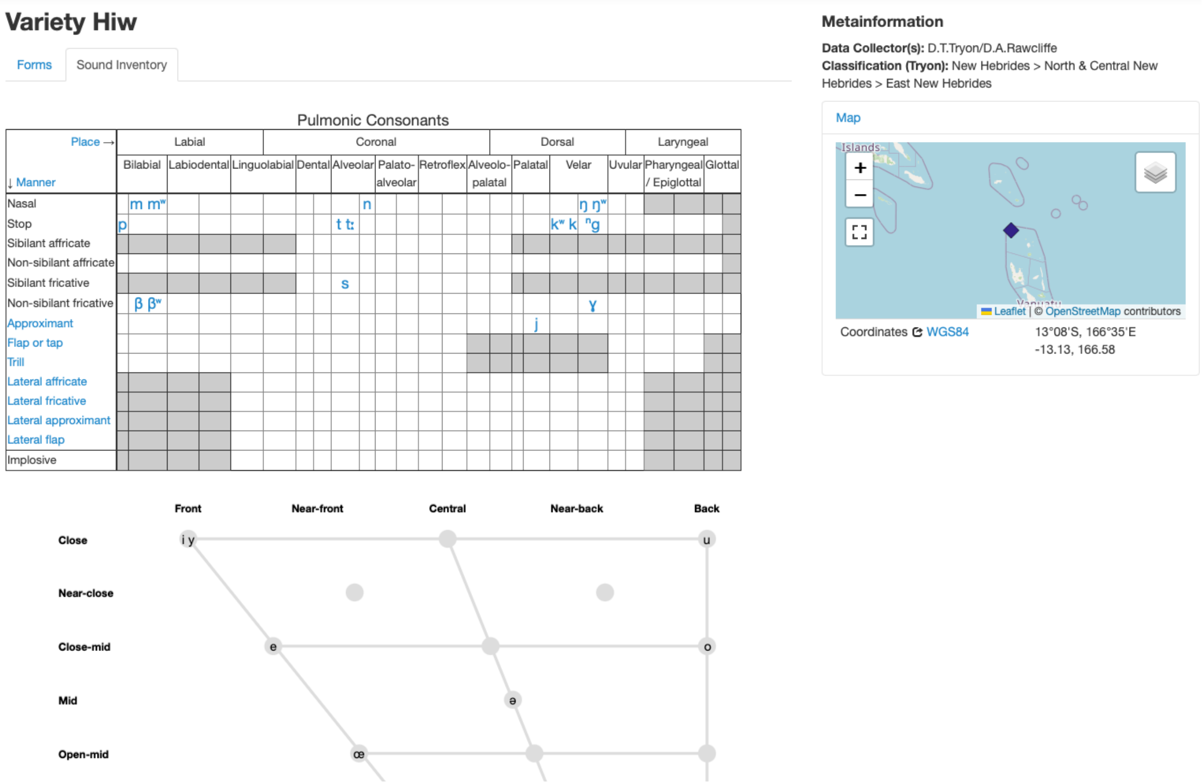


Showing 1 to 100 of 179 entries

	Number	Name	Name (abbr.)	Locality	Glottocode	Latitude	Longitude	Subfamily	Group	Subgroup	Collector
	Search	Search	Search	Search	Search	Search	Search	--any--	--any--	--any--	Search
	1	Hiw	Hiw	Torres Is.	hiww1237	-13.13	166.58	New Hebrides	North & Central New Hebrides	◆ East New Hebrides	D.T.Tryon/D.A.Rawcliffe
	2	Toga	Toga	Torres Is.	loto1240	-13.33	166.63	New Hebrides	North & Central New Hebrides	◆ East New Hebrides	D.T.Tryon/D.A.Rawcliffe
	3	Lehali	Lehali	Banks Is.	leha1243	-13.50	167.33	New Hebrides	North & Central New Hebrides	◆ East New Hebrides	D.T.Tryon/D.A.Rawcliffe
	4	Lehalurup	Lehalurup	Banks Is.	leha1244	-13.54	167.34	New Hebrides	North & Central New Hebrides	◆ East New Hebrides	D.T.Tryon/D.A.Rawcliffe
	5	Motlav	Motlav	Banks Is.	motl1237	-13.66	167.71	New Hebrides	North & Central New Hebrides	◆ East New Hebrides	D.T.Tryon

Figure 5: Geographic map view of language varieties in the CLLD application.

Figure 5 shows an overview of the varieties presented in the dataset, which are displayed on an interactive map, with colors and shapes representing the groups and subgroups established by Tryon (1976: 87-93). The geocoordinates are mostly obtained from Glottolog, with a few missing datapoints having been manually added based on the maps in Tryon (1976: 82-86) and the location of the respective settlements. Below the map, different types of metadata are represented in an interactive table that conveniently allows users to search for certain varieties or filter by a given parameter, e.g. according to the groups established by Tryon's lexicostatistical classification. By default, varieties are sorted according to the order in which they appear in Tryon's wordlist, but the web application also allows sorting by any other column. Similarly, concepts are displayed in a sortable and searchable table with external references to Concepticon.



