

Article

Digital Dictionary Approaches for Endangered Languages

Sevgi Çalışır Zenci 

Department of Turkish Language and Literature, Faculty of Letters, Anadolu University, Eskişehir 26470, Türkiye; sevgicalisir@anadolu.edu.tr

Abstract

This study examines digital lexical resources developed for endangered languages through a qualitative comparative framework grounded in documentary linguistics and digital lexicography. Focusing on four representative cases—the Māori Dictionary, the Ojibwe People's Dictionary, Talking Dictionaries, and the Endangered Languages Archive—the study analyzes how different models of digital dictionaries are structured and how they support language documentation, pedagogy, and community engagement. Rather than treating these resources as directly comparable entities, the study introduces a typological distinction between stand-alone community-oriented dictionaries and large-scale documentary infrastructures. Based on an operationalized analysis of lexicographic structure, contextualization, technological features, and participation models, the findings demonstrate that digital lexical resources are functionally differentiated rather than hierarchically ordered. The results suggest that these resources operate along a continuum between pedagogical accessibility and documentary depth. Community-based dictionary models tend to emphasize user-oriented design, culturally embedded content, and language learning, while archival platforms prioritize scalability, metadata standardization, and long-term preservation. Importantly, community participation emerges as a variable rather than uniform feature, particularly within large-scale infrastructures. The study contributes to the literature by proposing a functional-typological framework for analyzing digital dictionaries of endangered languages and by highlighting the need to move beyond monolithic and evaluative comparisons. It further emphasizes the importance of integrating community perspectives, contextual richness, and technological adaptability in the design of sustainable digital language resources.

Keywords: endangered languages; digital lexicography; documentary linguistics



Academic Editors: K. David Harrison,
Lisa Lim and Anvita Abbi

Received: 1 March 2026

Revised: 29 June 2026

Accepted: 7 July 2026

Published: 21 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

A large proportion of the approximately 7000 languages in the world are endangered; according to [UNESCO \(2003\)](#), approximately 40% of these languages are at risk of disappearing within the next few generations. Documenting and maintaining endangered languages is critically important not only for the preservation of linguistic diversity but also for the preservation of cultural and social memory ([Himmelmann, 1998](#)). In this context, digital technologies make it possible to store oral and written data in an easily accessible, updatable, and multimedia-supported format, taking language documentation beyond traditional printed methods ([Abel & Meyer, 2013](#); [Bird & Simons, 2003](#); [Lew, 2012](#)).

Digital dictionaries go beyond the functions of classic dictionaries, positioning themselves as holistic digital ecosystems that present lexical information along with its linguistic, cultural, and discursive context ([Bower, 2015](#)). These ecosystems support the pedagogical

teaching, academic documentation, and community ownership processes of endangered languages. In particular, digital dictionaries, thanks to their capacity to offer multimodal content, can include not only word meanings but also pronunciation, usage context, cultural references, and narrative examples (Austin & Sallabank, 2011).

Ogilvie's (2011) work demonstrates that dictionaries for endangered languages are not only language preservation tools but also dynamic tools that integrate lexicography, language documentation, and teaching contexts. This study goes beyond traditional lexicography, emphasizing the integration of technology, native speaker participation, and pedagogical materials into dictionary design. Cablitz (2011) examines how cultural knowledge is documented in dictionaries of endangered languages and what requirements this imposes on dictionary design. Based on fieldwork conducted in the Marquesan language, this article argues that dictionaries should include not only word definitions but also culturally relevant descriptions. Recent studies investigate the processes of developing digital applications and generating data in endangered languages through models that integrate linguistics and artificial intelligence. For example, a study conducted in the Toto language focuses on trilingual digital applications, Unicode integration, and morphological analysis, and examines how this can be used for language documentation and animation (Guha et al., 2025). Studies also show that digital dictionaries function as pedagogical tools in language revitalization processes. For example, digital dictionary projects for Māori and Ojibwe communities contribute to language preservation by not only providing dictionary entries but also encouraging the active participation of the speaking communities (Hinton, 2013). Therefore, digital dictionaries play a central role in both academic documentation and community-centered language policies.

