K. DAVID HARRISON is UNESCO Chairholder in Environmental Leadership, Cultural Heritage, and Biodiversity, VinUniversity, and Research Affiliate at the Center for Plants, People, and Culture, New York Botanical Garden, United States ORCID: https://orcid.org/0000-0002-4002-7668

HOÀI TRẦN is Senior Lecturer at the School of Interdisciplinary Sciences and Arts, Vietnam National University, Hanoi, and researcher at UNESCO Chair in Environmental Leadership, Cultural Heritage, and Biodiversity, VinUniversity, Vietnam. ORCID: https://orcid.org/0000-0002-6654-6757

XƠM A is a Bahnar community member and culture expert from Kon K'tu, Vietnam

K. DAVID HARRISON
HOÀI TRẦN
XƠM A

Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction

Indigenous dictionaries have a dual existence: they are simultaneously a community's linguistic heritage and a machine-readable dataset available for ingestion by generative artificial intelligence (AI) systems. A dictionary can be a grift*—an act of extraction from a community, mostly to the benefit of outsiders—or it can be a* gift *that the community creates for itself and then offers to the world on its own terms. AI raises the stakes of that choice considerably. Drawing on fieldwork with three communities—the Bahnar of Vietnam's Central Highlands and the Aneityumese and Futuna-Aniwa speakers of southern Vanuatu—this essay examines what happens when AI encounters Indigenous lexicons. We argue that AI applies a flat ontology to materials that encode a hierarchical and relational one: it cannot distinguish a common word from a sacred name or recognize that certain knowledge is held only by elders or ritual specialists. The result is what we call epistemic flattening; the transformation of living, context-dependent, socially distributed knowledge into retrievable, context-free information. We further*

This is an uncorrected advance copy of an essay that has been formally peer reviewed and accepted for publication in *Journal of American Folklore* 140(556) but has not received final copyediting, typesetting, pagination, or proof corrections. The final published version may differ. The definitive, citable version will be published in Spring 2027.

Terms of Use: This manuscript is protected by copyright held by the Board of Trustees of the University of Illinois and is shared for personal, non-commercial use only. It may not be uploaded to public repositories, commercial sharing sites, or distributed further without explicit permission from the publisher.

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Accepted, Uncorrected Manuscript

argue, however, that the same AI capability creates a methodological opportunity when deployed under community control: semantic clustering of legacy dictionary entries by topic allows the appropriate community expert, the plant specialist, the ritual elder, the river navigator, to guide word collection, validation and recording, respecting rather than bypassing the relational structure of Indigenous knowledge. The difference between AI as extractor and AI as servant lies not in the technology but in ethical lexicographic practice, who controls the process, and to what end.

Keywords (from the AFS Ethnographic Thesaurus):
Artificial intelligence, Bahnar, epistemology, ethnobotany, folk classification, lexicography, Vietnam, Vanuatu

Artificial Intelligence and Lexical Vulnerability

An Indigenous language dictionary has a dual existence. It is simultaneously a community's linguistic heritage and a machine-readable dataset. These two uses coexist in tension. Digitization of Indigenous languages raises questions about access, ownership, and appropriate use, while the application of AI sharpens that tension considerably. AI tools such as ChatGPT-5 or Claude Sonnet 5 can process an Indigenous lexicon in seconds, inferring semantic relationships, constructing thematic clusters, and producing plausible descriptions of concepts, cosmologies, or traditional practices. AI tools do all of this without knowing which words are sacred, which terms may only be spoken in ritual, which knowledge belongs to elders, and which is held differently by men and women.

This essay argues that AI-mediated engagement with Indigenous lexicons creates both epistemic risk and methodological opportunity. The difference between the two lies entirely in who controls the process and to what end. We write as three researchers who are currently leading a community-based lexicography project with the Bahnar community of Vietnam's Central Highlands, which serves as our first case study in this essay.[1] Co-author Xơm A is a community member and native speaker of Bahnar. Our second case study is drawn from Aneityum and Futuna Islands in Vanuatu where co-author K. David Harrison worked with eight communities to build Talking Dictionaries. Drawing on our fieldwork in Vietnam and Vanuatu, we argue that a dictionary can be seen as a *grift*—an act of extraction from a community, mostly to the benefit of outsiders—or it can be a *gift* that the community creates for itself and offers to the world on its own terms. AI raises the stakes of that choice, making the distinction between AI as extractor and AI as servant more consequential than ever for Indigenous language documentation and folklore studies.

Dictionaries as Knowledge Ecosystems

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

To understand what AI does when it ingests an Indigenous dictionary, you first need to understand what a dictionary contains. The answer, especially in the case of languages spoken in oral, nature-based cultures, is considerably more than a lexico-semantic inventory with definitions.

Consider two Bahnar-French legacy dictionaries: the earlier one compiled by the Catholic missionary Pierre X. Dourisboure (1889) with 363 pages and ca. 3,600 entries, and the later work by French colonial administrator turned ethnographer Paul Guilleminet with Catholic priest Jules Alberty (1959–63) with 991 pages and over 5,000 entries. An entry in either of these dictionaries—*hyam* (n.; a tree), *xakeng* (n.; a type of black fish), *Yang Dak* (n.; the spirit of the water)—is not merely a lexical item. Each carries layers of meaning: folk taxonomic information, seasonal and ecological context, ritual significance, spiritual status, restrictions on who may use the word and in what circumstances. Knowing the *hyam* tree [*Engelhardia roxburghiana* Lindl.] entails knowing bark harvesting techniques, traditional *sik* (rice wine) fermentation, and the *yang* (forest spirits) tree guardians. A landscape term *adral* (n.) denotes a "boundary between two rice fields" but also signals the respective owners' reciprocal duty to demarcate that boundary. The Bahnar dictionaries, read this way, are an archive, a knowledge base of ecological, ritual, and social knowledge encoded at the lexical and phrasal levels.

