

Citizen-linguists and Decolonial Lexicography: Co-creative Dictionary-building in Grassroots Digital Language Documentation

Anna Luisa Daigneault^{#/^} & Gregory D. S. Anderson[^]

[#]Université de Montréal & [^]Living Tongues Institute for Endangered Languages

annaluisa@livingtongues.org, gdsa@livingtongues.org

ABSTRACT

Many endangered, under-represented, minority and Indigenous language communities around the world need access to multilingual online resources to survive in the digital age. The Living Dictionaries platform provides a collaborative online space for professional linguists and citizen-linguists alike to produce their own grassroots digital dictionaries that include multimedia such as audio recordings and images. These online lexica can play an important role in assisting present and future generations in combatting language loss and creating visibility for their languages and cultures on the Internet.

1 Introduction

While state-run language programs often serve as vectors of total assimilation to dominant languages and the abandonment of

heritage ones (Skutnabb-Kangas, 2000; 2023), grassroots digital projects can serve as a counterbalance and bring visibility to lesser-known languages. Access to high-quality digital resources is essential for language communities in the modern age, as information is increasingly consumed and disseminated digitally, specifically through mobile platforms. Assisting communities in developing such accessible resources is a tangible contribution by linguists in response to the colonialist underpinnings of linguistics. Relying on institutional actors – state, academic, juridical – to act in the interests of linguistic minority communities and to enforce linguistic human rights (not just on paper) has proven to be largely ineffective to date, except in very few contexts where governments have successfully helped revitalize languages that are typically the sole or main minority Indigenous language of the nation (e.g., in Wales, Ireland, New Zealand).

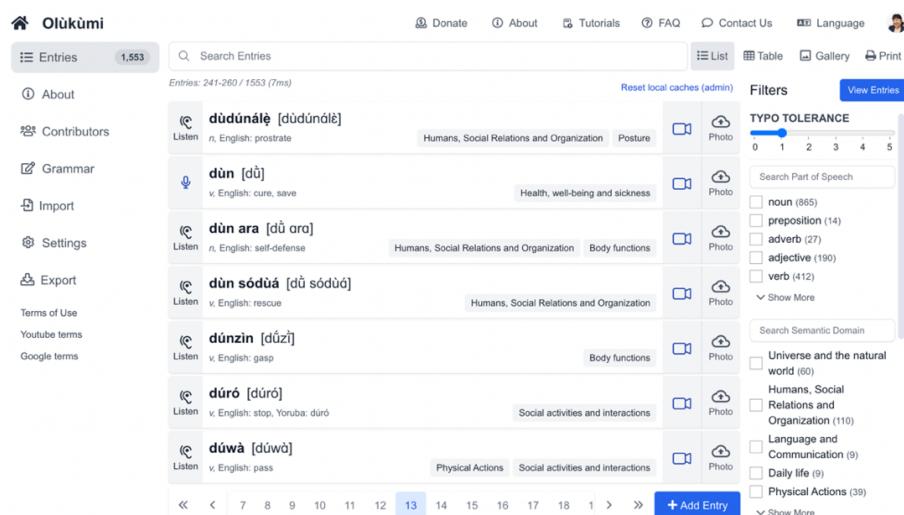


Figure 1: The Living Dictionary for Olùkùmi [ISO 639-3 code: ulb], an endangered Niger-Congo language of Nigeria, with glosses in English and some in Yoruba. It contains 1,553 entries tagged with semantic domains and parts of speech and includes multimedia. It was built by Dr. Bolanle Orokoyo (University of Ilorin) in close collaboration with scholars at Living Tongues Institute for Endangered Languages between 2012 and 2024. <https://livingdictionaries.app/olukumi/entries>

As such, grassroots efforts that combine technology, collaboration and community stewardship provide a meaningful path to combat language loss. This paper shows the global applicability of the Living Dictionaries platform, which was engineered for straightforward co-creation of between grassroots collaborators, to create accessible digital resources for all underrepresented and endangered languages.

2 The Living Dictionaries platform

The Living Dictionaries platform is an online, source-available dictionary-builder that serves a wide range of underrepresented, endangered, Indigenous, creole and diaspora languages around the world, with the goal of providing communities with efficient online access to systematic language materials that benefit language learners as well as scholars. All languages, lects and regional varieties are welcome to be represented on the platform.¹ The Living Dictionaries website is an innovative tool for in-person as well as remote collaboration because it provides an accessible, interoperable² and user-friendly way for community language activists and linguists to work together to document, store and share large amounts of high-quality lexical data paired with digital images, audio, video and GPS coordinates.