This study aims to analyze digital dictionary approaches developed for endangered languages in terms of structure and content, technology use, and community participation. This study examines distinct models of digital lexical resources developed for endangered and revitalizing languages. Rather than comparing equivalent dictionary products, it analyzes how different digital infrastructures support language documentation, learning, and revitalization.

The study addresses the following research questions:

1. How do different models of digital lexical resources vary in terms of structure, functionality, and intended use?
2. How do these models balance documentation, pedagogy, and community engagement?
3. What design principles emerge as most effective for supporting endangered language sustainability?

The analysis distinguishes between two broad types of digital language resources: Stand-alone community dictionaries (e.g., Māori Dictionary, Ojibwe People's Dictionary).

Documentary and archival platforms (e.g., Talking Dictionary, Endangered Languages Archive).

This typological distinction enables a more methodologically robust comparison by focusing on digital dictionary models rather than treating all platforms as directly equivalent entities.

Documentary Linguistics

Documentary linguistics is a research approach that aims to systematically record the structural features of language, as well as its context of use, cultural references, and social function (Himmelmann, 1998). In this framework, dictionaries are considered not only tools that present the meaning of words, but also fundamental outputs that enable the documentation of linguistic and cultural information. Austin and Sallabank (2011) emphasize that documentary linguistics is particularly critical for languages with few

speakers and those in danger. This perspective highlights the need for digital dictionaries to be enriched with multimedia data integration and contextual examples.

Digital Lexicography

Digital lexicography goes beyond traditional lexicography, focusing on developing interactive platforms that bring together audio, video, visual, and textual data. This approach offers the following advantages for endangered languages: Multimodal data presentation; pronunciation, usage context, and cultural narratives can be integrated into dictionary entries. Updatability; dictionary contents can be regularly updated online. Accessibility and sustainability; open-access and archival platforms ensure the long-term preservation of language resources (Lew, 2012).

Digital dictionaries also function as pedagogical tools in language revitalization processes, providing targeted content to both community members and second language learners (Bowern, 2015).

Language Revitalization and Community Engagement

The literature on language revitalization emphasizes that community participation is critical to the sustainability of endangered languages (Hinton, 2013). Digital dictionaries support community-centered language policies by enabling speakers to directly participate in content production and validation processes (Abel & Meyer, 2013). This participatory process enhances both the pedagogical function and cultural legitimacy of dictionaries (Nathan, 2006).

Recent scholarship has further expanded the theoretical and methodological landscape of digital language documentation and lexicography. Current research increasingly emphasizes that digital resources for endangered languages should move beyond preservation-oriented models toward revitalization-oriented, community-based approaches. In particular, the literature highlights the importance of community participation, digital sustainability, and ethical stewardship of linguistic data in the development of digital language resources (Hinton et al., 2018). Similarly, computational approaches such as corpus-driven morphological analysis, speech technologies, and small language models are demonstrating substantial potential for enhancing the accessibility and usability of endangered-language resources (Guha et al., 2025; Bartley & Ragni, 2025). At the same time, scholars have stressed that technological innovation must remain grounded in community priorities, cultural context, and ethical stewardship. This shift reflects a broader movement in documentary linguistics: from the mere digital preservation of linguistic materials toward the creation of dynamic, multifunctional ecosystems that support language learning, intergenerational transmission, and community empowerment (Shi et al., 2026). Recent large-scale archival initiatives, such as the renewed development of the Pangloss Collection (Vasile et al., 2020), also illustrate how digital infrastructures are increasingly designed to serve both specialist researchers and broader community audiences through differentiated interfaces and open-access models (Adamou et al., 2025).

In this context, the analytical framework of the study is structured around the dimensions of (i) structure and content, (ii) technology use, and (iii) community participation. These three dimensions have been identified as key criteria for evaluating the contribution of digital dictionaries to the documentation, teaching, and revitalization processes of endangered languages.

2. Materials and Methods

This study employs a qualitative comparative case study design. The aim is not to rank digital resources, but to identify and analyze distinct models of digital lexical resource design for endangered languages.

The study is exploratory and interpretive, drawing on principles from documentary linguistics, digital lexicography, and participatory language revitalization.