We apply Houde's (2007) framework which identifies multiple dimensions of Indigenous knowledge systems—ecological, cultural, spiritual, and societal—and offers insights into sustainable practices, environmental stewardship, and community-based management. A single lexical entry may carry meaning across all these dimensions simultaneously. The Bahnar tree name *hyam* necessitates knowledge about sustainable bark harvesting, the symbiotic relationship between bark and yeast organisms, technologies for brewing rice wine, Bahnar concepts of healthy foodways, sacred landscapes, and obligations owed to forest spirits. Our Bahnar language consultants, former Chief Bên A (born 1961) (fig. 1) and his nephew Xơm A (born 1996) (fig. 2), both possess deep knowledge of the landscape, agricultural practices, and ritual life of their community. They demonstrated their expertise vividly in sessions where a single word could open into extended explanations touching on forest ecology, family history, and sacred obligation in the same breath. Our interviews with Bên A (fig. 1) covered in detail topics such as the more than 20 kinds of baskets Bahnar men weave (fig.e 2), the complex vernacular architecture of the *rong* common house, and the more than two dozen forest plants collected to make houses and baskets (Trần et al. 2024a). Participatory activities with Xơm A revealed the currents one must know to safely pole a dugout canoe on the Đăk Bla river (Trần et al. 2025), the hidden pathways from the village to the rice pond fields, bird sounds that serve as forest omens (Trần et al. 2024b), edible insects and snails, and diverse Bahnar technologies for growing rice (Harrison et al. 2026).

Multidimensionality of Bahnar wor(l)ds is fundamental, not a lexicographic artefact. It reflects how knowledge is transmitted in the Bahnar community, which uses writing occasionally but remains a primary oral society. Bahnar environmental knowledge circulates within both a

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

hierarchical and a relational ontology. Knowledge is not simply possessed; it is held differentially by acknowledged experts. Some general knowledge is shared by everyone; some is gender-specific; some belongs to age cohorts or ritual roles; some is sacred and context-restricted, valid only when spoken by the right person, in the right ceremony, in the right circumstances. Knowledge does not exist independently of the social relationships and contexts of use that give it meaning and power.

The Bahnar-English Talking Dictionary that we are now compiling (Trần et al. 2023), even if done with the utmost care and sensitivity to which we aspire, can only partially reflect this knowledge structure. Entries may note that a word is connected to ritual, or flag a term as associated with spirits, but the notation is inevitably a flattening. The word is recorded, but the web of relationships, restrictions, and contexts that constitute its meaning in practice cannot be fully transferred to the printed or digital page. Our multimedia (text, audio, photography, video) dictionary is already, in a sense, a limited translation of a living knowledge system into a static representational form. And, as with all dictionary projects, it will never contain the complete lexicon of Bahnar, nor the world view it expresses.

The inadequacy of standard lexicographic formats for capturing multidimensional, relationally organized knowledge is not a new observation. Anthropologist Harold Conklin, whose foundational dissertation *The Relation of Hanunóo Culture to the Plant World* (1954) transformed ethnobotany from a cataloguing exercise into the linguistically informed study of a cultural knowledge system, argued that the lexicographical treatment of folk taxonomies requires formats fundamentally different from those designed for European alphabetical dictionaries (Conklin 1962). He shows that to impose the latter on the former is not merely a practical inconvenience but an epistemological distortion. Writing at a time when cognitive anthropology was establishing that Indigenous naming systems encode entire worldviews, Conklin diagnosed precisely the problem this essay addresses: that the organizational logic of the outside scholar's reference tool and the relational logic of Indigenous knowledge are structurally incompatible. The Bahnar experience we describe above—in which alphabetical presentation of words produced halting, hesitant responses from expert knowledge holders, while topically organized word lists produced animated, expansive ones—directly illustrates the dilemma Conklin identified over 60 years ago. The stakes, however, have changed considerably. The colonial dictionaries Conklin's generation critiqued were consulted by scholars one at a time, and often in collaboration with the community of origin. But today, digitized lexicons can be ingested by AI systems in seconds, making the epistemological distortions he described not merely analytically troubling but ethically urgent. This critical tradition has recently found an unexpected institutional echo: the École française d'Extrême-Orient—the very institution that published the Guilleminet and Alberty Bahnar-French dictionary we discuss—has itself begun to reckon with the colonial lexicographic legacy it helped create, with a recent volume examining how European dictionaries of Asian languages from the sixteenth to nineteenth centuries were "not neutral linguistic repositories but rather embedded in systems of power" (Bussotti and Lachaud 2023:21). That the institution responsible for producing the Bahnar-French dictionary now

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

frames such projects as instruments of colonial power rather than neutral scholarship marks a significant, if belated, disciplinary turning.

In Vietnam, site of our first case study, scholars have approached minority language dictionaries not merely as lexical reference tools but as "living museums" that preserve Indigenous knowledge and ethnic cultural identities. This perspective aligns closely with the notion of "knowledge ecosystems," in which each lexical entry functions as a node connecting language, cultural practices, and ethnobiological knowledge. Indeed, seminal works by Hữu Hoành Nguyễn (2013) and by Văn Lợi Nguyễn and Hữu Hoành Nguyễn (1998) laid important foundations for the systematization of minority languages such as Katu, Mảng, and Ruc, thereby illuminating deeper layers of meaning related to origin narratives and typological features. In the field of lexicography, Văn Thông Tạ (2018) emphasizes that Vietnamese–minority bilingual dictionaries are essential instruments for mitigating language loss, while also serving as repositories of cultural forms that are gradually disappearing. Dictionaries such as the *Tày–Vietnamese Dictionary* by Bèn Lương et al. (2011) are valued for their ability to convey rich cultural contexts through lexical definitions embedded in everyday livelihoods and local worldviews.

