Introducing the Historical African Languages Database: A Translingual Resource of Crosslinked Dictionaries

James Law, Brigham Young University
Daren E. Ray, Brigham Young University
Earl Brown, Brigham Young University

jimlaw@byu.edu

Abstract

The earliest written documentation of most African languages comes in the form of dictionaries and field notes prepared by European missionaries and linguists, with the assistance of African informants, in the nineteenth and early twentieth centuries. These resources have been difficult to access and compare, existing only in either print or unprocessed scans. We present a fully searchable and interconnected online database that makes such resources more easily accessible for study. It currently contains seven bilingual dictionaries, with many more sources to be added as they are processed. We explain the database's design, in which processed entries are separated and their fields tagged according to a consistent structure, maximizing query options and facilitating translingual connections. We describe the functionality of the website through which users can access the data in a variety of ways. We discuss the database's construction process, including particular challenges related to these historical data sources, and outline the development of a scalable procedure for its future expansion. We also present three case studies illustrating potential uses of the database by historians, linguists, and educators. Finally, we identify a roadmap for the resource's continued improvement through additional features.

Keywords: lexicography, historical linguistics, colonialism, database

1 Introduction

This paper presents a new digital resource that makes accessible a significant source of data on the history and diversity of African languages: nineteenth- and early twentieth-century colonial dictionaries and linguistic field notes. Beginning in the nineteenth century, European missionaries, linguists, and lexicographers recruited African informants to prepare descriptions of the grammar and vocabulary of African languages, often in connection with Bible translation efforts (Robinson, 2022). For most African languages, the publications that resulted represent the earliest written documentation of any kind (Nkomo, 2020). Some of these historically important documents have been scanned, with PDFs available on Google Books, HathiTrust Digital Library, Archive.org, or elsewhere on the internet, while others remain only in print form in archives. Our database collects these sources in a searchable, comparable form for the first time. We present this tool in an easy-to-use website for research by historians, linguists, teachers and learners of African languages.

The first dictionaries already added to the database mostly describe Bantu languages in Eastern and Southern Africa. However, the eventual scope of the database will include any language from the continent for which such colonial-era descriptions are available. This could include hundreds of individual sources.

In the following, we describe the historical context in which these dictionaries and notes were produced as well as our process for digitizing and assembling them into a useful database. We also propose three brief case studies illustrating how the database can be used for different research purposes.

2 Historical background

Protestant missionaries began the first coordinated efforts to compile dictionaries of African languages in the mid-nineteenth century. Previous travelers from Europe had recorded samples of African languages as early as the sixteenth century, usually in the form of wordlists and translations of Catholic catechisms (Wonderly and Nida, 1963:123). Protestant missionaries aimed to teach potential African converts in their vernacular languages (Constantine, 2013). They compiled extensive linguistic resources, both to prepare translations of the Bible and train future missionaries.

Missionary lexicographers were also motivated by advances in the study of linguistics and emerging standards for lexicography (Nkomo, 2020). In line with other modern dictionaries that had just begun to appear in Europe, they arranged entries in alphabetical order. Many of them also included parts of speech, definitions, sample sentences, grammatical notes, and etymological information.

Missionary lexicographers in Africa came from nearly every Christian denomination (Wonderly and Nida, 1963; Mkenda, 2018). They relied extensively on African collaborators who provided the raw material of word lists, usage, and grammatical knowledge to make their dictionaries (Robinson, 2022). However, they often worked independently of other missionaries because of their isolation in remote locations. They reported on their work to the Christian mission societies that funded their work, as well as to other missionaries working on similar languages. They often debated how best to represent the African language phonology and elicit vocabulary. Their correspondence included drafts of word lists, dictionaries, and language training manuals. In some cases, they left journals and other records in the archives of missionary societies that give insight into their language work (Paas, 2011; Krapf, 1860).

The Christian mission societies that raised funds for missionary work in Africa also funded editors to prepare language resources for publication. They then formed partnerships with publishing organizations such as the Society for the Promotion of Christian Knowledge. This charity published dictionaries alongside language lessons and pamphlets for distribution in Europe to preachers preparing for mission travel. As such, almost all of the dictionaries of African languages produced in the nineteenth century were bilingual rather than single language dictionaries.