¹ When dictionaries are configured, language identifiers such as ISO 639-3 and Glottocodes can be added or updated “Settings page” of the dictionary. These identifiers also help index the dictionary online so that researchers can correlate the content in the dictionary with existing linguistic literature on the language.

² The Living Dictionaries team is working to expand interoperability between data types, formats and software. They can currently import dictionary data from spreadsheets, .CSV files and FLEEx files (Standard

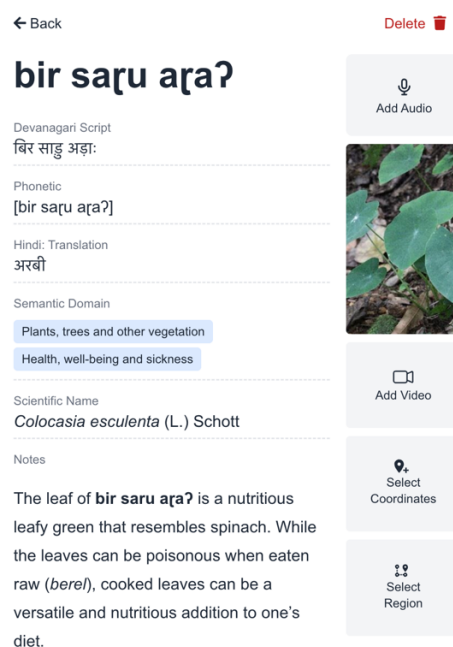


Figure 2: A detailed ethnobotanical entry from the Birhor Living Dictionary, with the headword represented in IPA as well as the Devanagari script (locally dominant in India), the Hindi translation, two semantic domain tags, the scientific name in Latin, an image, and a “notes” section with relevant information about the plant’s culinary use. Birhor [ISO 639-3 code: biy] is an endangered Munda language of India and this project was created by Living Tongues Institute for Endangered Languages in close collaboration with the Birhor tribal community of India.
<https://livingdictionaries.app/birhor/entries>

The platform differs from other digital dictionary programs and online platforms in many distinct ways. For example, it functions in any web browser on any device, so there is no software to download, and any updates to the platform are visible instantly. While designed by linguists and usable by professional linguists, the platform’s functionality is straightforward and easy to learn for citizen-linguists who may not have formal training in linguistics. To date, the Living Dictionaries platform houses

Format), and they aim to improve their existing import functionality by accommodating all legacy dictionary import types (such as formats coming from Toolbox, Shoebox, Lexique Pro, TshwaneLex and other dictionary programs) so that they can import any existing legacy data into the Living Dictionaries. Dictionaries may also be exported as .CSV files and professionally designed .PDFs for printing entire dictionaries. An “Offline mode” is also being developed in 2025.

dictionaries for over 400 languages and has a user interface that is available in fourteen languages, making it accessible to professional linguists and citizen-linguists in many parts of the world. The platform allows citizen-linguists to determine how they want their language to be named in the dictionary's title, and they can edit it at any time. The platform includes multimedia audio and video recording and uploading functionality, which is rare on other platforms. It includes a built-in list of tags for parts of speech and semantic domains, as well as customizable tags, which allow dictionary editors to tag, filter and sort according to their own categories. There is no paywall or fee of any kind to build digital dictionaries on the platform.

Unlike Wiktionary (where each dictionary has different user experience and layout features), every Living Dictionary has the same front-end layout and the same set of systematic linguistic features available to its editors (and they can choose what data fields they wish to fill out). Furthermore, the platform allows for citizen-linguists to configure and modify the "Settings" of their dictionary (including naming, language codes and locations) concerning the language(s) they are working on and decide what glossing languages they wish to include. Another notable feature of the Living Dictionary platform is that it includes geo-mapping for dictionary entries, allowing place names and other entries to be correlated to the dictionary's map.

The screenshot shows a web form for adding a new entry. The title 'adudai' is at the top. Below it, the English translation is provided: 'the Mortlocks (lit. "islands to the west of Nukuoro")'. The form has several sections: 'Part of Speech' with a dropdown menu set to 'noun'; 'Semantic Domain' with a dropdown menu set to 'Place names'; and 'Source' with a text input containing 'Carroll & Soulik 1973' and an 'Add' button. On the right side, there are four buttons: 'Add Audio' (with a microphone icon), 'Upload Photo' (with a cloud and photo icon), 'Add Video' (with a video camera icon), and a blue button with a location pin icon and the text 'Add Map Location'.

