Resource Acquisition for Understudied Languages: Extracting Wordlists from Dictionaries for Computer-Assisted Language Comparison

Frederic Blum[1], **Johannes Englisch**[1], **Alba Hermida-Rodríguez**[2], **Rik van Gijn**[2], **Johann-Mattis List**[1,3]
[1] Max-Planck Institute for Evolutionary Anthropology, Leipzig, [2]Leiden University, [3]University of Passau
{frederic_blum, johannes_englisch}@eva.mpg.de,
{a.h.r.hermida.rodriguez, e.van.gijn}@hum.leidenuniv.nl, mattis.list@uni-passau.de

Abstract

Comparative wordlists play a crucial role for historical language comparison. They are regularly used for the identification of related words and languages, or for the reconstruction of language phylogenies and proto-languages. While automated solutions exist for the majority of methods used for this purpose, no standardized computational or computer-assisted approaches for the compilation of comparative wordlists have been proposed so far. Up to today, scholars compile wordlists by sifting manually through dictionaries or similar language resources and typing them into spreadsheets. In this study we present a semi-automatic approach to extract wordlists from machine-readable dictionaries. The transparent workflow allows to build user-defined wordlists for individual languages in a standardized format. By automating the search for translation equivalents in dictionaries, our approach greatly facilitates the aggregation of individual resources into multilingual comparative wordlists that can be used for a variety of purposes.

Keywords: Cross-Linguistic Data Formats, dictionary parsing, computer-assisted language comparison

1. Introduction

Before the 20th century many Western linguists, missionaries, and archaeologists, often unified in one person, documented languages by recording comparative wordlists. Such wordlists formed the basis for historical language comparison and the reconstruction of ancestral languages. For example, the Linguistic Survey of India (LSI) documented 363 languages from southern Asia using such comparative wordlists (Grierson, 2023). Many of those languages have since become dormant and such documents are sometimes the only resource about them. In contrast, the late 20th and 21st century have seen a steep rise in extensive documentation efforts of individual languages, serving a diverse set of important community-oriented goals such as providing educational material for speaker communities or revitalizing obsolescent languages (Himmelmann, 1998; Gippert et al., 2006; Woodbury, 2014; Seifart et al., 2018). These documentation projects have led to an increased number of dictionary publications.

For historical linguistics, comparative lists of basic vocabulary are still the backbone for both classical and computational methods of language comparison (Durie and Ross, 1996; Greenhill and Gray, 2012; Blevins and Sproat, 2021; Blum et al., 2023b). Aggregated datasets of such wordlists also form the basis for interdisciplinary studies on cognitive aspects of language (Blasi et al., 2016; Jackson et al., 2019). Despite many efforts in automating steps of the comparative method (Wu et al., 2020; Blum and List, 2023), there are no standardized or transparent workflows for the compilation of comparative wordlists from dictionaries. Large comparative projects exist, but they are rare.

We propose a new approach for compiling such wordlists from individual sources, since no method exists for this purpose except the manual collection. In this study we present a computer-assisted method that allows for converting dictionaries into wordlists in a semi-automatic, transparent way that preserves references to the original dictionary. Apart from making wordlist extraction from dictionaries more transparent, the workflow can speed up the process of wordlist compilation and thus contribute to studies in which comparative wordlists have to be compiled from scratch or extended.

2. Background

Dictionaries and wordlists differ in their structure. In its most general representation, a dictionary consists of a *headword* and a *gloss*. The headword provides a form (or a lemma) in the language that the dictionary describes, and the gloss provides a hint to the meaning. The meaning itself can consist of multiple individual *senses*. Dictionaries may provide further information in addition to headword and gloss, such as the part-of-speech of a word, or example sentences that show how the word can be used. While the distinction between headword and gloss is present in nearly all dictionaries for individual languages, glosses differ widely and specifically sense descriptions are rarely standardized.