Variety Hiw

Forms | Sound Inventory

Pulmonic Consonants

Place → / ↓ Manner	Labial			Coronal					Dorsal			Laryngeal	
	Bilabial	Labiodental	Linguolabial	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal / Epiglottal	Glottal
Nasal	m mʷ				n					ŋ ŋʷ			
Stop	p				t tː					kʷ k ⁿg			
Sibilant affricate													
Non-sibilant affricate													
Sibilant fricative					s								
Non-sibilant fricative	β βʷ									ɣ			
Approximant									j				
Flap or tap													
Trill													
Lateral affricate													
Lateral fricative													
Lateral approximant													
Lateral flap													
Implosive													

Metainformation

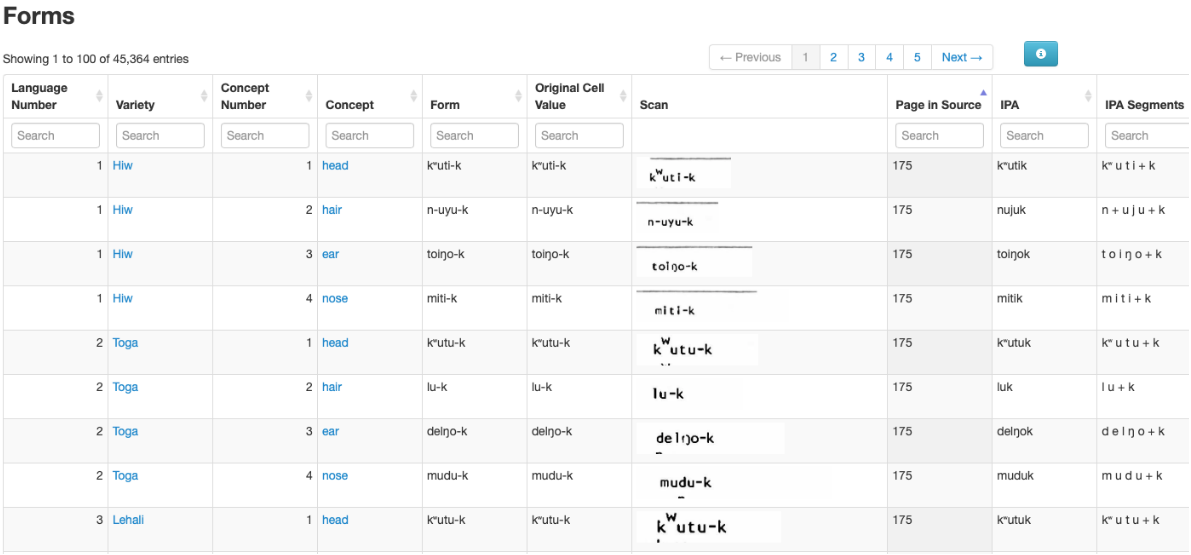
Data Collector(s): D.T.Tryon/D.A.Rawcliffe
Classification (Tryon): New Hebrides > North & Central New Hebrides > East New Hebrides

Map

Coordinates WGS84	13°08'S, 166°35'E -13.13, 166.58

Figure 6: Detail view of the language variety Hiw in the CLLD application.

All varieties and concepts are further displayed on a detail page that provides additional information, as well as a table of forms filtered according to the variety or concept in question. For each variety, a sound inventory is derived from the data and presented on the respective detail page**.** This gives users and editors a convenient functionality to inspect a variety's sound inventory and/or identify sounds that are phonologically or typologically interesting, or even pointing towards potential issues with the source data or their transcription. For instance, in the sound inventory of Hiw (Figure 6) [tː] stands out as the only geminate. However, Hiw does allow consonant gemination (François 2010: 399), but it only appears once in Tryon's data, in the form <ɣitteye> 'axe'. This confirms that this sound should actually be interpreted as a geminate and does not require further editorial intervention.



Forms

Showing 1 to 100 of 45,364 entries

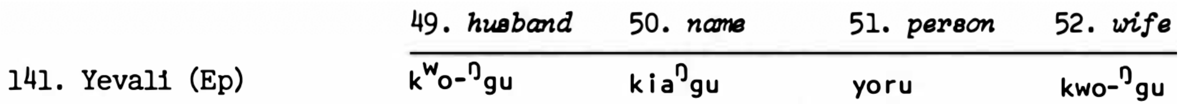
Language Number	Variety	Concept Number	Concept	Form	Original Cell Value	Scan	Page in Source	IPA	IPA Segments
Search	Search	Search	Search	Search	Search		Search	Search	Search
1	Hiw	1	head	kʷuti-k	kʷuti-k	kʷuti-k	175	kʷutik	kʷ u t i + k
1	Hiw	2	hair	n-uyu-k	n-uyu-k	n-uyu-k	175	nujuk	n + u j u + k
1	Hiw	3	ear	toiŋo-k	toiŋo-k	toiŋo-k	175	toiŋok	t o i ŋ o + k
1	Hiw	4	nose	miti-k	miti-k	miti-k	175	mitik	m i t i + k
2	Toga	1	head	kʷutu-k	kʷutu-k	kʷutu-k	175	kʷutuk	kʷ u t u + k
2	Toga	2	hair	lu-k	lu-k	lu-k	175	luk	l u + k
2	Toga	3	ear	delŋo-k	delŋo-k	delŋo-k	175	delŋok	d e l ŋ o + k
2	Toga	4	nose	mudu-k	mudu-k	mudu-k	175	muduk	m u d u + k
3	Lehali	1	head	kʷutu-k	kʷutu-k	kʷutu-k	175	kʷutuk	kʷ u t u + k

Figure 7: Example of the critical representation of forms in the CLLD application.

Figure 7 shows a screenshot of the first nine forms in Tryon's wordlist (under the Forms tab), in the order in which they appear in the original document. As can be seen, multiple representation layers are displayed, transparently separating Tryon's transcription (in the Original Cell Value column) from the interpretation. For cells indicating phonetic or lexical variants, individual forms are extracted by splitting up the original cell value onto two separate rows: e.g., Tryon's <peɣonə/yamə> (Hiw: 'sea'; p. 335) yields the two forms <peɣonə> and <yamə> each reported in a separate Form cell. In most cases, however, Tryon provides only one form; in these cases, the data in the Form and the Original Cell Value columns are identical.