Figure 3: An entry for the place name “adudai” (a Nukuoro term that refers to the Mortlock Islands) with its map location in the Living Dictionary for Nukuoro [ISO 639-3 code: nkr], an endangered Austronesian language of the Federated States of Micronesia. It was led by linguist Emily Drummond (UC Berkeley) in close collaboration Nukuoro speakers, with assistance from Living Tongues Institute for Endangered Languages: <https://livingdictionaries.app/nukuoro/entries>

Living Dictionaries are unlimited in size, may contain as many glossing languages as one wants, and may represent entries in up to five local orthographies or scripts, which can be very useful in contexts where there are multiple competing orthographies and users may want to type in the search bar with their preferred script or writing system. Also, the platform allows users to generate and print a professionally designed .PDF of the dictionary directly within the browser. The number of columns, font size, data fields and optional inclusion of images and QR codes (linking directly to the Living Dictionary entry) can all be configured within the ‘Print View’ of each dictionary. With many ‘sorting’ filters on the right-hand side bar, it is easy to generate and print small and/or thematically targeted sets of lexical materials for pedagogical purposes.

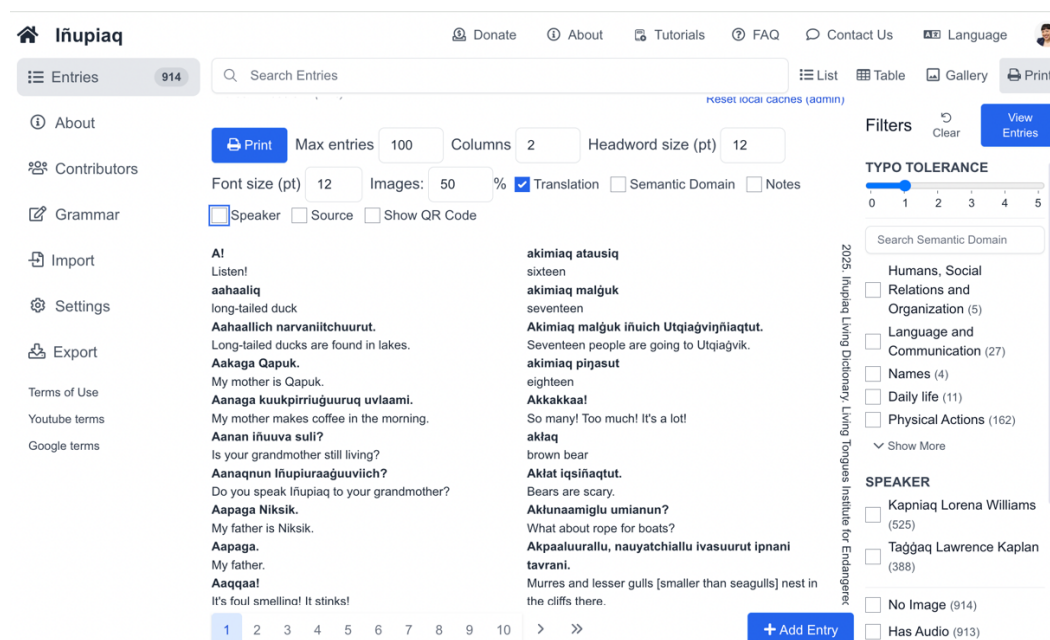


Figure 4: The ‘Print View’ of the Living Dictionary for Iñupiaq [ISO 639-3 code: ipk], an endangered Indigenous language of Alaska. It was created through a close collaboration between three nonprofit organizations: Doyon Foundation (Alaska), 7000 Languages and Living Tongues Institute for Endangered Languages. It is available here: <https://livingdictionaries.app/inupiaq/entries>

3 Co-creation on the Living Dictionaries platform

Collaboration is made easy since dictionary creators can send an automated email invitation to colleagues and assistants who wish to work on the dictionary. Once invited, collaborators join a project and they can add or edit content, record audio, and much more. These tools and features can also help facilitate collaboration and language exchange across generations, as younger, tech-savvy community members partner with experts from middle and elder generations in natural social contexts to record their voices on devices. The ‘History’ tab used for internal workflow makes it easy for various collaborators to edit a Living Dictionary at the same time and visualize what edits have been made and by whom. Through the availability of video tutorials³ on the platform as well as scheduled Zoom trainings organized by the platform’s leadership team, digital skills and best practices in citizen lexicography are made accessible to all users and indeed stand at the core of this initiative.