Because of the time in which they were produced, these dictionaries are an invaluable resource for understanding the pre-colonial state of African languages. Colonialism has had a profound effect on the lexicons of African languages, not only through the borrowing of words from European languages but also through the cultural changes imposed or influenced by colonial power structures (Peterson, 1997). Although we are in most cases without written documentation of African languages from a pre-colonial time, the dictionaries included in our database often represent a state of the language at the very earliest stages of colonialism, when European influence was relatively limited.

It is important to acknowledge the significant limitations of these historical dictionaries as a data source. Their compilers had varying degrees of linguistic and anthropological training (Nkomo, 2020). While some dictionaries show remarkable detail and consistency, others are rife with errors or provide scant descriptions. It is especially important to recognize the European and Christian bias displayed in these sources (Peterson, 1997). The African informants who supplied the data for these dictionaries are rarely acknowledged, and in most cases little is known about them (Bank and Bank 2013). This sometimes has the effect of obscuring the particular variety that is being described, since the linguistic background of the informants is unclear and a dictionary may assemble data from multiple sources, some of them second-hand. As an example, Bishop Edward Steere's compilation of the Zanzibar dialect of Kiswahili was collected from students in his school who learned the language as a second or third language (Robinson, 2022). We must exercise caution to evaluate how faithfully the definitions provided represent the usage of African informants and the degree to which European missionaries inserted their own perspective (Peterson, 1997). For some users of the database, this may be an explicit object of inquiry. The content of the dictionaries in our database should not be understood as a fully objective portrait of the languages described, and caution must be exercised in interpretation of the data.

Since the publication of the dictionaries in the time period that is our focus (roughly up to 1930), African lexicography has advanced considerably.

Linguists with a scientific rather than religious objective produced grammars and dictionaries of African languages throughout the twentieth century (Nkomo, 2020). In the twenty-first century, many of these dictionaries have been put online, and new online resources have been developed (de Schryver, 2003). These efforts have naturally been more extensive for those languages that have higher numbers of speakers and politically official status, such as Kiswahili and isiZulu. There have also been projects bringing dictionaries of different languages together for comparison, such as the Comparative Bantu Online Dictionary (CBOLD[1]).

Despite this progress, for some very low-resource languages, dictionaries made in the colonial period may still be among the most detailed documentation yet published. Even for higher-resource languages such as Kiswahili, colonial dictionaries have significant historical and comparative value. While many of these colonial dictionaries have been scanned and put online, none have been digitally transcribed so that their contents can be searched like a modern online dictionary. This makes them largely inaccessible for comparison and research. For these reasons, our database primarily focuses on historical sources from the nineteenth and early twentieth centuries, despite their limitations, setting it apart from comparable databases of more recent sources.

3 Database design

We selected a set of these colonial dictionaries to digitally transcribe, structurally parse, and organize into the initial database. We first prioritized a variety of Bantu languages to ensure that our database structure accounted for representations of Bantu noun classes in dictionary entries. We also prioritized dictionaries that displayed relatively more complexity in the structure of their entries (i.e., including additional fields beyond the standard headword, part of speech, and definition), in order to develop a database framework that could account for this complexity. Additional details regarding the selected dictionaries and parsing process are provided in the next section.

The database is accessible via a website[2] and is regularly updated as additional features and dictionaries are added. The web portal for accessing the database is designed to facilitate comparison across dictionaries and languages, increasing its usefulness for language students, historians, linguists, and other users. The site currently provides three main functions: browsing individual dictionaries, searching entries across multiple dictionaries, and accessing metadata about the dictionaries and their creators.

Individual dictionaries can be browsed by selecting a source and a starting letter. Corresponding entries are then presented in order, with each of their fields (headwords, parts of speech, definitions, example sentences, etc.) presented in a consistent format. Many dictionaries include internal references to related words in the same dictionary; these words are clickable and direct to the corresponding related entry. A button beside each entry displays the image of the PDF page on which it is found, so that users can compare the digitally transcribed text to its original source side-by-side.