Recent Vietnamese lexicography is increasingly oriented toward digitization and the development of multimedia corpora. Thị Mỹ Lệ Hoàng and Duy Linh Trần (2024) propose a four-stage processing model for the Hrê and Cua languages and explore the application of blockchain technology to ensure the integrity of linguistic data. Such innovations reconfigure dictionaries as dynamic, interactive systems that support both bilingual education and cultural preservation. Accordingly, dictionary-making in Vietnam is inherently interdisciplinary, integrating anthropology, history, and linguistics, and grounded in long-term ethnographic engagement with community knowledge holders. However, as the Bahnar case demonstrates, the transformation of relational, context-dependent, and socially stratified knowledge systems into lexicographic formats, whether print or digital, inevitably entails processes of reduction and flattening.

How Can Generative AI Assist in Reading a Dictionary?

Our project leverages AI tools to enhance human ability to extract lexical meaning and patterns from dictionaries, in the service of community-based language maintenance. To assess the usefulness of AI tools for this task, we scanned the legacy dictionaries and structured the data as tab-delimited columns labeled by language; Bahnar, French, and English respectively. These labeled columns were fed to ChatGPT (version 5). Multiple methods were applied in parallel to analyze and visualize the digitized lexicon. Word clouds and frequency tables were generated from both the Bahnar headwords and their English and French equivalents, revealing patterns in the vocabulary that were not visible from the dictionary entries alone. Vector maps—visual diagrams that plot words and concepts according to their semantic proximity to one another—were generated using the SentenceTransformers library (Reimers and Gurevych 2019) and the

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

MPNet transformer architecture (Song et al. 2020), which encoded the English definitions into numerical representations capturing their meaning and relationships. A third workstream used ChatGPT prompts to extract thematic sets—environmental vocabulary, kinship terms, ritual language—using the language's own ontological categories (e.g., animal, plant, forest) as organizing frameworks, rather than imposing external classificatory systems (Ye and Goddard 2026). Throughout all stages of our research, findings were shared with and validated by native Bahnar speakers, including Xơm A who is a coauthor of the present essay. These parallel, complementary datasets were then brought together to build a richer and more culturally grounded picture of the Bahnar lexicon than any single method could have produced. Full methodological details are provided in Harrison et al. (2026).

When we use AI tools to assist us in reading or creating a dictionary, the tools apply what we will call a flat ontology. All entries are equivalent data points, equally accessible, retrievable, and available for inference and combination. The tools can identify semantic clusters, infer relationships between terms, and generate coherent-sounding descriptions of the knowledge system the dictionary represents. AI is, in a narrow technical sense, good at this task, certainly surpassing what a print dictionary can offer in terms of semantic clustering.

But AI cannot know what the relational ontology of Indigenous knowledge requires it to know, or that Indigenous naming systems may have categories that are both different and more precise than scientific taxonomies (Davis et al. 2007). We infer that AI, working only from headwords and translation equivalents, lacks the contextual and cultural knowledge necessary to distinguish a plant name from a sacred name. It has no way of identifying that a given word is used only by women, or that a certain utterance is the exclusive domain of the *pojâu* (shaman, ritual specialist). Whether AI, given richer input such as extended narratives, ritual texts, or stories, might detect such patterns remains an open question. For primary oral languages like Bahnar, however, such corpora do not yet exist, which returns us to the recursive problem described above. AI processes lexemes according to the logic of data; without the embodied, socially distributed knowledge that Indigenous experts carry and that constitutes the actual meaning of the terms.

The consequences are not merely technical errors. Consider what an AI system does with legacy Bahnar lexicons. We digitized dictionaries by Guilleminet and Alberty (1959–63) and Dourisboure (1889), both of which contain many entries relating to forest spirits, landscapes, animals, and plants. We simplified the dictionaries' structure while digitizing, such that the input to ChatGPT was reduced to just (1) the Bahnar headword, and (2) the French or English definition. Information such as example sentences (with basic or metaphorical uses of a given term), but which were not glossed, or cross-references to other entries, could not feasibly be included at this stage of digitization; which we acknowledge as a limitation. A future iteration could do more to link related definitions, and to attempt (semi-)automated interlinear glossing of example sentences. ChatGPT instantly maps semantic similarities among these entries,

Accepted: Unproofed Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

categorizes them, and can generate a narrative about Bahnar cosmology, one that seems internally consistent, and plausible, but is entirely the AI tool's construction.

As a test, we asked ChatGPT to generate a narrative about Bahnar culture and cosmology to see what the model would do when given the digitized lexical data as its sole source.[2] But the resulting narrative was not a true representation of Bahnar culture; it was filtered through the model's tendency to confabulate and to produce authoritative-sounding output regardless of the limitations of its input (Ji et al. 2023; Farquhar et al. 2024). There is a deeper issue here than technical limitation. AI-generated text is seductive precisely because it mimics the surface features of human expression—fluency, coherence, confidence—while remaining indifferent to the cultural registers, relational structures, and community contexts that give an expression its meaning. The danger is not only that AI produces inaccurate accounts, but that it produces inaccurate accounts that seem authentic. Xơm A, a coauthor of this essay and a native Bahnar speaker, evaluated the narratives generated by ChatGPT and judged them to be inaccurate representations of Bahnar culture and cosmology. We have therefore chosen not to reproduce them here. To do so would risk circulating an AI-generated account that is not community validated, and that may flatten or distort dimensions of knowledge that the dictionaries themselves could not adequately capture. The decision not to reproduce the narrative is not an omission; it is an application of the ethical principles this essay argues for.

AI-generated content thus represents an accuracy problem. It also recalls an ethical dilemma long central to folklore studies. What happens when the knowledge is separated from its knower, devoid of social context, and the community loses the right to control its transmission? Folklorists have long insisted on precisely this point. Richard Bauman's foundational argument that verbal art is inseparable from its performance event established context as constitutive rather than incidental to folkloric knowledge (Bauman 1977). From this perspective, a tale cannot be separated from its telling, sacred speech is only potent in its proper context, and the right to tell is constitutive of the story itself. All these principles may be undermined when AI generates an account of a community's knowledge from a digitized word list. AI, in this mode, risks becoming the interpreter that replaces the village elder: seemingly fluent and confident, yet wrong in ways that matter.