By referencing the source page and directly including the relevant region of the scan, we enable users to trace each form from the original document to its standardized interpretation. For instance, consider the second form listed in Figure 7, <n-uyu-k> ('hair' in Hiw): We show our reading of the original cell value, alongside the scanned image of the form. Since Tryon uses <y> to indicate the palatal glide, the word's standardized and segmented IPA representation is [n + u j u + k]. This step allows the data to be compared and aggregated with forms from different sources. At the same time, such standardization constitutes a separate, derived representation layer — all information regarding the source document is retained and displayed faithfully.

Retaining these separate representation layers is crucial for ensuring that the entire digitization process is maximally transparent to the end user. However, not all cases are as clear-cut as the one presented above: Some forms are difficult to read or interpret correctly, as discussed in § 4.2 and § 4.3. If a user's interpretation of a form differs from ours, they can identify at which step disagreement arises: whether they *read* the form differently (column "Original Cell Value") or *interpret* it differently (columns "IPA", "IPA Segments").

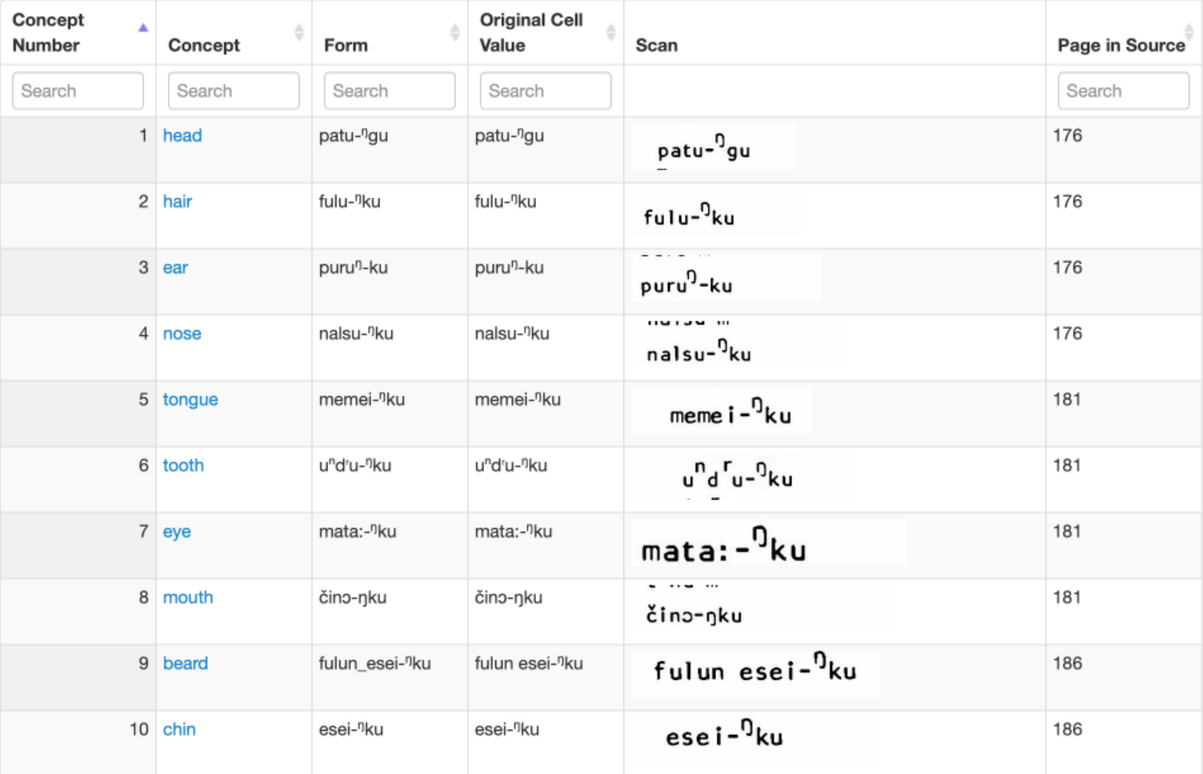


	49. *husband*	50. *name*	51. *person*	52. *wife*
141. Yevali (Ep)	k^{W}o-^{ŋ}gu	kia^{ŋ}gu	yoru	kwo-^{ŋ}gu

Figure 8: Transcription inconsistencies in the forms for ‘husband’ and ‘wife’ in Yevali (Tryon 1976: 238).

This gives us as editors the liberty to perform textual criticism and address some forms individually. Considering the vast amounts of data provided by Tryon, it is not surprising that the original document contains some errors. The different layers of representation are always based on the initial faithful digital transcript, while our edition amends evident errors in the interpretative layers (columns “IPA”, “IPA Segments”) and explains our interventions in the column “Comment” (this would be the equivalent of a critical apparatus in a critical edition, cf. § 2.2).

Figure 8 shows an example of an obvious inconsistency in the source data: In Yevali, Tryon reports the form <k^{w}o-^{ŋ}gu> for ‘husband’ (second column), but <kwo-^{ŋ}gu> for ‘wife’ (rightmost column). It is apparent that these two slightly different forms are the same lexeme: the two concepts are clearly semantically related and cross-linguistically often colexified to mean ‘spouse’. We can therefore correct this error and represent both forms as [k^{w} o + ^{n}g u] in our edition, while retaining the orthographic distinction in the digital transcript (the Original Cell Value column).