Sample

The research unit consists of digitally accessible dictionary projects. The sample was determined using purposeful sampling. The following criteria were considered in the selection process: focus on endangered or revitalized languages, online accessibility, multimedia integration, representation of different technological infrastructures, coverage of different geographic regions, visibility in the literature, and widespread use.

Cases were selected through purposive sampling according to the following criteria:

- relevance to endangered or revitalizing languages;
- public digital accessibility;
- significance in the scholarly and professional literature;
- representation of different infrastructural and institutional models;
- inclusion of multimedia and/or community-oriented features.

The selected cases represent two analytically distinct categories:

A. Stand-alone Community Dictionary Models

Māori Dictionary (MD)

Ojibwe People's Dictionary (OPD)

B. Documentary and Archival Infrastructure Models

Talking Dictionary (TD)

Endangered Languages Archive (ELAR)

Importantly, TD and ELAR are analyzed not as single dictionaries, but as infrastructural frameworks that host multiple language documentation projects.

Although major archival platforms such as Pangloss Collection and PARADISEC (Barwick & Thieberger, 2003) make highly significant contributions to endangered language documentation, they were not included in the present analysis for methodological reasons. The primary aim of this study is not to provide an exhaustive survey of all digital language archives, but rather to examine representative models of digital lexical resources across distinct functional orientations. Specifically, the selected cases were chosen to reflect two complementary types of digital language infrastructure: stand-alone, community-oriented dictionary projects (MD and OPD) and large-scale, multilingual documentary platforms that incorporate lexical resources within broader archival ecosystems (TD and ELAR). While Pangloss and PARADISEC are exemplary archival repositories, their principal emphasis lies in corpus preservation, long-term data stewardship, and research accessibility rather than dictionary-centered lexical interface design. Including them would certainly enrich a broader survey of documentary archives; however, their exclusion here reflects a deliberate effort to maintain analytical coherence and comparability among cases selected to represent specific models of digital lexicographic practice. Recent scholarship on Pangloss further confirms its importance as a premier open-access archive for endangered and underdocumented languages (Adamou et al., 2025).

Data Collection Process

Data were collected from multiple complementary sources, including systematic examination of dictionary platform interfaces, project documentation and technical descriptions, published scholarly studies, and publicly available statements from developers and participating communities. Each platform was examined through repeated visits between December 2025 and February 2026 to ensure consistency and depth of observation, and the findings were recorded in a structured data collection form.

The collected data were analyzed using thematic content analysis. The analytical process proceeded in four stages: first, open coding was conducted to identify initial pat-

terns; second, the resulting codes were grouped into categories; third, broader themes were generated; and finally, the platforms were analyzed comparatively to identify similarities and differences across cases.

Unit of Analysis

The unit of analysis in this study is defined as the digital lexical resource model rather than individual lexical entries or entire platforms treated as monolithic entities. This distinction enables a more nuanced comparison across resources that differ in scale, scope, and institutional orientation, without reducing them to overly simplified units.

The analytical framework is structured along three interrelated dimensions. The first dimension focuses on lexicographic design and content architecture, examining how entries are structured, the extent to which they are contextualized, how cultural knowledge is integrated, and whether the resource reflects a pedagogical orientation. The second dimension addresses technological infrastructure and functionality, including the degree of multimedia integration, the effectiveness of search and navigation systems, the organization of metadata, and the presence of computational language tools. The third dimension considers community participation and sustainability, with particular attention to speaker involvement, governance and ownership models, access policies, and strategies for long-term preservation.

The analytical procedure followed a thematic comparative approach conducted in four stages. First, each case was documented descriptively to establish a detailed baseline. This was followed by within-case analysis to identify internal patterns and characteristics. Subsequently, cross-case comparisons were performed both within and across different resource types to detect broader trends and divergences. Finally, the findings were synthesized to derive emergent design principles. This multi-stage process makes it possible to identify both shared features and structural differences, while avoiding the imposition of artificial equivalences between inherently different types of resources.

3. Results

Table 1 presents the general characteristics of the platforms.

Table 1. General characteristics of the platforms.