The platform’s developers, linguists and curators work in close collaboration with citizen-linguists to create new Living Dictionaries; the co-creative workflow in the development of Living Dictionaries consists of five overarching phases: 1) planning and community consultation; 2) digital training, where citizen-linguists receive online or in-person training; 3) data collection and feature programming (if new features are being conceived in collaboration with a community); 4) data assessment and quality control, and lastly, 5) the import of large batches of linguistic and multimedia data to the Living Dictionaries platform and the deployment of new web features. Of course, these phases can be repeated as necessary over the course of a long project. All co-creators have editing access to the linguistic data in spreadsheets that are to be uploaded to the Living Dictionaries and editing access to the online dictionaries themselves. Imported dictionary entries may then be edited on Living Dictionaries manually (one by one) by any project participant directly to the platform or

³ Free video tutorials are available on the platform in English and Spanish (with subtitles in Hindi, Russian, Chinese, Arabic and French) so that prospective dictionary creators can get started

quickly without necessarily having to register for an upcoming Zoom training. The tutorials are available at: <https://livingdictionaries.app/tutorials>

may be uploaded in large batches from spreadsheets in collaboration with the lead digital curator. Once imported, any entry may be edited online at any time, and new entries can also be added. Entries also may include sample sentences with their translations into various languages, semantic domain tags and custom tags that allow for data organization and filtering, one or several audio recordings of the headword's pronunciation by local speakers, and one or multiple images related to the content of the dictionary entry. Entries may include a video that shows a speaker uttering the headword or using it in an example sentence, and some videos may include an explanation of the origin or etymology of the word, a demonstration of an event, etc. Each dictionary entry has its own unique URL, and is thus easily shareable via text messaging, email and social platforms.

Dictionary entries contain a headword represented in the language community's preferred orthography. The web development team works with community researchers to make sure the orthographies are displayed using the best current Unicode-compliant practices for web browsers. If community researchers eventually want to display alternate writing systems, that is not a problem. The system can accommodate up to five alternate orthographies within dictionary entries. The platform sends out regular community messages to all users on the platform several times a year, notifying them about upcoming Zoom trainings, feature updates and scheduled down time.

Living Dictionaries allow citizen-linguists and scholars to create, curate and expand digital lexicons that benefit present and future generations of speakers – locally and across their diasporas. The dictionaries can become large-scale, community-collaborative digital resources, incorporating extensively tagged multimedia materials that allow users to search, filter, sort, export and print data, thus providing language communities with free and shareable resources.

Speakers, regardless of their location, can use the multilingual interface⁴ to record their voices directly to the cloud and store their audio recordings within dictionary entries for easy playback. Audio and video may be recorded directly into the web platform, using any device. The dictionary platform allows dictionary entries to be correlated with maps (helpful for place names and topographic features in the landscape). Users who are working on a language for which there already exists a Living Dictionary can click on the convenient “Contact Us” button located in the top menu toolbar (see the upper portion of Figure 4) to send an automated email message to the dictionary authors and ask for permission to join an existing project as a fellow editor.⁵ Meetings with prospective editors may be convened to have a further discussion about collaboration.

4 Citizen-linguists and the broader impacts of decolonial lexicography

Linguistics is rightly critiqued as rooted in colonialist projects (Errington, 2001; Makoni, 2013; Zimmermann and Kellermeier-Rehbein (eds.), 2015; Hudley et al., (eds.) 2024). It is a moral imperative of the 21st century to address colonialist legacies and thus for linguists to make possible decentralized and decolonial approaches to language documentation, language resource development and linguistic analysis. Indeed, it is not enough to simply acknowledge and critique the colonialist underpinnings of the field, which, while an important step, if not tied to action, remains empty, self-defeating rhetoric. To be sure, documenting languages is not only important to the scientific field of linguistics, but also to speech communities who are urgently looking for tools to combat language loss. It is not an overstatement to assert that language documentation is crucial to conserving humanity's intangible heritage on Earth. Living Dictionaries offer a central point where different communities of stakeholders can contribute equally and respectfully. The platform was created to make

⁴ At the time of writing, the platform's interface can be accessed in fourteen different languages, including Spanish, French, Hindi, Russian, KiSwahili, Bangla, Assamese, Portuguese, Malay, Bahasa Indonesia, Vietnamese, Hebrew, Chinese and English. More interface languages (Arabic, Thai, Italian) are under development.