Concept Number	Concept	Form	Original Cell Value	Scan	Page in Source
Search	Search	Search	Search		Search
1	head	patu-^{ŋ}gu	patu-^{ŋ}gu	patu-^{ŋ}gu	176
2	hair	fulu-^{ŋ}ku	fulu-^{ŋ}ku	fulu-^{ŋ}ku	176
3	ear	puru^{ŋ}-ku	puru^{ŋ}-ku	puru^{ŋ}-ku	176
4	nose	nalsu-^{ŋ}ku	nalsu-^{ŋ}ku	nalsu-^{ŋ}ku	176
5	tongue	memei-^{ŋ}ku	memei-^{ŋ}ku	memei-^{ŋ}ku	181
6	tooth	u^{n}d^{r}u-^{ŋ}ku	u^{n}d^{r}u-^{ŋ}ku	u^{n}d^{r}u-^{ŋ}ku	181
7	eye	mata:-^{ŋ}ku	mata:-^{ŋ}ku	mata:-^{ŋ}ku	181
8	mouth	činɔ-ŋku	činɔ-ŋku	činɔ-ŋku	181
9	beard	fulun_esei-^{ŋ}ku	fulun esei-^{ŋ}ku	fulun esei-^{ŋ}ku	186
10	chin	esei-^{ŋ}ku	esei-^{ŋ}ku	esei-^{ŋ}ku	186

Figure 9: Forms in Narango with inconsistencies in the source.

The function to filter forms according to the variety or concept in the web app is particularly useful for spotting such errors or inconsistencies in the original document, since it enables users to view, side by side, a number of forms that would be distributed over several pages in the original document. The benefit of this is illustrated in Figure 9, where ten Narango forms are displayed. While these forms are found on different pages in the original document, they can be inspected together in one table on the web application. This reveals some issues with the source data: all of these forms are for body-part nouns, which in Oceanic languages are usually bound forms that require a possessive suffix. In all ten forms, we clearly see reflexes of the Proto-Oceanic first-person possessive suffix **-gu* (Lynch et al. 2002: 40, 76). The most common form is clearly <-ᵑku>. The <-ᵑgu> in <patu-ᵑgu> 'head' likely reflects some free variation that is not problematic. More interestingly, the suffix in <činɔ-ŋku> 'mouth' should arguably more precisely be transcribed with a superscript <ᵑ>, while in <puruᵑ-ku> 'ear', the dash indicating the morpheme boundary should precede <ᵑ>. The latter form had already been flagged as problematic in the orthography profile during the CLDF conversion, as discussed in § 4.3 — illustrating once again the benefit of computational methods for testing whether the data is well-formed.

Another form that showcases the benefits of the proposed tools to identify and transparently correct errors is Araki <r̃oβu> 'sugar-cane' (cf. Footnote 3). Several components that have previously been discussed come into play here. The automatic inference of phoneme inventories identified Araki as the only variety where all three rhotic characters <r, ř, r̃> occur. The function to filter forms by variety and a search for specific characters then revealed that <r̃> is only attested in 'sugar-cane' in Araki, while the other two characters are used more consistently. Considering that only Araki displays the three different rhotic characters and their occurrence is not justified by more recent data (François 2002), <r̃> is very likely a typo.

Finding more potential mistakes in Tryon's data requires further documentation, philological and/or comparative work. Here we highlight the potential of computational tools for detecting errors. A scholarly digital edition as presented here provides such opportunities owing to its interactive functionality. Likewise, it allows editors and users to interpret, standardize, modify and correct the original data without losing any information, since full provenance is always provided for each data point.

5. Conclusion and Outlook

We have presented a new workflow for creating scholarly digital editions in comparative linguistics, following the core principles of efficiency, transparency and integrability, and discussing computer-assisted tools to realize them. The use of good, fine-tunable OCR tools ensures that large collections of data can be digitized with a manageable effort. Operating in a framework that keeps humans in the loop while validating the integrity of the data ensures that the transcriptions are highly accurate.

Clearly separating the initial, faithful transcription from the following interpretive layers allows the data to be integrated with comparative wordlists from other sources that follow common standardization techniques, all without changing or overwriting any part of the original data. Computational workflows for standardizing the data in a rule-based, reproducible way ensure that all interpretative steps are explicitly recorded; the resulting web application transparently showcases all the decisions made in transcription and interpretation, allows the user to browse the data in an interactive way, and provides the full provenance for each form.

This work is a rare example of a scholarly digital edition of texts in non-Western languages. The presented workflow and tools are closely guided by the FAIR principles for managing data (Wilkinson et al. 2016), ensuring that the resulting edition makes the data findable, accessible, interoperable and reusable. Given that FAIR data is very sparse for many Indigenous and minority languages, we hope to set an example for improving the management of linguistic data that gives the respective languages a higher visibility and accessibility.

The workflow presented here should not be understood as a strict pipeline, but rather as a set of principles for creating scholarly digital editions of comparative linguistic data. The tools we propose are certainly not the only ones that can achieve our goals — in fact, one of the limitations of the current workflow is that it relies on the Transkribus platform, whose full feature set requires a paid subscription, thus rendering a complete open-source replication of the workflow impossible. Future research would benefit from exploring open-source alternatives for OCR and tabular layout recognition that would make the entire workflow reproducible.

We believe that our digital edition of Tryon's (1976) data can serve as the base for many future analyses and studies. For instance, colexification patterns can be easily extracted from the data and compared to global-scale colexification patterns to identify culturally specific semantic patterns (cf. Jackson et al. 2019). Since morpheme segmentations are already provided in the data, partial colexification patterns (List 2023) can easily be inferred as well without relying on heuristics (Blum et al. 2025). Since partial colexifications reveal different patterns that often provide semantic information complementary to full colexifications (Rubehn & List 2025; Bocklage et al. 2025), the data would serve as a good starting point for investigating common patterns in word formation, or for evaluating the quality of automatically-inferred partial colexifications.