Features	MD	TD	OPD	ELAR
Language	Māori	Tuvan, Chamacoco, etc.	Ojibwe	Various languages
Geography	New Zealand	Global (America, Asia, etc.)	USA and Canada	Global
Speaker Status	revitalizing language	Mostly severely endangered	Endangered, yet with a strong community	Including critically endangered languages
Website	maoridictionary.co.nz	talkingdictionary.org	ojibwe.lib.umn.edu	elararchive.org
Structure-content	Bidirectional English–Māori structure	Fieldwork-based entries and cultural notes	Narrative-based usage examples and cultural context descriptions	Dictionary and corpus integration with rich metadata
Multimedia	Native speaker audio recordings, pronunciation-oriented design	Native speaker audio, photographs, and video in selected projects.	Audio, video, photographs, and maps	Audio, video, and field recordings
Technology	Advanced search, user-friendly interface, mobile-compatible	Talking Dictionary infrastructure, web-based search, metadata tagging	Advanced filtering, morphological search, thematic navigation	Digital archive infrastructure, persistent DOI system, standardized metadata

MD; maoridictionary.co.nz is a comprehensive online dictionary for te reo Māori (the Māori language of New Zealand). This platform offers a dictionary interface with Māori-English and English-Māori translation capabilities and includes both everyday words and culturally specific terms. In addition to basic vocabulary, it also contains encyclopedic information on plants, animals, celestial bodies, important Māori figures, ancestors, and place names; thus, it serves as more than just a word list, but a valuable resource for understanding the Māori language and culture. It also provides examples of idioms, proverbs, and grammatical explanations. The site is also available as a mobile application, allowing users to save their favorite words and reinforce their learning through quizzes. The dictionary is based on the comprehensive printed dictionary *Te Aka Māori-English, English-Māori Dictionary and Index* by Moorfield (2003–2026), and its online version provides quick access and practical ease of use for Māori language learners.

TD; is a free online dictionary project developed by Swarthmore College. The platform aims to document and preserve endangered languages worldwide. The concept was conceived in 2005 under the leadership of K. David Harrison, and since then, dictionaries have been created for over 200 different languages. Each dictionary presents phrases, words, and idioms both as text and through audio recordings of native speakers, making language learning and pronunciation understanding easier. The project is offered free of charge thanks to support from language communities, academics, and donors. The platform includes a search interface allowing users to search between English and the target language, and typically features rich content such as audio examples and sometimes images. Independent dictionary pages for different languages, such as Ho (Anderson et al., 2010) or Wamesa (Gasser, 2015), contain thousands of words and their corresponding audio recordings. This structure both presents the vocabulary of the language and preserves native pronunciation information, thus providing a living record of the language.

OPD; is an online Ojibwe (Anishinaabemowin) language dictionary developed by the University of Minnesota (Livesay & Nichols, 2012–2026). This platform offers English-to-Ojibwe search capabilities, audio recordings, example sentences, photos, videos, and cultural content, making it much more than just a simple translation dictionary. Users can search for words in both Ojibwe and English and listen to pronunciations from native speakers, facilitating pronunciation and learning. The interface also provides historical documents and materials, as well as cultural context, demonstrating the language's connection to daily life and cultural practices.

The project was initiated in partnership between academics from the University of Minnesota's American Indian Studies department and the Minnesota Historical Society, and its goal is to contribute to the documentation, preservation, and teaching of the Ojibwe language. With its comprehensive search capabilities, visual and audio content, and continuously updated database, the dictionary aims to be a valuable resource for language learners, teachers, and community members.

Data for this study were also sourced from the ELAR. This site aims to preserve and publish multimedia content such as audio, video, transcriptions, and related documents for endangered languages worldwide. The archive contains rich content including everyday conversations, ritual narratives, songs, oral literature, and significant cultural information about the language. Users can access, search, and explore these collections for free, including materials available for many languages. Researchers, language communities, and the public can also access this content free of charge by registering.