⁵ Each request is also forwarded to the platform administrators, who can assist the original authors in evaluating the inquiry and deciding if a new colleague should join the project or start a new Living Dictionary of their own instead.

building linguistic resources from the ground up easy and accessible to anyone who knows how to operate a smartphone or tablet. Equitable resource development is at the heart of this work. As platform administrators, we have sought to lessen the burden of colonial and capitalistic frameworks (such as bureaucracy and subscription models) so there are as few limitations as possible for newcomers to lexicography.

In particular, citizen-linguists play a huge role in the process of building these digital dictionaries. A *citizen-linguist*⁶ is here understood as a person who is actively engaged in their speech community, believes in safeguarding their native (or heritage) language and works towards transmitting it to future generations. Citizen-linguists are motivated to create accessible language resources that are tailored to their community's needs. They are people who fulfill the multi-faceted roles of documentarian, language activist and digital content creator, whether they have formal training in linguistics or not. Some are educators, some are students; some are people with advanced training in other academic fields who see the value in protecting their language, whether they are fluent speakers or not.

In general, citizen-linguists take on the difficult challenge of recording and sharing their language, which may be under-studied by scholars and under-valued by their own community and the wider public. Many citizen-linguists are leaders in other areas of community life and undertake language work as volunteers in their spare time, because they understand that the act of preserving their language is connected to their cultural group's

well-being, identity, and survival. Like other citizen scientists who study and celebrate local phenomena (e.g., flora or fauna, etc.), citizen-linguists are grassroots actors who may bridge divides between diverse groups of people and help bring local knowledge forward, while also feeling a sense of personal pride and connection to the language in question. Citizen-linguists are often excellent (co-) creators of language materials because they see the long-term value of the work they are doing and know that teamwork can benefit large, detailed undertakings like language documentation.

Living Dictionaries can serve as community-based access points to under-studied endangered traditional knowledge encoded in languages, in diverse domains such as flora and fauna, textile practices, spiritual traditions, food production, sacred sites and other elements of the landscape, and much more. For example, the Werikyana-Tiriyó-Portuguese-English Living Dictionary (see Figure 5) has over 3,000 entries with accompanying audio recordings and contains an impressive array of content regarding local culture, fauna and flora. This quadrilingual dictionary was created by the various Indigenous communities who speak the Werikyana language in the Brazilian Amazon, using solar-powered mobile devices and satellite Internet via Starlink. It is the first-ever digital resource of this kind for the Werikyana language and represents years of hard work led by many Werikyana speakers. This Living Dictionary was made publicly accessible to the world with the authorization of AIKATUK (Associação Indígena dos Kaxuyana-Tunayana-Kahyana).

⁶ This term, which can also be used interchangeably with other broad designations like *language champion*, *language activist* and *community language activist*, expresses the notion that a member of a local speech

community is contributing to science with their own resources and time, whether they have formal training or not.