Another obvious direction for future research is to use these data in historical linguistics. Since the data can now be fed into computational analyses, the regular sound correspondences identified by (Tryon 1976: 11-50) can be validated against automatically-inferred sound correspondence patterns (List 2019). Alternatively, phylogenetic inference could produce language trees based on shared cognacy, the same basic metric underlying Tryon's (1976: 87-93) subgrouping hypotheses. It would be interesting to compare the resulting phylogenetic trees with Tryon's lexicostatistical tree, as well as with reference trees.

In conclusion, while the concept of scholarly digital edition is hardly associated with typology and comparative linguistics, we have presented a case study that illustrates a possible workflow to prepare scholarly digital editions of legacy linguistic materials and the exciting research perspectives that improved digitization techniques will open up in linguistics.

Supplementary Material

The digital scholarly edition presented in this study is available online under https://tvl.digling.org. The CLDF dataset and the code for creating and validating it is hosted and curated on GitHub under https://github.com/calc-project/tryonvanuatu (v.0.1). The code for creating and hosting the CLLD web application is hosted and curated on Codeberg under https://codeberg.org/calc/tryonvanuatu-clld.

Author Contributions

JML initiated and supervised the study. JML and RR were responsible for the transcription workflow with Transkribus, with annotations by AR, TR and JB. AR and JML were responsible for the CLDF conversion and standardization. AR implemented the CLLD web application. TR, LC and JB contributed detailed information on the language varieties in question. AR wrote a first draft of the manuscript, with revisions and contributions by TR (§ 3.1), LC (§ 2.2) and JML (§ 2.3). All authors contributed to discussing the editorial choices discussed in the paper and agree on the current version of the manuscript and the digital scholarly edition.

Acknowledgements

We would like to thank Robert Forkel for his advice in creating the CLLD web application. This project was funded by the European Union through the ERC Consolidator Grant ProduSemy (AR, LC and JML, Grant No. 101044282, see https://doi.org/10.3030/101044282) and the Erasmus Plus Grant "Linguistics reimagined: Applying digital tools in language research" (JML, Project 2025-1-MT01-KA220-HED-000353724). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Co-funded by
the European Union

References

Anderson, Cormac & Tiago Tresoldi & Thiago Costa Chacon & Anne-Maria Fehn & Mary Walworth & Robert Forkel & Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting* 4(1). 21–53. (doi:https://doi.org/10.2478/yplm-2018-0002) (https://clts.clld.org)

Apollon, Daniel & Claire Bélisle & Philippe Régnier (eds.). 2014. *Digital critical editions*. Urbana, Chicago and Springfield: University of Illinois Press.

Auderset, Sandra. 2020. Interrogatives as relativization markers in Indo-European. *Diachronica*. John Benjamins Publishing Company 37(4). 474–513. (doi:10.1075/dia.19030.aud)

Austin, Peter K. 2017. Language documentation and legacy text materials. *Asian and African Languages and Linguistics* 11. 23–44.

Blanke, Tobias & Michael Bryant & Mark Hedges. 2012. Open source optical character recognition for historical research. *Journal of Documentation* 68(5). 659–683. (doi:10.1108/00220411211256021)

Blum, Frederic & Carlos Barrientos & Johannes Englisch & Robert Forkel & Simon J. Greenhill & Christoph Rzymski & Johann-Mattis List. 2025. Lexibank 2: pre-computed features for large-scale lexical data [version 2; peer review: 3 approved]. *Open Research Europe* 5(126). 1–19. (doi:https://doi.org/10.12688/openreseurope.20216.2)

Bocklage, Katja & Thanasis Georgakopoulos & Kellen Parker van Dam & Luca Ciucci & Frederic Blum & Alžběta Kučerová & Arne Rubehn & Abishek Stephen & David Snee & Johann-Mattis List. 2025. Testing the Potential of Automatically Inferred Affix Colexifications for Linguistic Typology. *Humanities Commons* 1–35. (doi:10.17613/a06m1-c9939)

Buzzoni, Marina. 2016. A protocol for scholarly digital editions? The Italian point of view. In Driscoll, Matthew James & Elena Pierazzo (eds.), *Digital scholarly editing*, 59–82. Cambridge: Open Book Publishers. (https://books.openedition.org/obp/3397)

Capell, Arthur. 1962. *A Linguistic Survey of the South-Western Pacific. New and Revised Edition. [With Maps.]* (Technical Paper 136). South Pacific Commission.

Chomé, Ignace. 1958. Arte de la lengua zamuca: Présentation de S. Lussagnet. *Journal de la Société des Américanistes* 47. 121–178.

Ciucci, Luca. 2018. Lexicography in the Eighteenth-century Gran Chaco: The Old Zamuco dictionary by Ignace Chomé. In Čibej, Jaka & Vojko Gorjanc & Iztok Kosem & Simon Krek (eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 439–451. Ljubljana: Ljubljana University Press.

Ciucci, Luca. 2021. How historical data complement fieldwork: New diachronic perspectives on Zamucoan verb inflection. *Studia Linguistica* 75(2). 289–327.

Ciucci, Luca. in prep. Ignace Chomé: Vocabulario de la lengua zamuca – Edición crítica y comentario lingüístico. Madrid and Frankfurt: Iberoamericana Vervuert Verlag.

Crowley, Terry. 2000. The language situation in Vanuatu. *Current Issues in Language Planning* 1(1). 47–132.