ELAR was established in the early 2000s in response to the need to preserve linguistic diversity and initially operated within SOAS (Seyfeddinipur, 2003). Responding to the global language crisis, where half of the world's approximately 7000 languages are at risk of extinction, the platform ensures the long-term digital preservation and accessibility

of language documents. The archive not only provides content but also offers training, guidance on metadata creation, and digital preservation services for language documenters, along with support services. Since July 2021, ELAR has been officially housed and managed at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) in Berlin, Germany. ELAR serves as a primary digital repository for multimedia documentation of endangered languages, providing preservation and access to unique linguistic and cultural materials worldwide.”

While these four dictionary/archive sites (TD, MD, OPD, and ELAR) fundamentally share the goal of preserving and making endangered languages accessible, they differ significantly in terms of target audience, scope, and functionality (Table 2).

Table 2. Key features of digital lexical resource platforms.

Platforms	Scope and Focus	Content Types	Target Users	Access and Use
TD	Multilingual platform comprising 200+ language-specific dictionary projects; primarily focused on lexical documentation and pronunciation	Lexical entries; audio recordings; limited example usage	Language learners, researchers, and documentation practitioners	Web-based interface; relatively simple search functionality; audio-supported access to lexical items
MD	Comprehensive single-language resource (Māori) integrating lexical, cultural, and encyclopedic information	Lexical entries; example sentences; idiomatic expressions; cultural and encyclopedic content; audio	Community members, students, and language learners	User-oriented interface; advanced dictionary search; mobile compatibility; interactive learning features
OPD	Single-language platform (Ojibwe) emphasizing culturally embedded lexical knowledge and narrative-based usage	Lexical entries; audio recordings; example sentences; images; videos; narrative and cultural content	Community members, educators, and language learners	Advanced search and filtering; thematic navigation; multimedia-rich interface
ELAR	Multilingual archival infrastructure focused on language documentation and preservation across diverse linguistic contexts	Audio recordings; video; textual documents; transcriptions; metadata-rich collections	Researchers, language communities, and documentation specialists	Archive-oriented access; metadata-driven search; partial access restrictions depending on collection

The categories used in Table 3 are based on a qualitative operationalization of observed platform features. Rather than assigning numerical scores, each dimension was evaluated according to the presence, extent, and integration of relevant features within the platform. For example, contextualization level refers to the degree to which lexical entries are supported by usage examples, discourse context, and cultural information, while community participation reflects the extent of speaker involvement in content creation, validation, and governance. Importantly, for large-scale infrastructures such as ELAR and TD, evaluations reflect general tendencies observed across multiple collections rather than uniform characteristics of all individual projects.

Table 3. Operationalized comparative analysis of digital lexical resource models for endangered languages.

Dimension	TD	MD	OPD	ELAR
Language status focus	Primarily severely endangered languages across multiple regions	Revitalizing language (Māori)	Endangered but actively revitalized language	Broad range (from vulnerable to critically endangered)
Lexicographic structure	Entry-based, fieldwork-derived lexical units with audio alignment	Definition-centered entries with bilingual structure and examples	Narrative-enriched entries integrating stories and usage contexts	Metadata-linked lexical materials embedded within larger corpora
Contextualization level	Moderate: lexical items supported by limited contextual notes	Moderate to high: examples, idioms, and cultural references	High: extensive narrative, discourse-based, and culturally embedded usage	Variable: depends on collection; often high in research-oriented datasets
Multimedia integration	Audio consistently; visual media in selected projects	Audio with limited visual support	Audio, video, images, and maps systematically integrated	Extensive multimedia (audio, video, field recordings), but not always lexicon-centered
Search and navigation	Basic keyword search; bilingual lookup	Advanced dictionary-style search; user-oriented interface	Advanced filtering, thematic navigation, partial morphological search	Metadata-driven search; archive-oriented navigation system
Technological features	Standard web-based infrastructure; limited computational tools	Stable platform; minimal computational processing tools	Partial computational support (e.g., morphological awareness tools)	Archival standards (DOI, metadata schemas); limited end-user linguistic tools
Community participation	Moderate: community involvement varies by project	High: strong integration with language users and educators	Very high: community-driven content, storytelling, and validation	Variable: ranges from researcher-led to community-collaborative projects
Access model	Open access	Open access	Open access	Tiered access (open + restricted collections depending on permissions)
Pedagogical orientation	Moderate: supports vocabulary acquisition and pronunciation	High: explicitly designed for language learning and teaching	High: supports immersive, culturally grounded language learning	Limited to moderate: primarily research-oriented, pedagogical use depends on dataset
Sustainability model	Project-based, grant-supported	Institutionally maintained and continuously updated	University–community partnership model	Long-term digital archiving with institutional and international support

Table 3 presents an operationalized comparison of four digital lexical resource models across key analytical dimensions, including lexicographic structure, contextualization, technological infrastructure, and community participation. Rather than assigning fixed

quantitative scores, the comparison is based on the presence, extent, and integration of features observed across platforms.