The search function permits searching across all dictionaries (by default) or across a subset or a single dictionary. Queries, which can target exact matches or partial matches with the beginning, middle, or end of an expression, can target headwords, particular fields (definitions, example sentences, etc.), or all fields together. Resulting entries matching the query are grouped by dictionary, with the same presentation format as the browse function: consistently structured entries with a button to display the original page side-by-side with the digital text.

Metadata about each dictionary are presented on individual pages of the site. These include the available information about the informants, linguists, lexicographers, and publishers involved in its creation and the relevant historical and ethnographic context. These pages also discuss how we adapted the information and formatting of linguistic data in each dictionary to our database structure. That is, we describe the kinds of information typically found in entries of each dictionary and how they are presented, and we explain how we categorized them to fit into the standard data structure of our database. Any information necessary to interpret a source's entries in the database, such as the meaning of abbreviations used by the lexicographer, is also explained on these pages.

In addition to the entries that comprise the main content of each dictionary, sources generally

[1] http://www.cbold.ddl.cnrs.fr/

[2] http://hald.byu.edu

include prefaces, lists of terms, grammatical descriptions, and other such details in frontmatter or appendices. These contents are provided on these metadata pages in their original PDF forms.

4 Process

The processing of dictionaries for incorporation into the database always begins with an analysis of the document's structure. We examine sample entries closely to identify how they are organized, the kinds of information they contain, and the typographical conventions they follow. These historical documents vary in their formality and level of copyediting, and we must often account for exceptions and mistakes in the layout of dictionary entries, such as inconsistent use of abbreviations or indentation. This analysis informs the rest of the process.

For most of the dictionaries we are interested in, quality scans are already available online; where necessary, we perform our own scans. Scanned PDFs are first preprocessed to maximize their computer readability. Preprocessing steps used vary according to the quality of individual scans, but may include contrast enhancement, binarization, deskewing, or noise removal.

Optical character recognition (OCR) is performed using the Tesseract 5.5 engine (Smith, 2007), which we selected over commercial OCR software because of the ability to fully adjust parameters for each dictionary to produce the best results. In addition to the text itself, our OCR approach outputs data on the position of each word on the page (useful for identifying indentations and other positional characteristics that mark meaningful elements of entry structure) as well as a confidence score, which allows us to automatically delete low-confidence items (typically stray marks) and flag moderate-confidence words for manual verification and correction.

We use Python scripts to convert the raw OCR text output to structured JSON representations of each dictionary. These scripts are individual to each dictionary, as they must account for variations in the fields included in entries and the ways those fields are typographically distinguished. However, because most of the dictionaries follow common formatting patterns (such as indenting the headword of each entry or numbering alternate definitions), we are able to minimize coding time by reusing some core functions and modifying parameters as needed. Unlike some OCR engines, Tesseract 5.5 does not perform font recognition, so data about font style (bold, italics) is unavailable for use in parsing entries; however, positional and character data has been sufficient for parsing entry structure in all dictionaries included so far.

One common element in many of the dictionaries is the inclusion of example sentences in the target language illustrating the use of each word. To identify the example sentences and separate them from their translations and other entry elements such as definitions, we use machine learning models trained with PyTorch (Ansel at al., 2024) for language classification. Of course, low-resource languages can pose a challenge for language classification tools developed using machine learning. This is alleviated by the fact that the included dictionaries pair a (potentially low-resource) African language with a high-resource European language. We can therefore achieve acceptable results simply by distinguishing English/French (for definitions and translations) from not-English/French (for example sentences). Furthermore, for Bantu languages, we have successfully relied on a single Kiswahili language model to distinguish target language text from translation language text, eliminating the need for additional model training resources. For example, in a Mijikenda-English dictionary, strings are classified based only on their similarity to Kiswahili and English. Mijikenda example sentences are classified as Kiswahili by the model due to linguistic similarity between these two Bantu languages, and they can thus be separated from the English definitions and translations.