SIGUL2024 Workshop, pages 300–306

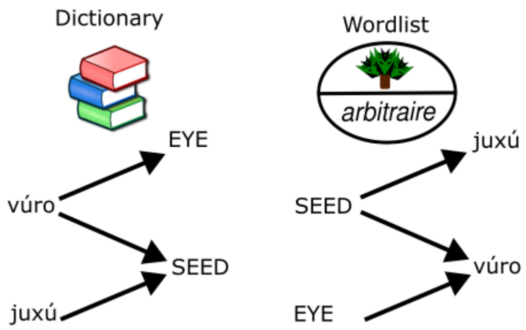

Figure 1: The structure of dictionaries and wordlists contrasted through the colexification of EYE and SEED in Amawaka (Case Study II).

In contrast to a dictionary that starts from the *word form*, taking a form-based or *semasiological* perspective, a wordlist starts from a list of *concepts* (or senses), taking a concept-based or *onomasiological* perspective (compare Lehmann 2004, 197 and List 2014, 22–24). A wordlist offers *translation equivalents*, based on a concept list in which individual concepts are referenced with short elicitation glosses (List et al., 2016). Since the relation between signifier (word form) and signified (meaning) can be complex, with forms denoting meanings consisting of multiple senses, there is no one-to-one relation between the elicited concepts in a wordlist and the glossed meanings in a dictionary. As a result, the same word form can occur several times in the same wordlist, each time representing different concepts, while at the same time one concept can be expressed by several different word forms.

An important part of the presented workflow is the standardization of data using the Cross-Linguistic Data Formats (CLDF), an initative for making linguistic data linked and re-useable (Forkel et al., 2018). CLDF comes with many different modules and provides the backbone for diverse datasets. For example, CLDF can represent lexical datasets (List et al., 2022), grammatical datasets (Skirgård et al., 2023; Blum et al., 2023a), or corpus data (Seifart et al., 2023). One of the core components of CLDF is the linking of data to other datasets through reference catalogues like Glottolog (Hammarström et al., 2024). The linking to those catalogues makes it possible to unambiguously identify points of comparison with other datasets that also use CLDF.

One such standardized reference catalogue that is especially relevant for this study is Concepticon, a repository for concepts and conceptlists (Tjuka et al., 2023; List et al., 2023). This reference catalogue stores lists of basic vocabulary and maps the entries to concepts, which establish translation equivalents across different source languages. For example, both English 'lake', German 'See' and Spanish 'lago' map to LAKE in Concepticon (`https://concepticon.clld.org/parameters/624`). This mapping process makes it possible to compare the meaning of lexical forms across different datasets with different source languages.

3. Method

3.1. Workflow

Linguistic dictionaries are published in many different formats. While more recent dictionaries are presented in a machine-readable form, older dictionaries are often only available as books where any information needs to be extracted manually. In other cases, proprietary tools like *Toolbox* or *Fieldworks Language Explorer* have been used to create dictionary files on a computer. But even when two different dictionaries are available as machine-readable files, the lack of standardization can lead to differently structured dictionaries, a lack of translation equivalents for dictionary entries, and different ways of presenting the same information. The manual extraction of comparative information is thus highly dependent on tedious and time-consuming manual work.

We present a workflow which extracts such wordlists from dictionaries of different source formats. Our method proceeds in four steps, as visualized in Figure 2. As a first prerequisite, a dictionary must be represented in machine-readable formats. This includes the digitization and parsing of data from different source formats. In a second step, the dictionary has to be converted to the specific dictionary representation of CLDF (Forkel et al., 2018). In a third step, the meaning descriptions in the dictionary are automatically mapped onto a user-defined selection of Concepticon concept sets (List et al., 2023). In this step we can easily create the translation equivalents for different source languages that have been used in the respective dictionaries. In a fourth step the mappings are used to extract a wordlist from the dictionary, which is then standardized following the guidelines underlying the Lexibank repository (List et al., 2022). The resulting dataset can be used as a starting point for comparative studies of many different kinds.