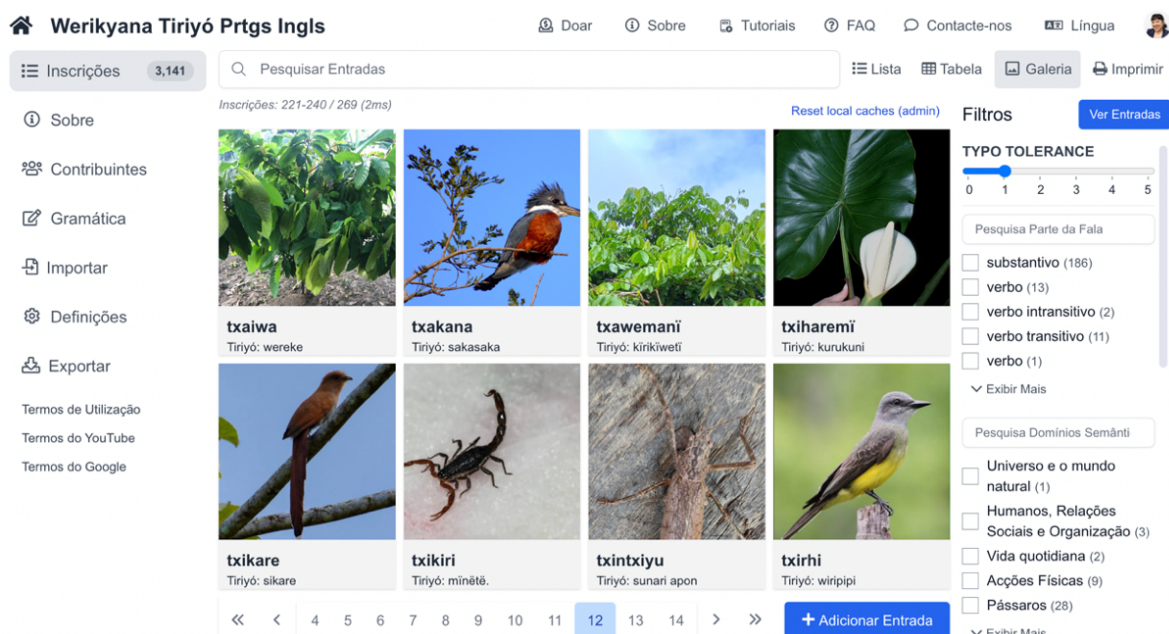


Figure 5: ‘Gallery view’ of the Werikyana-Tiriyó-Portuguese-English Living Dictionary, which was created by Werikyana speakers and Prof. Spike Gildea (University of Oregon) with technical assistance from Living Tongues Institute for Endangered Languages. Werikyana [ISO 639-3 code: kbb] is an endangered Indigenous language of the Brazilian Amazon of the Carib family. It is displayed in the Portuguese language interface in this screenshot (<https://livingdictionaries.app/werikyana/entries>)

The website offers the insights and systematicity of linguistic science in a user-friendly context, free of cost to citizen-linguists, with no institutional or other administrative roadblocks preventing access to this online tool.⁷ It highlights and preserves essential ecological, social, and linguistic knowledge that lie at the foundation of cultural survival. All intellectual property rights associated with dictionary data remain with the language community from which the data originates.

This platform can help mitigate the global language extinction crisis by opening the door to linguistic documentation for all, thereby side-stepping colonial structures that seek to oppress minority languages. The platform supports citizen science by providing STEM opportunities to community activists to document their own languages and help them gain access to technical guidance from our team of professional linguists, anthropologists and web developers. The team also prioritizes racial equity by promoting access to (and awareness of) this platform to diverse communities of color worldwide, particularly in the Global South, with a focus on

supporting academic colleagues and citizen-linguists long-term through digital literacy training online.

This platform provides an easy-to-use framework for systematically storing and sharing dictionary data for at-risk languages, thus increasing their viability for survival in the long-term. This comes with big implications: studies in North America and Australia show that weaving a connection to one’s heritage language leads to better mental health, better performance in schools, and expanded economic opportunity (Zuckerman, 2020; Olko et al., 2022; van Beek, 2016; Olko and Andrason 2023). Therefore, pride in ethnolinguistic identity has numerous socio-economic and psycho-social-political benefits. Living Dictionaries, being multilingual tools at their core, also help promote bilingualism and multilingualism, which, in addition to social benefits, have positive biological outcomes such as improved cognition and protection against the onset of dementia (Bialystock et al., 2007; Perani and Abutalebi, 2015).

⁷ Except for in nation-states that block access to certain websites for political/ideological reasons, like in China.

5 Code Accessibility on the Living Dictionaries platform

Accessible code is also important for future generations of developers building such tools. The Living Dictionaries code base is available on its GitHub page, and its license operates under a source-available, non-commercial license also listed in GitHub. Data in the Living Dictionaries are stored in a PostgreSQL database, backed up daily. Media is stored in a Google Cloud Storage bucket that is also backed up regularly. Dictionary managers can download their dictionaries as a .JSON file for their use in other applications. The .JSON is structured to make it easy for consuming applications to connect all relevant data points such as mapping speakers to dictionary entries. Dictionaries can also be exported as a .PDF file for easy printing. The site relies on popular, open-source technologies that make it easy to maintain and upgrade in the long-term. Two web developers who are specialized in mobile-friendly web applications help the platform's lead digital curator answer technical questions, improve the display and functioning of the platform on mobile devices and desktop computers, fix bugs on the platform, upload batches of content, manage the database, make necessary changes to the front-end and back-end of the website, plan and code new features and keep track of all technical issues on GitHub. They ensure that the code stays source-available at all times and stays up to date with the latest web technologies.