As we parse each dictionary, its content is converted to a standard JSON structure. Because not all dictionaries contain the same fields, we have opted for a maximalist structure; for example, although most dictionaries do not include etymological information in their entries, this field is available for those that do.

The digital text in the database is primarily a faithful representation of the scanned text. However, there are some exceptions to this. For languages with noun classes (such as Bantu languages), we convert variable information about noun class (for example, some dictionaries list affixes and particles associated with each noun, while others use custom numbering systems) to a maximal Bantu noun class numbering system that includes most scholarly variants (Maho 1999). We

Language Pair	Source
Maa - English	Erhardt & Krapf (1857)
Sotho - English	Kruger (1876)
Mijikenda - English	Krapf & Rebmann (1887)
Yao - English	Maples (1888)
Kiswahili - French	Sacleux (1939 [1888])
Luganda - English	Pilkington (1892)
Kikuyu - English	McGregor (1904)

Table 1: Dictionaries currently in the database.

also expand abbreviated parts of speech (e.g., converting *n.* to *noun*) and standardize them (e.g., converting *s.* for 'substantive', a synonym for noun, to *noun*) so that users can easily target words of a particular part of speech in their search queries. Although such modifications to conform to a consistent structure are minor, they are one reason we prioritized easy access to the original PDF for users of the site. Our standardized representation of dictionary entries can be quickly compared to the original with the click of a button.

Research assistants use a custom software tool to manually verify and correct each dictionary's content before adding it to the database. The software tool displays parsed entries according to the standard JSON structure we have adopted and allows for easy navigation between entries, side-by-side comparison with the PDF, and rapid correction. Errors can arise from mistakes in the OCR or from oversights in the parsing script that neglect to account for entries that have unusual content or formatting. We have found that correction proceeds most efficiently in two stages. During a first pass, assistants correct any errors in parsing (separating erroneously grouped entries, moving incorrectly assigned text to the proper field, etc.). Then during a second pass, they correct OCR errors, focusing on words with low confidence scores that are more likely to contain mistakes (misspellings, cut off words, etc.). They also identify issues that are consistent enough to be resolved through a mass edit, which they pass on to the project directors to implement programmatically. To clarify, the goal at this stage is not to correct any perceived errors on the part of the dictionary's creators, but simply to ensure that the digital representation accurately reflects the content of the document.

As of this writing, seven dictionaries have been processed for inclusion in the database (although final manual corrections are ongoing for some of these). They are listed in Table 1. They represent six Bantu languages and one Nilotic language, with either English or French as a translation language. All derive from data collected in the latter half of the nineteenth century or the early twentieth century.

5 Challenges

We have encountered several challenges specific to our data source. First, because these older dictionaries are printed in a variety of fonts that may not be standard in modern texts, the OCR output contains more errors than would be typical in a modern text. As explained above, manual correction is therefore a crucial step.

Another challenge is the variable entry structure across dictionaries. For example, the Mijikenda dictionary by Krapf and Rebmann (1887) organizes entries hierarchically, with some words subordinate to other words they are derived from, and includes the source language for many words; by contrast, the Yao dictionary by Maples (1888) has a flat entry structure and does not include source language notes. In order to place such varied sources together into a single database, we had to create a standard data structure that would accommodate all of the fields and relationships that the dictionaries include. We did our best at the beginning of the project to anticipate the fields that would be required by dictionaries added later on, designing a data structure that includes a maximal set of fields (many of which can be left blank for simpler dictionaries). However, we have also had to modify the data structure over the course of the project to add fields that we originally missed but that are included in certain dictionaries, such as the etymological notes included in the Kiswahili dictionary by Sacleux (1939 [1888]).