3.2. Parsing Dictionaries

The first step in our workflow is about converting the dictionary into a file that can be parsed computationally. If the raw data is available in machine-readable format, such as in our Case Study I, this may be skipped. More often than not, however, the dictionary is published as a PDF and requires some form of parsing or even a previous OCR scan,

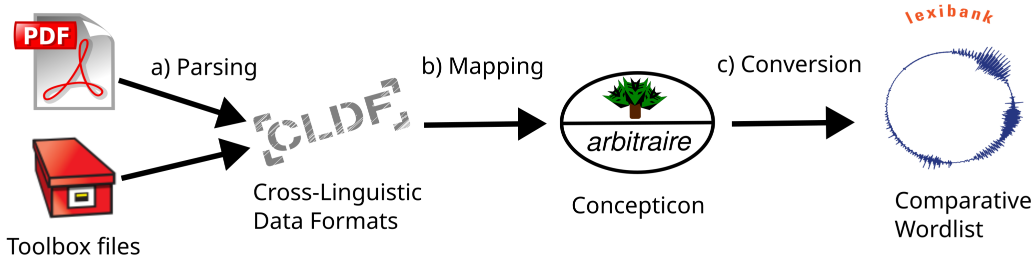

Figure 2: Overview of the workflow for the parsing of dictionaries and extraction of comparative wordlists.

as in our Case Study II. As these tasks are highly dependent on the source format, we will discuss them in each case study individually. As a general requirement for the CLDF conversion, we recommend having the dictionary parsed as a CSV file to easily iterate through the data. Other file formats, such as Toolbox text files, might also offer this option, and are another possible source format for the CLDF conversion.

3.3. Converting Dictionaries to CLDF

One of the cornerstones of our workflow is the creation of a CLDF dictionary. This is the step where all the different input formats get funneled into a uniform output format. For this purpose, we use the CLDFBench package (Forkel and List, 2020) to create the necessary metadata (`https://pypi.org/project/cldfbench`). CLDFBench projects can deal with a variety of diverse dictionary formats, be it Toolbox files, custom Excel sheets, or CSV files. Dictionary-specific support comes from the PyDictionaria package (`https://pypi.org/project/pydictionaria/`), which forms the back-bone of *Dictionaria*, an online journal for CLDF dictionaries (`https://dictionaria.clld.org`).

Depending on the source format, this process differs from dictionary to dictionary. For toolbox-dictionaries, a mapping file between the *Standard Format Markers* (SFM) markers and CLDF features is built (Case Study I). The SFM markers are the core of the toolbox-format and store all information of the entry in pre-defined headers. For example, '\lx' commonly presents the lexical form. Other markers can specify glosses in different languages or grammatical information. However, there are no enforced standards, and the mapping has to be adapted to each dataset. For dictionaries that have been parsed into tabular format, the script iterates through each line of the input format based on an established separator (e.g. tab or comma) and splits the input line into entry, senses, and other features such as part-of-speech tags, if available (Case Study II). CLDFBench is then used to create the final CLDF dataset.

The resulting CLDF dictionary contains a collection of linked tables, most relevantly an *Entry Table* and a *Sense Table*. The Entry Table contains the word form and additional – mostly grammatical and phonological – information. The Sense Table contains the different meanings of an entry and other semantic information. Note that the meaning descriptions provided in the Sense Table can be quite prosaic and vary between dictionaries. For comparative work, these descriptions need to be linked using a set of common concepts. This is the subject of the following section.

3.4. Automated Concept Mapping

Now that the CLDF dictionary is complete we can proceed to create the wordlist. For this step we choose a list of basic vocabulary from Concepticon that we want to use for our language comparison (Tjuka et al., 2023). If the desired list is not on Concepticon yet, one can easily follow a tutorial to contribute to this project (Tjuka, 2020). Once we have chosen the concept list, we map the entries from the dictionary to the list of concepts using a new Python package we wrote for this purpose, called `GetCL`, published in Version 0.1 along with this study (`https://pypi.org/project/getcl`).

The package uses a straightforward mapping algorithm available in the PySEM package (List, 2024) to map the dictionary entries to the concepts from the concept list (`https://pypi.org/project/pysem`). This is done through scoring the mapping of an entry to concepts in Concepticon based on previous mappings that have been established in the Concepticon workflow (List, 2022).