We conclude that the rich content of Bahnar legacy dictionaries is for now only partially accessible to computational processing. Dourisboure, as well as Guilleminet and Alberty, were not mere word-collectors. They lived among the Bahnar for decades, achieved fluency, and their dictionaries reflect that immersion in a wealth of example sentences, metaphorical or idiomatic expressions, and explicit conceptual links between entries. A careful reader can discern something approaching a web of knowledge, not merely an alphabetical inventory.

In the present project, however, we were constrained to extract only headwords and their French or English equivalents. The example sentences carry the most contextual and cultural value but could not be incorporated because they lack interlinear glosses. Aligning them to create

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

a usable parallel corpus would require translation models that do not yet exist for Bahnar. This limitation reflects a challenge shared by almost all primary oral languages. The computational tools needed to unlock the full depth of the legacy dictionaries cannot be built without large parallel corpora. And the materials needed to construct those corpora exist, in incipient form, inside the very dictionaries whose depths we cannot yet fully access.

Efforts are underway to improve digitization and corpora building for oral languages. The SEA-LION project (Ong and Limkonchotiwat 2023) is building language models for 11 Southeast Asian languages that already have substantial text corpora awaiting fuller digitization, while other projects partner directly with primary oral communities in South Asia (Reitmaier et al. 2024) to develop speech and language technologies without relying on written texts at all. This distinction matters because the technology industry collapses these scenarios under a single label, "low resource," meaning a language lacks enough digitized text or annotated data to train effective language models or translation systems. That description fits two very different situations. Filipino, Vietnamese, Lao, and Burmese are "low resource" only in a technical or temporary sense. They have tens of millions of speakers, established literary traditions, and substantial text corpora that have simply not yet been fully ingested into language models. Bahnar belongs to a different category. Bahnar functions as a primary oral language, and this is its normal state, not a deficiency, as is true of most of the world's languages. Bahnar written records consist almost entirely of a partial Bible translation, a hymnal, and three legacy dictionaries. The "low resource" label, applied indiscriminately across both categories, performs its own kind of flattening. It treats a pipeline problem as equivalent to an ontological one, and risks generating the same extractive approaches for primary oral communities that this essay critiques more broadly.

What ChatGPT received was not the Bahnar knowledge system, nor even the legacy dictionaries in their full complexity. We provided merely a partial extraction consisting of headwords and equivalents. And the epistemic flattening the AI produced was not only inherited from colonial archival materials. It was compounded at the digitization stage by structural constraints, and then further compounded by AI's observed tendency to produce authoritative output or even confabulations, regardless of the limitations of its input (Ji et al. 2023; Farquhar et al. 2024).

There is an additional dimension that the Bahnar case makes visible: the colonial inheritance of the dictionaries themselves. Both Dourisboure (1889) and Guilleminet and Alberty (1959–63) were compiled in the French missionary tradition; not designed for the Bahnar community but for outsiders: linguists, administrators, Bible translators. They reflect outsider perspectives: Dourisboure's dictionary labels certain spirit names as "Superstition," while certain plants of ritual or nutritional importance are simply glossed as "a kind of plant," presumably because the dictionary compiler lacked the botanical expertise to identify plants. Community members contributed their knowledge but were not named as coauthors; the dictionaries were published in Hong Kong and Paris, were expensive and difficult to obtain, and never found their

Accepted, uncorrected manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

way back to the Bahnar people whose language they documented. Copyright was held by authors and publishers, not the community. When we as researchers digitize legacy dictionaries and feed them to AI systems, rather than neutralizing these colonial power dynamics, we risk prolonging them.

Returning the Dictionary

One of the first obligations of any linguist or outside scholar undertaking language documentation work is to locate existing legacy resources and to bring copies back to the community. This is not merely a courtesy. It is a recognition that any prior dictionary was built of the community's knowledge, and that community members should have access to and authority over it. An act of return can itself be a significant moment in the relationship between researcher and community.

When Harrison brought copies of the legacy Aneityumese-English dictionary, compiled by Scottish missionary John Inglis (1882), to the community of Anelcauhat on Aneityum Island in Vanuatu, one community member responded emotionally: "Thank you for returning our language to us." Others browsed through the dictionary with evident joy and interest, noting obsolete forms, laughing at unfamiliar spellings, and commenting on how the language had changed. Aneityum remains a primarily oral community; written texts are known and used, but literacy practices are not central to daily life. The dictionary was, for most people, a curiosity and a historical record; interesting, even moving, but not seen as authoritative.

The Aneityum community are avidly practicing folk lexicography. Kirk Keitadi, an elementary school teacher, told how he had independently compiled the "100 most important words", such as *nelco* ("outrigger canoe," also metaphorically "family"), and had been teaching his pupils from it. David Nasauman shared his handwritten notebooks of lexicographic information, including gems like the "sugar cane calendar," an environmental calendar (Balick et al. 2023) organized around plant phenology, precisely the kind of multidimensional ecological knowledge that a standard alphabetical dictionary tends to omit. And Spokesman to Chief Simon of Anelcauhat Reuben Neriam, while modestly describing himself as "not a plant expert," identified nearly 1,000 local plant species and recorded their vernacular names (Plunkett et al. 2024, Balick et al. 2025). He also told traditional stories of sun worship (Harrison et al. 2025) and weather magic (Balick et al. 2022), each containing unique lexical items that encode the Aneityumese cosmology. These individuals were not awaiting outside expertise; they were already engaged in documentating and transmitting their language guided by their own priorities. And they were willing to share their knowledge by coauthoring the Aneityum Talking Dictionary (Harrison et al. 2015).