A further set of challenges posed by the data have been the ambiguous intentions of dictionary compilers in their presentation of the dictionaries, which have required interpretation. While most of the dictionaries include at least some frontmatter that explains the abbreviations used and other important context, these introductions are sometimes lacking in detail. Outdated or unusual linguistic terms such as *neuter verb* (generally meaning a kind of stative verb) are often used. In cases such as these, we generally err on the side of a faithful representation of the dictionary's contents, even if those contents may be unclear to modern users of the database; in such cases, we explain our interpretation of the terms used in the dictionary descriptions included on the website. However, in order to provide a structured and comparable analysis of each dictionary, we have had to commit to certain interpretations of ambiguous notation. For example, in Krapf and Rebmann's (1887) Mijikenda dictionary, similar words are noted in two ways: either the word *See* or an equals sign. Because this notation is not explained in the dictionary's frontmatter, we had to decide, based on consideration of many examples, to interpret these markers in two different ways: *See* indicates related words elsewhere in the dictionary (which are linked in our database for users to easily cross-reference), while an equals sign indicates comparable words that are not necessarily included in the dictionary and may be from other languages. We have no way of knowing if this is exactly the meaning that the compilers intended with this notation, but it is a reasonable interpretation of the data that allows the dictionary to be incorporated smoothly into the rest of the database.

6 Case studies using the database

In this section, we provide three examples of how this database could be used by historians, linguists, and other researchers, or by language learners and teachers. Our intention here is to illustrate the kinds of uses we had in mind as we created the database, although we hope it could be useful in other ways as well.

First, although the database is certainly a valuable resource for information about the languages described, it is even more directly useful for comparison of the dictionaries themselves. The database includes dictionaries that document the same or very similar languages, published at different times and places and compiled by different authors with their own motivations, biases, and sources. As an example of the kind of historical research through dictionary comparison that could be performed with the database, consider how one might compare entries for the same word or cognate words. The word *koma* is defined in the Mijikenda dictionary by Krapf and Rebmann (1887) as follows: "An evil spirit, supposed to be of some dead person. The chief idea of religion among the Wanyika seems to be to appease the koma." The cognate word *k`oma* is defined in Sacleux's (1939 [1888]) Kiswahili dictionary somewhat differently: "Esprit de mort, mânes" ('Spirit of a dead person, ancestral spirit'). The former dictionary thus describes a more pejorative sense, while the latter is more neutral. Supplementary evidence would of course be required to determine whether these differences in definition are due to actual differences in meaning at the time and place when the data was collected or due to bias on the part of the lexicographers (or some combination of these factors). The database facilitates this kind of research by allowing searches to filter results to a specific dictionary or dictionaries and to search for exact or partial matches in various fields. For instance, one could look for related evidence by searching for topical words like *spirit* or *god* or pejorative words like *evil* (and their French equivalents) in the definitions and example sentence translations of these two dictionaries. This kind of evidence provided by the database could be useful for studying the way colonial lexicographers interpreted African linguistic and cultural concepts through a European lens.

A second line of research that could be aided by the use of this database is the investigation of lexical change. The meaning or form of words can change for a variety of reasons, including language-internal factors such as phonological erosion and language-external factors such as contact with other cultures. By comparing the historical dictionaries included in this database to more recent dictionaries of the same languages, it is possible to observe and analyse these kinds of changes. A particularly fruitful line of inquiry might concern the influence of colonialism and related cultural shifts on the lexicon, since many of the dictionaries in this database represent a state of the documented languages prior to large-scale European expansion into Africa. To present just one example of the kind of lexical change we are thinking of, consider the word *chikulundine* from Maples' (1888) Yao dictionary, defined as "The third party who accompanies the bridegroom in asking the bride". Over a century later, this word is defined (written as *cikulundiine*) by Ngunga (2001) as "marriage agreement with the woman's relatives". Of course, there is the possibility that either or both of these definitions misrepresent the full scope of this word's meaning, given the scarcity of other data for confirmation. However, taking these definitions at face value, it appears as though the meaning of this word shifted over time from designating a specific role in a marriage negotiation to the marriage negotiation itself. This would be a case of broadening or generalization, or simply metonymization in Traugott and Dasher's (2002) categorization of semantic change. The new meaning could be due to natural semantic drift but could also possibly be tied to cultural changes surrounding marriage negotiations. By making the earliest sources for many African languages accessible and searchable, this database allows for diachronic lexical analysis of these languages to be performed at a larger scale than previously possible.