This step includes the option to use mappings from other languages that are already part of Concepticon. In our case studies, for example, we have used Spanish in addition to English to provide an automated mapping to our concept list, since the dictionary of Amawaka was published in Spanish.

The mapping should be followed up by two rounds of manual checks: First, we assure that all automated mappings are actually correct. Some ambiguous forms (e. g. 'bark') may have been mapped erroneously, and it is crucial for the comparative linguist that the mappings are corrected. Second, we check if any missing concepts can be found in the dictionary, for example by considering translations that are not yet part of the Concepticon mappings. By back-feeding this information to Concepticon we can improve the mapping process continuously.

3.5. Wordlist Extraction

The final step is the creation of the wordlist as a CLDF component. For this, we make use of the Lexibank specifications (List et al., 2022). This includes the selection of a concept list, mapping the languages to Glottolog (Hammarström et al., 2024), and ensuring that all sounds are represented in CLTS (List et al., 2021). The mapping to a concept list and the mapping of the described language to Glottolog are already part of the previous steps. The last feature that needs to be added is the standardization of the wordlist data through the creation of an *orthography profile* (Moran and Cysouw, 2018), a mapping table that maps from one orthography to another. In our case, the conversion is from the individual orthography used in a language resource to a phonetic transcription following the standard conventions of CLTS, which is derived from the International Phonetic Alphabet and compatible with it (Anderson et al., 2018).

The result of this procedure is a new CLDF dataset consisting of both the original dictionary and a standardized wordlist, which can be integrated with additional CLDF wordlists for the purpose of historical language comparison (Blum et al., 2024) or for computational approaches in lexical typology (Tjuka et al., 2024).

4. Case Studies

4.1. Workflow and Sample

The sample of two languages has been chosen out of convenience. We can showcase the workflow from two different sources: An existing pydictionaria repository, as well as a parsed PDF dictionary. The workflow is applicable to any dictionary that has a suitable input format available. In both case studies we use Swadesh's traditional concept list of 100 items (Swadesh, 1955). As mentioned before, it is possible to use any of the conceptlists in Concepticon for this purpose, or to create a new concept list if a study requires so. Table 1 summarises the total number of dictionary entries and senses as well as the number of mapped concepts for the target wordlist in both case studies.

4.2. Case Study I: Daakaka

In the first study we extract a comparative wordlist from a dictionary of Daakaka (von Prince, 2017), a language spoken by around 1000 speakers on Ambrym, Vanuatu (von Prince, 2022). Dictionaria already has a CLDF version of the dictionary, which we use as a basis for wordlist extraction. This CLDF dictionary is generated from a Toolbox file, which boils down to a flat list of key–value pairs called *Standard Format Markers* (SFM). PyDictionaria splits the list into separate entries and maps SFM markers to CLDF table columns. After that GetCL takes over the data and matches the individual meaning descriptions in the Sense Table to concepts from the Swadesh list. The extracted concepts are combined with the headwords from the Entry Table to create a CLDF wordlist.

At the end the whole process produces a hybrid dataset: The dictionary part contains 2167 entries referring to a total of 2229 different senses, and the wordlist provides word forms for 79 of the 100 Swadesh concepts. These automated mappings were supplemented manually with another 10 forms. This includes cases like '(fresh) water', which could not be mapped correctly to WATER due to the presence of additional information. We also removed five entries from the mappings. They were erroneously mapped either due to complex senses that included the target concept (e. g. 'a dish made out of fish' mapped to FISH) and the homophony in which cases of 'lie' are mapped to both LIE (REST) and LIE (MISLEAD). In total, we could map a form to 89 of 100 concepts.