Dellert, Johannes & Thora Daneyko & Alla Münch & Alina Ladygina & Armin Buch & Natalie Clarius & Ilja Grigorjew et al. 2020. NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources & Evaluation* 54. 273–301. (doi:https://doi.org/10.1007/s10579-019-09480-6)

Dobrin, Lise & Saul Schwartz (eds.). 2021. *Special Issue on the Social Lives of Linguistic Legacy Materials*. *Language Documentation and Description*. Vol. 21.

Dockum, Rikker & Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description*. Aperio Press 16. 35–54. (doi:https://doi.org/10.25894/ldd112)

Fischer, Franz. 2019. Digital Classical Philology and the Critical Apparatus. In Berti, Monica (ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 203–220. Berlin, Boston: De Gruyter Saur. (doi:doi:10.1515/9783110599572-012) (https://doi.org/10.1515/9783110599572-012) (Accessed November 18, 2025.)

Forkel, Robert. 2014. The Cross-Linguistic Linked Data project. In Chiarcos, Christian & John Philip McCrae & Petya Osenova & Cristina Vertan (eds.), *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, 61–66. Reykjavik. (http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LDL2014%20Proceedings.pdf)

Forkel, Robert. 2024. *CLLD IPAChart Plugin [Software, version 0.3.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://pypi.org/project/clld-ipachart-plugin)

Forkel, Robert & Simon J. Greenhill & Hans-Jörg Bibiko & Christoph Rzymski & Tiago Tresoldi & Johann-Mattis List. 2024a. *PyLexibank: The Python curation library for Lexibank [Software, version 3.5.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://pypi.org/project/pylexibank)

Forkel, Robert & Johann-Mattis List. 2020. CLDFBench. Give your Cross-Linguistic data a lift. *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 6997–7004. Luxembourg: European Language Resources Association (ELRA). (http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf)

Forkel, Robert & Johann-Mattis List & Simon J. Greenhill & Christoph Rzymski & Sebastian Bank & Michael Cysouw & Harald Hammarström & Martin Haspelmath & Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5(180205). 1–10.

(doi:https://doi.org/10.1038/sdata.2018.205) (https://www.nature.com/articles/sdata2018205)

Forkel, Robert & Johann-Mattis List & Christoph Rzymski & Guillaume Segerer. 2024b. Linguistic Survey of India and Polyglotta Africana: Two Retrostandardized Digital Editions of Large Historical Collections of Multilingual Wordlists. In Calzolari, Nicoletta & Min-Yen Kan & Veronique Hoste & Alessandro Lenci & Sakriani Sakti & Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10578–10583. Torino, Italy: ELRA and ICCL. (https://aclanthology.org/2024.lrec-main.925)

Forkel, Robert & Christoph Rzymski & Johann-Mattis List & Gereon Kaiping. 2019. *PyConcepticon [Software library, version 2.3.0]*. Geneva: Zenodo. (doi:https://doi.org/10.5281/zenodo.3516955)

François, Alexandre. 2002. *Araki: A disappearing language of Vanuatu* (Pacific Linguistics). Canberra: Australian National University.

François, Alexandre. 2010. Phonotactics and the prestopped velar lateral of Hiw: Resolving the ambiguity of a complex segment. *Phonology*. Cambridge: Cambridge University Press 27(3). 393–434. (doi:https://doi.org/10.1017/S0952675710000205)

Franzini, Greta. 2012. Catalogue of digital editions. Vienna: Austrian Centre for Digital Humanities and Cultural Heritage. (https://dig-ed-cat.acdh.oeaw.ac.at/)

Gauchat, Louis & Jules Jeanjaquet & Ernest Tappolet (eds.). 1925. *Tableaux phonétiques des patois suisses romands.* Neuchâtel: Attinger.

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

Geisler, Hans & Johann-Mattis List. 2022. Of word families and language trees: New and old metaphors in studies on language history. *Moderna* 24(1–2). 134–148. (doi:https://doi.org/10.19272/202201902005)

Gengnagel, Tessa & Frederike Neuber & Daniela Schulz (eds.). 2023. Issue 16: Scholarly editions (FAIR criteria). *RIDE: A Review Journal for Scholarly Digital Editions and Resources* 16. (doi:10.18716/ride.a.16)

Greenhill, Simon J. & Robert Blust & Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4. 271–283.

Grierson, George Abraham (ed.). 1928. *Comparative Vocabulary*. *Linguistic Survey of India*. Vol. 1. Calcutta: Office of the Superintendent of Government Printing.

Hammarström, Harald & Robert Forkel & Martin Haspelmath & Sebastian Bank. 2025. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (doi:https://doi.org/10.5281/zenodo.14006617) (https://glottolog.org/)

Hirschmann, Hagen. 2019. *Korpuslinguistik: Eine Einführung.* J.B. Metzler.

Hull, D. L. 1988. *Science as a Process - An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: The University of Chicago Press.

IPA. 1999. *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet.* Cambridge: Cambridge University Press.

Italia, Paola & Giulia Raboni (eds.). 2021. *What is authorial philology?* Cambridge: Open Book Publishers. (https://books.openbookpublishers.com/10.11647/obp.0224.pdf)

Jackson, Joshua Conrad & Joseph Watts & Teague R. Henry & Johann-Mattis List & Peter J. Mucha & Robert Forkel & Simon J. Greenhill & Russell D. Gray & Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science* 366(6472). 1517–1522. (doi:https://doi.org/10.1126/science.aaw8160) (https://science.sciencemag.org/content/366/6472/1517)

Jauncey, Dorothy G. 2011. *Tamambom,the language of west Malo, Vanuatu* (Pacific Linguistics 622). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Kahle, Philip & Sebastian Colutto & Gunter Hackl & Gunter Muhlberger. 2017. Transkribus - A service platform for transcription, recognition and retrieval of historical documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19–24. IEEE. (doi:https://doi.org/10.1109/icdar.2017.307)

Koelle, Sigismund W. 1854. *Polyglotta Africana or Comparative Vocabulary of Nearly Three Hundred Words and Phrases in more than One Hundred Distinct African Languages*. London: Church Missionary House.