The analysis indicates that the examined resources differ not in terms of overall quality, but in their functional orientation and design priorities. Stand-alone dictionary models such as the MD and the OPD demonstrate a stronger pedagogical orientation, as reflected in their user-centered interface design, structured entries, and integration of examples and culturally relevant content. In contrast, infrastructural platforms such as TD and the ELAR prioritize documentation breadth, multilingual scalability, and long-term data preservation.

In terms of contextualization, OPD stands out for its narrative-based and discourse-rich structure, where lexical items are embedded within stories and culturally situated usage. MD similarly integrates cultural references and example-based explanations, although within a more conventional lexicographic framework. TD provides consistent audio-supported lexical documentation, though contextual depth varies across individual projects. ELAR, as an archival infrastructure, offers potentially high levels of contextualization; however, this is distributed unevenly across collections and is often oriented toward research rather than pedagogical use.

Technological comparison further highlights a distinction between user-oriented and archive-oriented design. MD and OPD provide more accessible search and navigation systems tailored to language learners, whereas ELAR relies on metadata-driven retrieval systems designed for research and documentation purposes. TD occupies an intermediate position, offering relatively simple search functionality combined with scalable infrastructure.

Community participation emerges as a key differentiating factor, though it varies across and within platforms. OPD represents a strongly community-embedded model, with speaker involvement in content creation and validation. MD also reflects sustained engagement with language users, particularly in educational contexts. In contrast, participation in TD and ELAR is more heterogeneous, ranging from community-collaborative to primarily researcher-led projects.

Overall, the findings suggest that digital lexical resources for endangered languages operate along a continuum between pedagogical accessibility and documentary depth, rather than representing directly comparable or hierarchical models.

4. Discussion

This study has shown that digital dictionaries developed for endangered languages are not merely lexicographical tools; rather, they function as multi-layered digital ecosystems situated at the intersection of language documentation, cultural transmission, and language revitalization processes. The findings demonstrate that digital lexicography is increasingly integrated with documentary linguistics and community-based language planning. Among the projects examined, the narrative-based and multimodal structure offered by the OPD, in particular, shows that dictionaries are evolving from the traditional “definition-centered” model into context-rich digital learning environments. This finding indicates that digital dictionaries need to be reconceptualized not only as repositories of lexical information but also as usage-based language resources. In contrast, the more structured and pedagogically oriented design of the MD shows that accessibility and learnability remain central to language revitalization processes. This reveals the necessity of functional differentiation according to the user profile in digital dictionary design.

The documentary linguistics perspective is most clearly observed in the TD. The systematic use of field-data-based dictionary entries and native speaker recordings reflects an approach consistent with Himmelmann’s principles of documentation. However, pedagogical user experience appears to be secondary in these projects. On the other hand, the robust archival standards and metadata structure of the ELAR model offer a critical

advantage in terms of long-term digital preservation. However, it is noteworthy that the archive-focused design can be limiting in terms of accessibility, especially for language learners and community members. One of the strongest findings of the research is the positive relationship between the level of community participation and the revitalization potential of dictionaries. The speaker-centered content production observed particularly in the OPD increases the sociolinguistic legitimacy of digital dictionaries and makes the living use of the language visible.