A third potential use of this database is simply as an additional internet source for low-resource languages. Setting aside the specifically historical value of these dictionaries, many of them are one of the only pieces of documentation for certain languages (or at least, once incorporated into our database, one of the only pieces of documentation that is easily accessible on the internet). Despite their flaws, they may also be more thorough in some ways than other resources or include information that is left out elsewhere. Consider Maa, a low-resource language with only one online dictionary of which we are aware (Payne and Ole-Kotikash, 2008). Any dictionary will include gaps that can be filled by others, and two examples will illustrate ways in which the Maa vocabulary in our database (Erhardt and Krapf, 1857) can be a useful supplement to this other online dictionary despite its early publication. First, there are words that appear in the 1857 dictionary and not in the 2008 dictionary, such as *mésera*, a type of tree described in some detail by Erhardt and Krapf. Second, words that appear in both dictionaries may include different information, together providing a fuller picture of the word's meaning. For instance, the verb *nuk* (Erhardt and Krapf's spelling) or *a-nʉ́k* (Payne and Ole-Kotikash's spelling) is defined by both dictionaries as 'bury'. However, the 1857 entry goes on to describe burial customs among the Maasai, while the 2008 entry lists several metaphorical extensions of the word's sense, such as 'to hide or conceal information'. Although separated by 151 years and subject to the same issues of bias and lexical change just discussed, these two dictionaries can still complement one another to provide users with the most complete information possible about this language. Because no dictionary can be entirely exhaustive, the use of multiple sources is often advisable to get the best information about a word; for some low-resource languages, our database of historical dictionaries makes this possible for the first time.

7 Future improvements and discussion

As stated earlier, the database currently contains the seven dictionaries listed in Table 1. An additional 11 sources have already been selected for processing in the database's initial phase of development. Some contain data on multiple languages, and together these will provide full dictionaries of 20 languages and more limited glossaries for more than 60 languages. Once these have all been processed, the database's expansion can continue. There are more than 100 similar sources from the nineteenth and early twentieth centuries that could be added over time. As our processing becomes more streamlined, we anticipate that each new source will be ready to add to the database within two weeks (plus time for manual correction as needed). We envision a database that includes historical dictionaries, glossaries, and field notes for languages from across the continent and representing several different language families.

The current architecture of the database and its website are close to being finalized. However, there are some features that we hope to develop further that would increase the resource's functionality. Currently, entries in search results can be visualized in their original form by displaying an image of the page; we plan to refine this visualization so that just the part of the page containing the entry is displayed. This will increase the ease of use by allowing more rapid comparison of the data and source.

Other planned features relate to the translingual nature of the resource. The dictionaries included in the database (and others that will be added to it in time) use several different translation languages including English, French, and German. We hope to facilitate searching across these different languages by allowing automatic translation of search terms. When this feature is added, a search for the word "book" in the definition field could return not only results such as *chuo* 'a book' from the Mijikenda-English dictionary but also *buku* 'livre' ('book') from the Kiswahili-French dictionary. We likewise intend to provide automated translations of entry contents from any of these European languages to any other to facilitate user access to materials when browsing.

A major goal of the project is to present our data in a way that connects languages together. Our data sources largely present African languages as discrete objects, siloed by a colonial European understanding of language and ethnicity (Makoni, 1998; Chimhundu, 1992; Harries, 1987). Each named language is presented in its own dictionary, a bilingual dictionary that prioritizes its connection to a European translation language over its connection to its sister languages. This presentation inherently creates a linguistic hierarchy, with European languages as a necessary intermediary through which African languages must pass in order to connect with one another. Indeed, the development of these dictionaries was part of a colonial project that imposed a compartmentalizing and hierarchizing conception of language onto African societies (Ndhlovu and Makalela, 2021: 53-54). In contrast, the concept of translanguaging, promoted e.g. by García (2019), reflects the reality that individuals draw on a linguistic repertoire to communicate that may include different languages as traditionally defined but which are not necessarily compartmentalized as such in the minds or in the linguistic outputs of speakers. Translingual competence, fluid movement among multiple languages, has been highlighted in both the African context (Ouane and Glanz, 2010) and elsewhere (Geisler et al., 2007) as an educational priority. If we are to support this new model of language education and use, we need to provide data in formats that allow for such interconnectedness (e.g., Parton et al., 2008).