Kurzmeier, Michael & James O'Sullivan & Mike Pidd & Orla Murphy & Bridgette Wessels. 2024. Visualising the catalogues of digital editions. *The Journal of Electronic Publishing* 26(1). (doi:10.3998/jep.3569) (https://doi.org/10.3998/jep.3569)

Ladefoged, Peter & Ian Maddieson. 1996. *The Sounds of the World's Languages*. Oxford & Cambridge: Blackwell.

Lahaussois, Aimée. 2025. Methods and tools for (meta)grammaticography. In Bradley, David & Katarzyna Dziubalska-Kołaczyk & Camiel Hamans & Ik-Hwan Lee & Frieda Steurs (eds.), *Contemporary linguistics: Integrating languages, communities, and technologies*, 235–247. Leiden: Brill. (doi:10.1163/9789004715608_020)

Levinson, Steven & Nicholas Evans. 2010. Time for a sea-change in linguistics: Response to comments on `The Muth of Language Universals'. Lingua 120. 2733–2758. (doi:https://doi.org/10.1016/j.lingua.2010.08.001)

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45(1). 137–161. (doi:https://doi.org/10.1162/coli_a_00344) (https://www.mitpressjournals.org/doi/full/10.1162/coli_a_00344)

List, Johann-Mattis. 2023. Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology* 14(1156540). 1–10. (doi:https://doi.org/10.3389/fpsyg.2023.1156540)

List, Johann-Mattis & Cormac Anderson & Tiago Tresoldi & Robert Forkel. 2024a. *PyCLTS [Software library, version 3.2.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://pypi.org/project/pyclts)

List, Johann-Mattis & Cormac Anderson & Tiago Tresoldi & Christoph Rzymski & Robert Forkel (eds.). 2024b. *CLTS. Cross-Linguistic Transcription Systems [Dataset, version 2.3.0]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://clts.clld.org/)

List, Johann-Mattis & Annika Tjuka & Frederic Blum & Alžběta Kučerová & Carlos Barrientos Ugarte & Christoph Rzymski & Simon Greenhill & Robert Forkel (eds.). 2025. *CLLD Concepticon 3.5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (https://concepticon.clld.org/)

Liu, Lin & Yinhu Li & Siliang Li & Ni Hu & Yimin He & Ray Pong & Danni Lin & Lihua Lu & Maggie Law. 2012. Comparison of next-generation sequencing systems. *BioMed research international*. Wiley Online Library 2012(1). 251364. (doi:10.1155/2012/251364)

Lynch, John & Malcolm Ross & Terry Crowley. 2002. *The Oceanic languages*. Richmond: Curzon.

Maas, Paul. 1960. *Textkritik*. 4th edn. Leipzig: Teubner.

Maddieson, Ian. 1984. *Patterns of Sounds* (Cambridge Studies in Speech Science and Communication). Cambridge: Cambridge University Press. (doi:https://doi.org/10.1017/CBO9780511753459)

Memon, Jamshed & Maira Sami & Rizwan Ahmed Khan & Mueen Uddin. 2020. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access* 8. 142642–142668.

Michel, Jean-Baptiste & Yuan Kui Shen & Aviva Presser Aiden & Adrian Veres & Matthew K Gray & Google Books Team & Joseph P Pickett et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*. American Association for the Advancement of Science 331(6014). 176–182.

Michelone, Francesca. 2021. The critical edition between digital and print: methodological considerations. *Umanistica Digitale* 5(10). 25–48. (doi:10.6092/issn.2532-8816/12626) (https://umanisticadigitale.unibo.it/article/view/12626)

Moran, Steven. 2012. *Phonetics Information Base and Lexicon*. University of Washington. (PhD Thesis.)

Moran, Steven & Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Berlin: Language Science Press. (http://langsci-press.org/catalog/book/176)

Morel, Benoit & Pierre Barbera & Lucas Czech & Ben Bettisworth & Lukas Hübner & Sarah Lutteropp & Dora Serdari et al. 2021. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution*. Oxford University Press (OUP) 38(5). 1777–1791. (Ed. Malik, Harmit.) (doi:10.1093/molbev/msaa314)

Paulsen, Aksel & Knut Harboe & Ingvild Dalen. 2020. Data entry quality of double data entry vs automated form processing technologies: A cohort study validation of optical mark recognition and intelligent character recognition in a clinical setting. *Health science reports* 3(4). e210. (doi:https://doi.org/10.1002/hsr2.210)

Pulini, Michele & Johann-Mattis List. 2025. Using Cross-Linguistic Data Formats to Enhance the Annotation of Ancient Chinese Documents Written on Bamboo Slips. In Anderson, Adam & Shai Gordin & Bin Li & Yudong Liu & Marco C. Passarotti & Rachele Sprugnoli (eds.), *Proceedings of the Second Workshop on Ancient Language Processing*, 31–37. The Albuquerque Convention Center, Laguna: Association for Computational Linguistics.

Rangelov, Tihomir. 2023. The acoustic and articulatory properties of the prenasalised coronal trill in two Oceanic languages. *Proceedings of the 20th International Congress of Phonetic Sciences*, 3417–3421. Prague: ICPhS.