4.3. Case Study II: Amawaka

In the second case study, we standardize the dictionary of Amawaka, a Panoan language spoken in the Peruvian and Brazilian Amazon, where it is spoken by around 500 to 600 persons. The digitization and scanning process for the Amawaka dictionary followed a systematic approach using an existing PDF. We made use of the proprietary OCR software ABBYY FineReader to convert the PDF file into searchable documents and then exporting them to TXT files. In the OCR recognition process the first step was to enhance PDF quality using ABBYY's scanning tool when needed, coupled with picture editing options to improve readability and reduce recognition errors. The second step comprised automatic format and text recognition, taking approximately 3 to 5 minutes for a 500-page dictionary. The third phase involved the verification and editing process. This step can be semi-automatic, as the software learns to recognize common mistakes, highlights recurrent 'unsure' characters, and those can be mass-changed in the search bar once identified. The final step involves exporting files to TXT files, maintaining the original format with automatic entry and subentry separation using tabs.

During the parsing of the extracted text data we take advantage of the consistent structure of the dictionary entries, which separates the senses and

Language	Glottocode	Source	Entries	Senses	Mapped
Daakaka	daka1243	von Prince (2017)	2167	2229	89/100
Amawaka	amah1246	Hyde (1980)	2106	2235	90/100

Table 1: Summary of both case studies: Number of dictionary and wordlist entries.

forms via part-of-speech tags. Apart from a handful of inconsistencies which needed manual solutions, this structure made it possible to iterate through the dictionary entry per entry with a clean separation of forms and senses by splitting the strings on the POS-tags. We strip the data of any whitespace and new lines, and export the final list to a TSV file of form `'Sense / POS / Form'`. The final table contains a list with the concept (e. g. LEAF), its form (/púhi/), as well as a link back to the sense-table of the dictionary ('1041-puhi'). In this case, the same form also links to FEATHER ('1605-puhi'), similar to the example provided in Figure 1.

We mapped 86 concepts to entries running the 'getcl' command. Following the manual check we removed two of those mappings (e. g. Spanish 'lengua' being mapped to TONGUE in cases where it means LANGUAGE) and added six concepts that were not mapped previously. In total we could successfully extract 90 of the 100 concepts of the Swadesh list from the dictionary.

4.4. Limitations

The main bottleneck for this workflow is the availability of machine-readable dictionaries. Even though OCR techniques have made huge progress, it is still difficult to digitize older dictionaries (e.g. from scans) in a quality that makes it reasonable to use them as resource for computer-assisted workflows.

Another limitation is the availability of languages for the mapping process for dictionaries with a source language other than English. While for some languages there is reasonable support (Spanish, Mandarin Chinese, German), the availability of high-quality mappings for many other languages in Concepticon is scarce. This is a direct consequence out of the fact that mappings are added through conceptlists that provide such a gloss, and most such lists are only presented in English, or other European languages. For example, there are 3756 available mappings for Spanish, 4612 for German, but only 28 for Marathi, and none for Hindi. Dictionaries written in languages for which no mapping resources exist are thus difficult to process with this specific workflow. A possible solution would be to pre-process the original data using automatic translations if available, but this would make it necessary to run even more quality checks after the mappings.

5. Conclusion

We offer a new standardized way to extract comparable wordlists from published dictionaries. Instead of going through dictionaries manually and typing out the relevant entries, our computer-assisted workflow establishes a reproducible way for offering a better analysis, for larger data. This reduces the error rate considerably, given that we avoid the chance of typos or missing an entry, making it necessary to go through the dictionary again. We expect that this workflow can reduce the workload for creating comparative wordlists considerably.

Mapping the entries to Concepticon ensures that we can directly compare data from different source languages with each other. For example, we could directly compare forms for a certain concept whose original publications were in Spanish, Portuguese, and English, because they all link to the same database. This can be used not only for historical language comparison and reconstruction, but also for studies that trace contact between languages. By maintaining the dictionary in CLDF format we also make it possible to re-use the dictionary data for other purposes, while computer-assisted steps assure the reproducibility of this effort.

6. Acknowledgements

This project was supported by the Max Planck Society Research Grant 'Beyond CALC: Computer-Assisted Approaches to Human Prehistory, Linguistic Typology, and Human Cognition (CALC^3)' (2022–2024, FB and JML), the ERC Consolidator Grant ProduSemy (Grant No. 101044282, see https://doi.org/10.3030/101044282, JML), and the ERC Consolidator Grant SAPPHIRE (Grant No. 818854, see https://doi.org/10.3030/818854, AHR and RvG). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for helpful comments and all people who share their data openly, so we can use it in our research.