In contrast, more researcher-centered models (especially ELAR) limit community interaction, relatively weakening the direct contribution of dictionaries to language revitalization processes. This situation demonstrates that the balance between documentary accuracy and community ownership is a critical parameter in digital dictionary design. Another important yet often overlooked aspect in the design of digital dictionaries for endangered languages is the choice of interface and translation languages. As language shift progresses, fluent speakers become increasingly rare, while younger generations tend to communicate primarily in the dominant regional or national language. Consequently, dictionaries designed exclusively in the endangered language may not adequately serve semi-speakers, heritage learners, or community members who are attempting to reclaim their ancestral language. For this reason, bilingual or multilingual digital dictionaries can significantly enhance accessibility, usability, and community participation. Incorporating translations into the dominant local language and, where appropriate, an international language such as English enables broader educational use while supporting intergenerational language transmission. The Great Andamanese digital dictionary illustrates this approach by presenting lexical entries in Great Andamanese alongside Hindi and English, thereby facilitating both community engagement and linguistic documentation despite the language having only a handful of remaining speakers during the compilation process (Abbi, 2012).

In line with these findings, it is recommended that the following design principles be prioritized in digital dictionaries to be developed for endangered languages: (i) multimodal and context-rich data architecture, (ii) systematic participation of speaker communities in the production and validation processes, (iii) bilingual or multilingual interfaces that reflect the linguistic ecology of the target community and support both heritage learners and semi-speakers, (iv) balancing archival resilience with a user-oriented pedagogical interface, and (v) controlled integration of Artificial Intelligence (AI)-powered language processing tools.

In the context of Turkic languages, significant advances have also been made in the field of digital lexicography and language technologies. In particular, digital dictionary initiatives developed for minority and regional Turkic languages such as Tuvan, Khakas, Altai, and Tatar demonstrate that innovation in this field is not limited to widely studied global cases. For Tuvan and Altai, online bilingual dictionaries and other digital language resources have played an important role in improving the documentation, accessibility, and transmission of lexical knowledge (Harrison, 2007; Grenoble & Whaley, 2005). For Khakas, electronic dictionaries and corpus-based resources provide not only lexical inventories but also morphological annotation, contextual usage examples, and corpus-query functions, thereby aligning with the core principles of contemporary digital lexicography (Lew, 2012). Among these languages, Tatar has developed one of the most sophisticated digital ecosystems, including comprehensive electronic dictionaries, national corpus projects, automatic morphological analyzers, and machine translation tools, all of which have significantly enhanced its digital presence and functional capacity (Gatiatullin et al., 2024). These examples demonstrate that digital dictionaries serve not merely as instruments of language preservation, but also as foundational infrastructures for the development of language technologies for low-resource languages. More broadly, the Turkic case illustrates how

digital dictionaries can function as multifunctional platforms that bridge language documentation, pedagogy, community engagement, and computational language processing. From this perspective, comparative research on digital dictionary initiatives across Turkic languages would make a valuable contribution to both digital lexicography and the long-term sustainability of endangered and minority languages.

This study examines representative models rather than an exhaustive sample of all digital dictionary initiatives. Future research should expand the comparative corpus to include additional platforms such as Pangloss and PARADISEC, as well as regionally specific initiatives.

User-centered studies involving learners, teachers, and community members would further strengthen understanding of effectiveness and impact. In addition, the emerging role of artificial intelligence in low-resource language lexicography warrants dedicated investigation.

5. Implications for Digital Language Preservation and Conclusions

Digital dictionaries for endangered languages should be understood as strategic digital infrastructures rather than merely reference tools. Their success depends on how effectively they balance documentation, pedagogy, preservation, and community ownership.

In Türkiye, scholarship on Turkic languages has traditionally relied on printed comparative dictionaries and historical lexicographic compilations, particularly those associated with classical works such as Sir Gerard Clauson and the comparative lexicographic tradition represented by Ahmet Bican Ercilasun and related scholars. However, recent digital initiatives in languages such as Tuvan, Khakas, and Altai demonstrate that contemporary lexicographic platforms have moved beyond the mere presentation of headwords. These systems increasingly integrate corpus access, morphological analysis, and, in some cases, community-oriented participation. For example, Tatar Corpus Project ([Saykhunov et al., 2010](#)) and *Speak Tatar* (2021) illustrate a model in which lexicographic data are embedded within broader digital language infrastructures. This stands in contrast to most digital dictionary systems in Türkiye—including Turkish Language Association—where users typically access lexical entries in isolation, with limited exposure to authentic usage contexts or dynamic language data.