While our database represents the very dictionaries that historically imposed a monolingualizing view of language, we hope to present the data in a way compatible with translingual philosophy and practice. This conception of language is supported in our database by the ability to simultaneously search across many sources that were previously confined to a single dictionary centering a single named language. This means that words (cognate or non-cognate) that have a formal, semantic or grammatical connection may appear together in search results, regardless of their association to particular named languages. By searching for partial matches, users can identify words with similar forms across different languages, facilitating the study of cognates. However, these connections can go much further. We aim to enrich the database by linking, to the degree possible, words in languages of the Bantu family to their Proto-Bantu roots and borrowed words to the corresponding words in their source languages. This would increase the interconnectedness of the database, allowing users to see cognate words at a glance without the need for more complex searches. It would in effect remove the necessity for a European intermediary language, as users could move from one African language to another directly.

8 Conclusions

We have presented a new online database that assembles fully digitized versions of African language dictionaries, and similar linguistic resources, from the nineteenth and early twentieth centuries. The database is already live and freely accessible in its initial release, with further updates to come.

For many African languages, the earliest linguistic documentation has been difficult to access and compare, locked in print form or in unparsed scans. These historical sources are of great value to historians, linguists, and even language learners and teachers. Although outdated in some ways and influenced by historical biases, they offer a rich data supplement for many low-resource languages and provide a way to study linguistic change in the African colonial context. For some languages, the dictionaries (with example sentences) and linguistic notes made accessible here may be used fruitfully to supplement the training of large-language models, a significant challenge for low-resource languages that currently rely prominently on Bibles and the few available contemporary texts (Alhanai et al., 2025). Our database allows precise search queries and easy visualization and comparison of these lexicographic texts. It is our hope that by making old data available in this new format, this resource can be of use to a wide variety of stakeholders in the research, teaching, and promotion of African languages.

Acknowledgments

The authors would like to express gratitude to the Brigham Young University Office of Digital Humanities for their support of this project, as well as to the Research Development Office for an Interdisciplinary Research Origination Award that funded the development of the database. We would also like to thank our student research assistants

Alek Spackman, Katelyn Shellman, Aaron Shaw, Ben Carson, Emma Law, and Nils Young.

References

Tuka Alhanai, Adam Kasumovic, Mohammad M. Ghassemi, Aven Zitzelberger, Jessica M. Lundin and Guillaume Chabot-Couture. 2025. Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27). 27802–27812.

Jason Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. A. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, and S. Chintala (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '24)*, Vol. 2. 929–947.

Andrew Bank and Leslie Bank. 2013. *Inside African Anthropology: Monica Wilson and Her Interpreters*. New York City, NY: Cambridge University Press.

Herbert Chimhundu. 1992. Early Missionaries and the Ethnolinguistic Factor During the 'Invention of Tribalism' in Zimbabwe. *Journal of African History*, 33(1). 87–109.

Simon Constantine. 2013. Phrasebooks and the Shaping of Conduct in Colonial Africa ca. 1884-1914. *The International Journal of African Historical Studies*, 46(2). 305–28.

James J. Erhardt and Johann Ludwig Krapf. 1857. *Vocabulary of the enguduk iloigob, as spoken by the Masai-tribes in East-Africa*. Ludwigsburg, Germany: F. Riehm.

Ofelia García. 2019. Decolonizing Foreign, Second, Heritage, and First Languages: Implications for Education. In Donaldo Macedo (ed.), *Decolonizing Foreign Language Education: The Misteaching of English and Other Colonial Languages*, 217–235. Oxford: Taylor & Francis.