Rangelov, Tihomir & Eleanor Ridge & Lana Takau. 2025. Linguistics in Vanuatu 45 years after Independence. *Te Reo – The Journal of the Linguistics Society of New Zealand. Special issue on Vanuatu languages in context* 68(2). (Ed. Rangelov, Tihomir & Eleanor Ridge & Lana Takau & Victoria Chen.)

Rehbein, Malte & Belen Escobari & Sarah Fischer & Anton Günsch & Bettina Haas & Giada Matheisen & Tobias Perschl & Alois Wieshuber & Thore Engel. 2025. Quantitative and qualitative Data on historical Vertebrate Distributions in Bavaria 1845. *Scientific Data* 12(525). (doi:https://doi.org/10.1038/s41597-025-04846-8)

Reynolds, Leighton D. & Nigel G. Wilson. 1968. *Scribes and scholars: A guide to the transmission of Greek and Latin literature*. London: Oxford University Press.

Ringmacher, Manfred (ed.). 2022. *La Conquista espiritual de Antonio Ruiz de Montoya (1639) y su traducción al guaraní.* (http://www.etnolinguistica.org/biblio:ringmacher-2022-conquista)

Rosselli del Turco, Roberto. 2023. Filologia digitale: le prossime sfide, gli strumenti per affrontarle. In De Blasi, Margherita (ed.), *Moving texts. Filologie e digitale*, 15–39. Napoli: UniorPress.

Rubehn, Arne & Johann-Mattis List. 2025. Partial colexifications improve concept embeddings. *Proceedings of the Association for Computational Linguistics 2025. Long Papers*, 20571–20586. (https://aclanthology.org/2025.acl-long.1004)

Rubehn, Arne & Jessica Nieder & Robert Forkel & Johann-Mattis List. 2024. Generating Feature Vectors from Phonetic Transcriptions in Cross-Linguistic Data Formats. *Proceedings of the Society for Computation in Linguistics (SCiL) 2024*, vol. 7, 205–216. Irvine, CA.

Sahle, Patrick. 2016. What is a scholarly digital edition? In Driscoll, Matthew James & Elena Pierazzo (eds.), *Digital scholarly editing*, 19–39. Cambridge: Open Book Publishers. (https://books.openedition.org/obp/3397)

Sahle, Patrick. 2020. A catalog of digital scholarly editions, v.4.047. Bergische Universität Wuppertal. (https://www.digitale-edition.de/exist/apps/editions-browser/$app/index.html)

Schütz, Albert J. 1969. *Nguna grammar* (Oceanic Linguistics: Special Publication 5). Honolulu: University of Hawaii Press.

Skirgård, Hedvig & Hannah J. Haynie & Damián E. Blasi & Harald Hammarström & Jeremy Collins & Jay J. Latarche & Jakob Lesage et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*. American Association for the Advancement of Science (AAAS) 9(16). (doi:https://doi.org/10.1126/sciadv.adg6175)

Stussi, Alfredo. 2002. *Introduzione agli studi di filologia italiana*. Bologna: Il Mulino.

Takau, Lana & Tom Fitzpatrick & Mary Walworth & Aviva Shimelman & Sandrine Bessis & Tom Ennever & Iveth Rodriguez et al. 2025. *Vanuatu Voices [Dataset, version 1.4]*. Geneva: Zenodo. (doi:https://doi.org/10.5281/zenodo.17407434)

Thun, Harald & Franz Obermeier & Leonardo Cerno (eds.). 2015. *Guarinihape tecocue: lo que pasó en la guerra (1704–1705). Memoria anónima en guaraní del segundo desalojo de la Colonia del Santo Sacramento/Uruguay de los portugueses por los españoles*. Kiel: Westensee-Verlag.

Timpanaro, Sebastiano. 2005. *The genesis of Lachmann's method*. Chicago and London: The University of Chicago Press. (Ed. Most, G. W.)

Tjuka, Annika & Robert Forkel & Christoph Rzymski & Johann-Mattis List. 2025. Advancing the Database of Cross-Linguistic Colexifications with New Workflows and Data. *Proceedings of the 16th International Conference on Computational Semantics*, 1–15. (https://aclanthology.org/2025.iwcs-main.1/)

Tryon, Darrell T. 1976. *New Hebrides languages: An internal classification* (Pacific Linguistics, Series C-50). Canberra: The Australian National University.

Vercruysse, Bas & Julie M. Birkholz & Krishna Kumar Thirukokaranam Chandrasekar & Derrick Muheki & Wim Thiery & Hans Verbeeck & Koen Hufkens & Kim Jacobsen & Christophe Verbruggen. 2025. Human-in-the-loop tabular data extraction methods for historical climate data rescue. *International Journal on Document Analysis and Recognition*. Springer. (doi:https://doi.org/10.1007/s10032-025-00524-y)

Villari, Susanna. 2014. *Che cos'è la filologia dei testi a stampa*. Roma: Carocci.

Villavicencio, Frida. 2009. De la paleografía a la edición crítica: ¿una ecdótica para lenguas indígenas? In Clark de Lara, Belem & Concepción Company Company & Laurette Godinas & Alejandro Higashi (eds.), *Crítica textual: un enfoque multidisciplinario para la edición de textos*, 279–296. Ciudad de México: Colégio de México and UNAM.

Wilkinson, Mark D. & Michel Dumontier & Ilsbrand J. Aalbersberg & Gabrielle Appleton & Myles Axton & Arie Baak & Niklas Blomberg et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. Nature Publishing Group 3. 1–9.

Contact

Arne Rubehn (arne.rubehn@uni-passau.de)