On Futuna Island, Vanuatu, our research team brought copies of the Futuna dictionary authored by Janet Dougherty (1983) to the community. They were happy to receive it. Some had known the dictionary existed and that elders had worked on it but had never seen the finished

Accepted, Unproofed Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

product. One commented that the author must have made money selling it, while the community received nothing. Communities may believe that benefits flow to authors or publishers, thus contributing to the perception of dictionary as extractive, a *grift*. Futuna Islanders examined a sketch of the wind compass included in the dictionary and identified what they perceived as gaps or errors. At a *toktok* (community meeting) elder Yaugane Misikofe (born 1960), carefully redrew the map and wind compass according to his current understanding, while teaching it to younger community members, including Cliffton Napapo and Nigasau Zeiake, who had only partial knowledge of it (Harrison et al. 2024). The moment was instructive: the dictionary had preserved a version of the compass, but the living authority over that knowledge resided with the elder, not the tome. Community members then engaged actively in creating a new Futuna-Aniwa Talking Dictionary (Harrison et al. 2016), recording several thousand words organized into semantic categories—fish, kinship terms, medicinal plants, weather magic—that reflected their own sense of how knowledge was organized, not the alphabetical logic of a legacy resource.

AI as Servant: The Case for Community-Controlled Lexicography

The same AI capability that creates the risks we have described also offers a useful tool for community-driven dictionary creation, if applied under community control and in service of their own language revitalization agenda. The Talking Dictionary platform, created by coauthor Harrison, now hosts over 200 endangered languages and represents a different approach to lexicographic documentation than the legacy dictionaries we have been discussing. Where those dictionaries were produced by outside scholars for outside audiences, the Talking Dictionary is open-ended, community coauthored, and community validated. Entries may include text, audio recordings, photographs, video, and even social media posts (Lillehaugen et al. 2019). The platform is designed to include dialect and orthographic variation that printed dictionaries often omit. The resulting resource is, in Clifford Geertz's terms, more "thick" than flat (Geertz 1973). It can accommodate register restrictions, expert knowledge, contextual annotations, and the kind of relational information that links a word to the context in which it is spoken. The platform is widely used by Indigenous and minority language communities for language revitalization and is freely accessible online.

The ChatGPT ontology, despite the limitations we have described, proved productive in one specific and helpful way: it generated thematic word lists—terms grouped by semantic domain—that could be taken to Bahnar specialists for correction, expansion, and validation. This process of AI-assisted elicitation, with native speakers fully in the loop at every stage, is precisely the "AI as servant" (rather than master) model we advocate. The thematic groupings gave specialists a concrete starting point for discussion, prompting the kind of contextual, relational knowledge that a direct request to "tell us about plants" or "tell us about spirits" might not have elicited as readily. The full methodology of this elicitation process is described in detail in Harrison et al. (2026). Going forward, we intend to continue using AI tools in the development of the Bahnar-English Talking Dictionary, with human collaborators—including native speakers—fully in the loop at every stage. We can use AI to detect connections and

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

patterns that accumulate as the dictionary grows, while retaining community authority over the meanings of those connections.

A central methodological challenge of any community dictionary project involving a legacy resource follows directly from this mismatch between alphabetical and relational organization: outside scholars can look up a word by letter, but community experts hold knowledge by domain, season, social role, and ecological context, not by initial letter. In our Bahnar work, alphabetical order produced halting, uncertain responses from Bên A and Xơm A, while presenting the Guilleminet and Alberty dictionary in topical groupings—rice vocabulary, bird checklists, forest-spirit names, buffalo-sacrifice terms—made the same consultants animated and expansive. The effect was strengthened further by pairing sessions with physical artifacts such as baskets or canoes, which served as memory prompts.





Figure 1: Bên A, Bahnar dictionary coauthor and culture expert, with a traditional Bahnar *plung* dugout canoe at the Đăk Bla river. Photo by the authors.

Figure 2: Xơm A (left), Bahnar dictionary coauthor and culture expert, observes Jo A (right) weaving a traditional Bahnar *reo* basket. Photo by the authors.

Accepted, Unedited Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05



Figure 3: Bahnar culture expert Banh A (left) working with researcher C. M. K. Hoàng (right) to validate a bird species checklist compiled from citizen science platform iNaturalist. Photo by the authors.

AI semantic clustering offers a solution to this problem. A generative AI system can process several thousand dictionary entries and sort them into thematic categories: plants, animals, fish, agricultural terms, kinship vocabulary, weather and environmental phenomena, terms related to ritual and ceremony. Some of these categories were suggested by AI; we had not thought to ask for them. This does not require the AI to understand the knowledge—it requires only that it identifies semantic proximity, which is precisely what AI systems do well. The output is a set of organized word lists that can then be brought to the appropriate community expert: the plant list to the person who knows plants, the fishing vocabulary to the specialist in marine knowledge, the ritual terms to the elder with authority over ceremonial life. Used in this way, AI reveals the hierarchical and relational ontology that it would otherwise flatten. The semantic clustering merely organizes, it does not interpret or generate. The interpretation and validation of vocabulary items remain entirely with knowledgeable community members. The elder who reviews the list of ritual terms can identify which are sacred, which may be recorded, which should be withheld. The plant specialist can correct errors, add detail, provide the multi-dimensional context—ecological, cultural, spiritual, social—that a legacy dictionary compresses into a minimal gloss. AI does the sorting; the human expert does the knowing.

This distinction between AI as extractor and AI as servant maps onto a more fundamental question about who controls the process. When AI ingests an Indigenous dictionary and generates interpretations, the community has no role in determining what is retrieved, what is combined, what is produced, or how it is framed. When AI is used to sort a legacy dictionary

Accepted Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

into semantic categories as a first step in a community-led documentation project, the community retains full control over what happens next. The difference is not in the technology but in the structure of the relationship.

The Talking Dictionary project for Futuna-Aniwa exemplifies what this looks like in practice. Community members recorded thousands of words grouped by semantic category (e.g., fish names). The organization was determined in consultation with the community, reflecting their sense of how knowledge was appropriately arranged. Speakers with expertise in particular domains recorded the relevant vocabulary. The elder's redrawing of the wind compass, correcting what he saw as mistakes in the legacy published version, and explaining the system to younger people who had only partial knowledge, became the record, not a marginal annotation to someone else's authority.