Michael Geisler, Claire Kramsch, Scott McGinnis, Peter Patrikis, Mary Louise Pratt, Karin Ryding and Haun Saussy. 2007. Foreign Languages and Higher Education: New Structures for a Changed World: MLA Ad Hoc Committee on Foreign Languages. *Profession*. 234–245.

Patrick Harries. 1988. The Roots of Ethnicity: Discourse and the Politics of Language Construction in South-East Africa. *African Affairs*, 87(346). 25–52.

Johann Ludwig Krapf. 1860. *Travels, Researches, and Missionary Labours, during an Eighteen Years' Residence in Eastern Africa*. London: Trubner.

Johann Ludwig Krapf and Johannes Rebmann. 1887. *A Nika-English Dictionary*. London: Society for the Promotion of Christian Knowledge.

Jouni F. Maho. 1999. *A Comparative Study of Bantu Noun Classes* (Orientalia et Africana Gothoburgensia 13). Goteburg, Germany: Acta Universitatis Gothoburgensis.

Sinfree Makoni. 1998. In the Beginning Was the Missionaries' Word: The European Invention of an African Language: The Case of Shona in Zimbabwe. In Kwesi Kwaa Prah (ed.), *Between Distinction and Extinction: The Harmonisation and Standardisation of African Languages*, 157–64. Witwatersrand, South Africa: Witwatersrand University Press.

Chauncy Maples. 1888. *Yao-English vocabulary*. Zanzibar: Universities' Mission Press.

A. W. McGregor. 1904. *English-Kikuyu vocabulary, compiled for the use of the C.M.S. missions in East Africa*. London: Soc. for the prom. of Christ. Knowledge.

Festo Mkenda. 2018. Jesuits, Protestants, and Africa before the Twentieth Century. In Festo Mkenda and Robert Aleksander Maryks (eds.), *Encounters between Jesuits and Protestants in Africa*, 11–30. Leiden, Netherlands: Brill.

Finex Ndhlovu and Leketi Makalela. 2021. *Decolonising Multilingualism in Africa: Recentering Silenced Voices from the Global South*. Bristol: Multilingual Matters.

Armindo Ngunga. 2001. *Yao Dictionary*. Comparative Bantu OnLine Dictionary (CBOLD). `http://www.cbold.ddl.cnrs.fr/`

Dion Nkomo. 2020. Vernacular Lexicography in African Languages: From Early Days to the Digital Age. *Dictionaries: Journal of the Dictionary Society of North America*, 41(2). 213–43.

Adama Ouane and Christine Glanz. 2010. Why and How Africa Should Invest in African Languages and Multilingual Education. Association for the Development of Education in Africa. Hamburg, Germany: UNESCO Institute for Lifelong Learning.

Steven Paas. 2011. *Johannes Rebmann: A Servant of God in Africa before the Rise of Western*

Colonialism. Nürnberg, Germany: Verlag für Theologie und Religionswissenschaft.

Kristen Parton, Kathleen R. McKeown, James Allan and Enrique Henestroza. 2008. Simultaneous multilingual search for translingual information retrieval. In James G. Shanahan (ed.), *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, 719–728. New York: Association for Computing Machinery.

Doris L. Payne and Leonard Ole-Kotikash. 2008. *Maa Dictionary*. University of Oregon. `https://darkwing.uoregon.edu/~maasai/`

Derek Peterson. 1997. Colonizing Language? Missionaries and Gikuyu Dictionaries, 1904 and 1914. *History in Africa*, 24. 257–72.

George Lawrence Pilkington. 1892. *Luganda-English and English-Luganda Vocabulary*. London: Society for Promoting Christian Knowledge.

Morgan Robinson. 2022. *A Language for the World* (New African Histories). Athens, Ohio: Ohio University Press.

Charles Sacleux. 1939 [1888]. *Dictionnaire Swahili-Français*. Paris: Institut D'Ethnologie.

Gilles-Maurice de Schryver. 2003. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos 13*.

Ray Smith. 2007. An overview of the Tesseract OCR engine. *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. 629–633.

Elizabeth C. Traugott and Richard B. Dasher. 2002. *Regularity in Semantic Change*. Cambridge, UK: Cambridge University Press.