All edits to dictionary entries can be visualized online in real-time, without having to refresh the page, which facilitates instant remote collaboration between dictionary editors who are editing a dictionary at the same time (whether they are side by side or

working at great distances). The design and engineering of the dictionary platform are created on an ongoing basis by an in-house web development team, guided by feedback from hundreds of scholars and citizen-linguists who attend our online workshops which provide training for community researchers involved in the editing, curation, recording process and construction of the dictionaries. The web developers working on Living Dictionaries have designed and implemented new features based on carefully assessing feedback from the user base and incorporating the needs of communities with limited digital literacy. Through engaging images, audio and video recordings, and the ability to add unlimited cultural information in the 'notes' section of each digital dictionary entry and grammatical information about the language in the 'Grammar' tab, the platform is able to showcase the unique features of each language and culture represented in the Living Dictionaries.

6 Summary

Living Dictionaries offer an inclusive, participatory citizen science approach to digital lexicography, thereby helping to decolonize and decentralize the process of language documentation. Grassroots digital language documentation is one of the only realistic approaches to combatting language loss in the long-term, and tools such as this one can benefit different communities of stakeholders. Living Dictionaries are playing an important role in helping under-represented, minority and Indigenous language communities worldwide to successfully claim space in the digital arena and thus safeguard their languages from extinction.

References

- Bialystok, E., Craik, F. I. M., and Freedman, M. 2007. Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, 45(2):459–464.
<https://n.neurology.org/content/81/22/1938>.
- Errington, J. 2001. Colonial linguistics. *Annual Review of Anthropology*, 30(1):19–39.
- Hudley, A. H. C., Mallinson, C., and Bucholtz, M., eds. 2024. *Decolonizing linguistics*. Oxford University Press.
- Makoni, S. B. 2013. An integrationist perspective on colonial linguistics. *Language Sciences*, 35:87–96.
- Olko, J., Lubiewska, K., Maryniak, J., Haimovich, G., de la Cruz, E., Cuahutle Bautista, B., Dexter-Sobkowiak, E., and Iglesias Tepec, H. 2022. The positive relationship between Indigenous language use and community-based well-being in four Nahua ethnic groups in Mexico. *Cultural Diversity and Ethnic Minority Psychology*, 28(1):132–143.
<https://doi.org/10.1037/cdp0000479>.
- Olko, J., and Andrason, A., eds. 2023. Introduction: Heritage languages and the well-being of speakers. In *Heritage languages and the well-being of speakers*, pages 5–16. Linguapax.
<https://www.linguapax.org/uploads/2024/01/Linguapax-2023-baixa.pdf>.
- Perani, D., and Abutalebi, J. 2015. Bilingualism, dementia, cognitive and neural reserve. *Current Opinion in Neurology*, 28(6):618–625.
https://journals.lww.com/co-neurology/Abstract/2015/12000/Bilingualism,_dementia,_cognitive_and_neural.12.aspx.
- Skutnabb-Kangas, T. 2000. *Linguistic genocide in education or worldwide diversity and human rights*. Mahwah, NJ/London: Lawrence Erlbaum Associates.
- Skutnabb-Kangas, T. 2023. Preventing the implementation of linguistic human rights in education. In T. Skutnabb-Kangas and R. Phillipson, eds., *The handbook of linguistic human rights*, pages 109–126. Hoboken, NJ: Wiley-Blackwell.
- van Beek, S. 2016. Intersections: Indigenous language, health and wellness. *First Peoples' Cultural Council*.
- Zimmermann, K., and Kellermeier-Rehbein, B., eds. 2015. *Colonialism and missionary linguistics*. Colonial and Postcolonial Linguistics Vol. 5. Walter de Gruyter GmbH & Co KG.
- Zuckerman, G. 2020. Our ancestors are happy. Language revival and mental health. In G. Zuckerman, ed., *Revivalistics: From the genesis of Israeli to language reclamation in Australia and beyond*, pages 9–36. Oxford University Press.
<https://doi.org/10.1093/oso/9780199812776.003.0009>.