7. Software and Data

All the code and data that was used in this study, including the case studies, is stored on Zenodo (v1.0.0, `https://doi.org/10.5281/zenodo.10948712`) and curated on GitHub (`https://github.com/FredericBlum/ExtractingWordlistsFromDictionaries`). The GetCL-package is available from pypi (`https://pypi.org/project/getcl/`).

8. Bibliographical References

Cormac Anderson, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4(1):21–53.

Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.

Juliette Blevins and Richard Sproat. 2021. Statistical evidence for the Proto-Indo-European-Euskarian hypothesis: A word-list approach integrating phonotactics. *Diachronica*, 38(4):506–564.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, Damián E. Blasi, and Roberto Zariquiey. 2023a. Grammars Across Time Analyzed (GATA): a dataset of 52 languages. *Scientific Data*, 10(835):1–11.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Zoe Poirier. 2023b. A phylolinguistic classification of the Quechua language family. *INDIANA - Anthropological Studies on Latin America and the Caribbean*, 40(1):29–54.

Frederic Blum, Carlos Barrientos, Roberto Zariquiey, and Johann-Mattis List. 2024. A comparative wordlist for investigating distant relations among languages in Lowland South America. *Scientific Data*, 11(92):1–9.

Frederic Blum and Johann-Mattis List. 2023. 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*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.

Mark Durie and Malcolm Ross. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press, New York, Oxford.

Robert Forkel and Johann-Mattis List. 2020. Cldfbench: Give your cross-linguistic data a lift. In *12th Conference on Language Resources and Evaluation*, pages 6995–7002.

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

Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel. 2006. *Essentials of Language Documentation*. Mouton de Gruyter, Berlin, New York.

Simon J. Greenhill and Russell D. Gray. 2012. Basic vocabulary and Bayesian phylolinguistics. *Diachronica*, 29(4):523–537.

George Abraham Grierson. 2023. CLDF dataset derived from Grierson's "Linguistic Survey of India" from 1928.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. *Glottolog database (v5.0)*. Max-Planck Institute for Evolutionary Anthropology, Leipzig.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Sylvia Hyde. 1980. *Diccionario Amahuaca*. Instituto Lingüístico de Verano, Yarinacocha.

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

Christian Lehmann. 2004. Data in linguistics. *The Linguistic Review*, 21(3-4):175–210.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List. 2022. How to map concepts with the pysem library. *Computer-Assisted Language Comparison in Practice*, 5(5):1–5.

Johann-Mattis List. 2024. *PySem: Python library for handling semantic data in linguistics [Software Package, Version 0.8.0]. With contributions by*

Johannes Englisch. MCL Chair at the University of Passau, Passau.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. Cross-Linguistic Transcription Systems Cross-Linguistic Transcription Systems (Version v2.2.0).

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon. A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2393–2400, Luxembourg. European Language Resources Association (ELRA).

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Annika Tjuka, Mathilda van Zantwijk, Frederic Blum, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2023. *CLLD Concepticon [Dataset, Version 3.2.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Frank Seifart, Ludger Paschen, Matthew Stave, and Robert Forkel. 2023. CLDF dataset derived from the DoReCo core corpus.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Annika Tjuka. 2020. Adding concept lists to concepticon: A guide for beginners. *Computer-Assisted Language Comparison in Practice*, 3(1).

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2023. Curating and extending data for language comparison in concepticon and NoRaRe. *Open Research Europe*, 2(141).

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2024. Universal and cultural factors shape body part vocabularies. *PsyArXiv Preprints*, pages 1–15.

Kilu von Prince. 2017. Daakaka dictionary. *Dictionaria*, 1(1):1–2167.

Kilu von Prince. 2022. *A Grammar of Daakaka*. De Gruyter.

Anthony C. Woodbury. 2014. Defining documentary linguistics. *Language Documentation and Description*, 1:35–51.

Mei-Shin Wu, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data*, 6(1):2.