Ethics, Ownership, and the Future of Community-Based Lexicography

The questions raised by AI-mediated engagement with Indigenous languages are not merely technical questions. They are questions about knowledge, ownership, orality, and the right to (re)interpret. Diverse disciplines dealing with language as cultural heritage have grappled with these questions. Nora and Richard Dauenhauer (1987) characterize Tlingit oral literature as "a treasury of spiritual gifts from which we can draw in times of need" (Dauenhauer and Dauenhauer 1987:ix). Yet they also observe that oral and written literatures have "widely differing and often conflicting aesthetics and rules for composition and publication," which presents challenges for scholars who seek to transmute the former into the latter. Richard Bauman and Charles Briggs (1990) introduce the concept of *entextualization*, the process by which discourse is extracted from its performative (oral) context and reduced to a text. Lexicography necessarily relies on both the performative and metalinguistic skills of the speaker, who in the process of creating a dictionary is asked to perform the atypical task of defining and translating individual words, outside of their usual discursive contexts. Considering ownership of knowledge, Kimberly Christen's work on tribal archives and digital cultural heritage shows how colonial collecting projects have "unhinged" knowledge, resulting in it being "removed from communities and detached from local knowledge systems" (2015:2), a dynamic that digitization and AI now risk repeating at scale. Victoria Sear and Mark Turin (2021) show that many contemporary users of early dictionaries of North American languages continue to grapple with how to utilize these resources for contemporary linguistic understanding, due to the disconnect between the extractive methods used to create these collections and the goals of living language communities. Steven Bird (2024) makes a crucial distinction between an *extractive* paradigm, in which language data is harvested from communities as raw material for computational systems, and a *relational* paradigm, in which technology is developed in partnership with speech communities, and respectful of their own linguistic ideologies and priorities. These ongoing debates about the ethics of fieldwork, the politics of archiving, and the obligations of the outside scholar who transports community knowledge into academic contexts are germane to the

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

questions we raise in this essay. The capabilities of generative AI give these questions new urgency.

Several issues merit attention from the field. First, the question of consent and control: who has the authority to decide whether and how a digitized Indigenous dictionary is fed into an AI system? If a missionary dictionary compiled in the nineteenth century is digitized by a national library and made freely available, the community whose language it documents may have no practical ability to prevent it from being ingested by AI systems and used to generate content about their culture. The CARE Principles for Indigenous Data Governance (Carroll et al. 2020) provide a helpful framework for thinking about these obligations, but the practical mechanisms for ensuring community authority over digitized legacy materials remain underdeveloped.

Second, the question of interpretive authority: once AI generates an account of a community's cosmology or traditional botanical knowledge, that account takes on an authority that is difficult to contest. It may circulate in contexts the community cannot monitor, be cited, and even become part of the putative record of what the community believes or practices. The risk of AI-generated accounts displacing authentic community voices is not hypothetical, nor is it a bug. It is a structural feature of AI-generated content.

Third, the field needs clearer guidelines for distinguishing between legitimate and illegitimate uses of AI in relation to Indigenous knowledge. The distinction we have proposed between AI as extractor and AI as servant is a starting point, but it needs elaboration and operationalization. What protocols should govern the use of AI tools in language documentation projects? What does meaningful community control look like when AI is part of the methodology? How do we build AI-assisted lexicographic workflows that vest interpretive authority with community members?

A digitized Indigenous language dictionary can be a grift, or a gift. The nineteenth-century Aneityumese dictionary became, when returned to the community, an occasion for reflection, comparison, and the recognition of change, a resource that community members like Kirk Keitadi and David Nasauman engaged with critically and on their own terms. The Bahnar dictionaries of Dourisboure (1889) and Guilleminet and Alberty (1959–63) carry within them the knowledge of ancestors of today's Bahnar speakers who contributed to them, possibly without credit or compensation. Fed into an AI system by a party with no relationship to the community, those same dictionaries become a dataset from which knowledge can be retrieved without permission, interpreted without community validation, and circulated without reciprocity. On a positive note, communities and linguists working together have shown that dictionaries can be showcased as objects of pride alongside material and performative culture in a folkloric setting. In the "One World, Many Voices" program at the Smithsonian Folklife Festival (Harrison and Hunt 2013), four of the invited delegations included Indigenous lexicographers who proudly showcased their community's in-progress dictionary projects: Garifuna (Reyes 2012, n.d.), Koro-

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Aka (Anderson and Harrison 2012), Siletz Dee-ni (Anderson and Harrison 2007), and Tuvan (Harrison et al. 2007). As the case studies we consider here amply attest, lexicography is a valued mode of cultural activism and heritage preservation. Dictionaries can be powerful symbols of community pride and cultural artifacts with multiple stakeholders (Wicherkiewicz 2025).

The problems this essay describes are not caused by AI. Epistemic flattening, extractive collection, and the detachment of knowledge from its community context are as old as the colonial dictionary. What AI does is raise the stakes: it accelerates, amplifies, and lends new authority to these processes, and it places their outputs within reach of anyone with an internet connection. The ethical obligations this essay argues for are not new, but are newly urgent. Folklore and cultural studies, which direct our attention to the politics of representation, the ethics of fieldwork, and the rights of communities over their expressive culture, are well placed to guide the conversation about what responsible AI-assisted lexicography looks like. AI does not determine the outcome, we do, and we can ensure that the communities whose knowledge is at stake remain at the center of it.

Acknowledgments

The authors thank the Bahnar community for their generosity in sharing their knowledge; and acknowledge funding from a VinUniversity grassroots grant (PI Myles Lynch), and an Explorers Club expedition grant (PI K. David Harrison).

References Cited

Anderson, Gregory D.S., and K. David Harrison. 2012. *Koro Aka Talking Dictionary*. Living Tongues Institute for Endangered Languages and Swarthmore College. Access this resource at http://www.talkingdictionary.org/koro (accessed March 2026).

Anderson, Gregory D.S., and K. David Harrison. 2007. *Siletz Dee-ni Talking Dictionary*. Confederated Tribes of Siletz, Living Tongues Institute for Endangered Languages and Swarthmore College. Access this resource at http://siletz.talkingdictionary.org (accessed March 2026).

Balick, Michael J., K. David Harrison, Neal Kelso, et al. 2022. Weather Magic as Environmental Knowledge in Southern Vanuatu. *Journal of Ethnobiology* 42(4):383–99.

Balick, Michael J., Gregory M. Plunkett, K. David Harrison, et al. 2023. Calendar Plants in Southern Vanuatu. *Economic Botany* 77:227–42.

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Balick, Michael J., Gregory M. Plunkett, Jean-Pascal Wahe, et al. 2025. *Plants, People and Culture in Tafea Province, Vanuatu.* 2 vols. New York Botanical Garden.

Bauman, Richard, and Charles L. Briggs. 1990. Poetics and Performance as Critical Perspectives on Language and Social Life. *Annual Review of Anthropology* 19:59–88.

Bauman, Richard. 1977. *Verbal Art as Performance.* Waveland Press.

Bird, Steven. 2024. Must NLP be Extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Bussotti, Michela, and François Lachaud, eds. 2023. *Mastering Languages, Taming the World: The Production and Circulation of European Dictionaries and Lexicons of Asian Languages (16th–19th Centuries).* École française d'Extrême-Orient.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodriguez, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* 19(1):43.

Christen, Kimberly. 2015. Tribal Archives, Traditional Knowledge, and Local Contexts: Why the "s" Matters. *Journal of Western Archives* 6(1):1–19.

Conklin, Harold C. 1954. *The Relation of Hanunóo Culture to the Plant World.* PhD diss., Yale University.

Conklin, Harold C. 1962. Lexicographical Treatment of Folk Taxonomies. *International Journal of American Linguistics* 28(2), Part 4, Monographic Supplement/Memoir 21:119–41.

Dauenhauer, Nora Marks, and Richard Dauenhauer, eds. 1987. *Haa Shuká, Our Ancestors: Tlingit Oral Narratives*. Classics of Tlingit Oral Literature, Vol. 1. University of Washington Press.

Davis, Wade, K. David Harrison, and Catherine Herbert Howell. 2007. *Book of Peoples of the World: A Guide to Cultures*. National Geographic Books.

Dougherty, Janet W. D. 1983. *West Futuna-Aniwa: An Introduction to a Polynesian Outlier Language.* University of California Press.

Dourisboure, Pierre X. 1889. *Dictionnaire bahnar-français.* Hong Kong.

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Farquhar, Sebastian, Jannik Kossen, Lorenz Kuhn, Adam Goldie, Mirac Suzgun, and Yarin Gal. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* 630:625–30.

Geertz, Clifford. 1973. Thick Description: Toward an Interpretive Theory of Culture. In *The Interpretation of Cultures: Selected Essays*. Basic Books.

Guilleminet, Paul, with Jules Alberty. 1959–63. *Dictionnaire bahnar-français.* 2 vols. Publications de l'École française d'Extrême-Orient 40. École française d'Extrême-Orient.

Harrison, K. David, Gregory D.S. Anderson, and Alexander Ondar. 2007. *Tuvan Talking Dictionary*. Swarthmore College. Access this resource at http://tuvan.talkingdictionary.org (accessed March 2026).

Harrison, K. David, and Marjorie Hunt. 2013. *One World Many Voices: Endangered Languages and Cultural Heritage*. Smithsonian Center for Folklife and Cultural Heritage.

Harrison, K. David, Ausdin Jackson, Takaronga Kuautonga, et al. 2016. *Futuna-Aniwa Talking Dictionary.* Swarthmore College. Access this resource at http://www.talkingdictionary.org/futuna_aniwa (accessed March 2026).

Harrison, K. David, Jack Keitadi, Kirk Keitadi, et al. 2015. *Aneityum Talking Dictionary.* Swarthmore College. Access this resource at http://www.talkingdictionary.org/aneityum (accessed March 2026).

Harrison, K. David, Neal Kelso, Dominik M. Ramík, et al. 2024. Wind Lore as Environmental Knowledge in Southern Vanuatu. *Journal of Marine and Island Cultures* 13(1).

Harrison, K. David, Neal Kelso, Dominik M. Ramík, et al. 2025. "Sun Brings All Things": Sun and Moon Lore as Biocultural Knowledge on Aneityum Island, Vanuatu. *PLoS ONE* 20(8):1-20.

Harrison, K. David, Hoài Trần, Công Minh Khang Hoàng, et al. 2026. Documenting Environmental Knowledge in the Bahnar Language of Vietnam. *Languages* 11(7): 141.

Hoàng, Thị Mỹ Lệ, and Duy Linh Trần. 2024. Định hướng xử lý tiếng dân tộc Hrê và dân tộc Co ứng dụng xây dựng kho ngữ vựng song ngữ. [Orientation for Processing the Hrê and Co Languages toward the Development of a Bilingual Lexical Database.] Tạp chí Khoa học và Công nghệ Việt Nam. Access this article in Vietnamese at: https://vjst.vn/dinh-huong-xu-ly-tieng-dan-toc-hre-va-dan-toc-co-ung-dung-xay-dung-kho-ngu-vung-song-ngu-viet-hre-hre-viet-viet-co-va-co-viet-67211.html (accessed March 2026).

Accepted Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Houde, Nicolas. 2007. The Six Faces of Traditional Ecological Knowledge: Challenges and Opportunities for Canadian Co-management Arrangements. *Ecology and Society* 12(2):34.

Inglis, John. 1882. *A Dictionary of the Aneityumese Language: In Two Parts. I. Aneityumese and English. II. English and Aneityumese. Also Outlines of Aneityumese Grammar.* London.

Ji, Ziwei, Nayeon Lee, Rita Frieske, et al. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12):1–38.

Lillehaugen, Brook D., Aurora Sánchez Gómez, Floriana Hernández Martínez, Savita M. Deo, Graham Mauro, Saul Ontiveros, and K. David Harrison. 2019. San Bartolomé Quialana Zapotec Talking Dictionary, version 1.0. Swarthmore College. Access this resource at http://www.talkingdictionary.org/quialana.

Lương, Bèn, Viết Toại Nông, Kim Dung Lương, and Hương Giang Lê. 2011. *Từ điển Tày – Việt* [Tày–Vietnamese Dictionary]. Nhà xuất bản Đại học Thái Nguyên.

Nguyễn, Hữu Hoành, ed. 2013. *Ngôn ngữ, chữ viết các dân tộc thiểu số ở Việt Nam (Những vấn đề chung).* [Languages and Writing Systems of Ethnic Minorities in Vietnam: General Issues.] Nhà xuất bản Từ điển Bách Khoa.

Nguyễn, Văn Lợi, and Hữu Hoành Nguyễn. 1998. *Tiếng Katu.* [Katu Language.] Nhà xuất bản Khoa học Xã hội.

Ong, David, and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, 245–246. Association for Computational Linguistics.

Plunkett, Gregory M., Michael J. Balick, Presley Dovo, et al. 2024. Message Plants of Southern Vanuatu. *Sociolinguistic Studies* 18(3–4):377–404.

Reimers, Nils, and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Reitmaier, Thomas, Dani Kalarikalayil Raju, Ondrej Klejch, et al. 2024. Cultivating Spoken Language Technologies for Unwritten Languages. In *Proceedings of the 2024 CHI*

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Conference on Human Factors in Computing Systems. Association for Computing Machinery.

Reyes, Ruben. 2012. *Garüdia: Garifuna Trilingual Dictionary (Garifuna-English-Spanish)*. CreateSpace Independent Publishing Platform.

Reyes, Ruben. n.d. *Ruben Reyes Online Garifuna Dictionary*. Garifuna Institute. https://www.garifunainstitute.com/index.php (accessed March 2026).

Sear, Victoria, and Mark Turin. 2021. Locating Criticality in the Lexicography of Historically Marginalized Languages. *History of Humanities* 6(1):237–259.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems 33*. Curran Associates Inc.

Tạ, Văn Thông. 2018. Biên soạn từ điển đối dịch trước nguy cơ mai một các ngôn ngữ ở Việt Nam. [Compiling Bilingual Dictionaries in Response to Language Endangerment in Vietnam.] *Tạp chí Khoa học Đại học Tân Trào* 10:46–54.

Trần, Hoài, Bên A, Hung A, Trường Huy Đỗ, and K. David Harrison. 2024a. Baskets of Wisdom: Bahnar Basketry, Folk Taxonomies, and the Maintenance of Upland Environmental Intelligence in Vietnam. *Sociolinguistic Studies* 18(3–4):433–66.

Trần, Hoài, K. David Harrison, Khang Minh Công Hoàng, and Hải Lâm Cao. 2024b. "Who Eats the Forest?": Forest Animacy among the Bahnar People of Vietnam. *Journal of Ethnobiology* 44(4):402–14.

Trần, Hoài, Thế Lâm Dương, Công Minh Khang Hoàng, Xơm A, and K. David Harrison. 2023. *Bahnar Talking Dictionary*. Swarthmore College. Access this resource at http://www.talkingdictionary.org/bahnar.

Trần, Hoài, K. David Harrison, Phước Lâm Huy Trần, Cong Minh Bùi, and Trung Kiên Phạm. 2025. The Bahnar Canoe and Environmental Knowledge in the Central Highlands of Vietnam. *Journal of Material Culture* 30(4):524–544.

Wicherkiewicz, Tomasz. 2025. Review of *Mastering Languages, Taming the World: The Production and Circulation of European Dictionaries and Lexicons of Asian Languages (16th–19th Centuries)*, edited by Michela Bussotti and François Lachaud. *Journal of Jesuit Studies*. 12(2):345–349.

Accepted Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05

Ye, Zhengdao, and Cliff Goddard. 2026. Conceptualizing the Environment: Words as Cultural Lenses for Seeing the World. *International Journal of Language and Culture.*

< Captions and Alt Text >

Figure 1: Bên A, Bahnar dictionary coauthor and culture expert, with a traditional Bahnar *plung* dugout canoe at the Đăk Bla river. Photo by the authors.
Figure 1 alt text: A man stands beside a large wooden dugout canoe on the bank of a river.

Figure 2: Xơm A (left), Bahnar dictionary coauthor and culture expert, observes Jo A (right) weaving a traditional Bahnar *reo* basket. Photo by the authors.
Figure 2 alt text: One man observes while another weaves a basket, with partially completed basketwork visible.

Figure 3: Bahnar culture expert Banh A (left) working with researcher C. M. K. Hoàng (right) to validate a bird species checklist compiled from citizen science platform iNaturalist. Photo by the authors.
Figure 3 alt text: Two men sit looking at a checklist, engaged in collaborative validation.]

[1] ISO 639-3 language codes: Bahnar [bdq]; Hrê [hre]; Co/Cua [cua]; Katu [kuf]; Mảng [zng]; Rục (no ISO 639-3 code assigned; classified under Chứt [scb]); Tày [tyz]; Aneityumese [aty]; Futuna-Aniwa [fut].

[2] To avoid data retention by AI models, we disabled the setting that allows it.

Accepted, Uncorrected Manuscript

How to Cite: Please cite the final published Version of Record (VOR) upon its release. Alternative Provisional Citation: Harrison, D., H. Trần, and X. A. (2027). Dictionary as Grift, or Dictionary as Gift? Indigenous Lexicons, Generative Artificial Intelligence, and the Ethics of Extraction. Accepted for publication in Journal of American Folklore 140(556). DOI: 10.5406/15351882.140.556.05