Online dictionaries developed for Tuvan and Altai are especially noteworthy because of their multilingual interfaces and partial support for collaborative updating. Platforms such as Tuvan-English Dictionary (Swarthmore) and *Altai-Khakas Dictionary* (2021); *Speak Tatar* (2021) indicate a shift in how minority Turkic languages are conceptualized no longer solely through Russian as an intermediary language, but increasingly as globally accessible linguistic resources. This transformation is significant because most Soviet-era dictionaries were designed primarily for local researchers and philologists. By contrast, contemporary digital platforms increasingly serve multiple audiences, including diaspora communities, language learners, and developers of artificial intelligence systems.

A particularly advanced case is the language resource infrastructure developed by the Tatarstan Academy (TATZET). Unlike many smaller-scale projects, TATZET functions not merely as an online dictionary but as a comprehensive digital ecosystem, incorporating electronic archives, a national corpus, dialect atlases, and machine translation tools. In this respect, Tatar occupies a distinct position among Turkic languages: whereas Tuvan and Khakas digital projects are primarily preservation-oriented, Tatar has transformed digitalization into an institutionalized language technology framework. This distinction reveals differing national and regional approaches to language planning in the post-Soviet context.

Despite these advances, several underexamined limitations remain. First, user experience design is generally weak. Most of these platforms appear to have been developed

from an academic rather than user-centered perspective. Interfaces are often outdated, mobile compatibility is limited, and pedagogical affordances are minimal. For younger researchers and students in Türkiye, this is a substantial barrier, as contemporary users increasingly expect digital dictionaries to provide integrated audio, contextual example sentences, visual support, automatic inflectional paradigms, and Application Programming Interface (API) access. Such features remain sparse in most Tuvu and Khakas platforms.

Second, the treatment of phonetic information is insufficient. For endangered languages, textual representation alone is inadequate; pronunciation, prosody, and dialectal variation are central to language preservation. Yet many of these digital dictionaries lack systematic audio recordings, and when recordings are available, they are often fragmentary and inconsistent. The absence of standardized phonetic annotation—particularly International Phonetic Alphabet (IPA)-based transcription—constitutes a serious limitation for Siberian Turkic languages, many of whose phonological distinctions cannot be reliably inferred from orthography alone.

Third, community participation remains limited. Successful contemporary language revitalization projects increasingly adopt a community-curated lexicon model, in which users can contribute lexical items, example sentences, oral histories, and dialectal variants. In contrast, most existing Turkic digital dictionaries still operate according to a top-down paradigm: academic institutions produce the data, while users passively consume it. Such a structure raises concerns about long-term sustainability, especially for endangered and under-resourced languages.

From the perspective of Turkological research in Türkiye, these platforms are still frequently treated as auxiliary reference tools. However, they have the potential to open several new research trajectories.

First, they create a basis for comparative digital Turkology. The digital platforms developed for Tuvan, Khakas, Altai, and Tatar invite comparative inquiry into how closely related minority languages are represented in digital space. A major unanswered question is the following: to what extent do digital dictionaries of minority Turkic languages function as instruments of language preservation, academic archiving, or community revitalization? This remains a significant gap in current scholarship.

Second, they are highly relevant to AI-ready lexicography. Although many of these resources were originally designed for human users, they contain data structures that are highly valuable for artificial intelligence applications, particularly in low-resource language modeling. Lemma-based lexical databases, morphologically tagged corpora, and aligned translation pairs are crucial for training large language models and other Natural Language Processing (NLP) systems. As AI research expands toward under-resourced languages, these digital dictionaries are likely to become foundational infrastructures rather than supplementary tools.

Third, these developments expose a notable gap within Türkiye itself. Comparable digital infrastructures are still lacking for many Turkic varieties and communities connected to Türkiye, including Gagauz, Karay, Ahıska Turkish varieties, and Iraqi Turkmen dialects such as those spoken in Tal Afar. Consequently, some minority Turkic languages in Siberia have, in certain respects, achieved more advanced digital lexicographic representation than several Turkic varieties that are historically and culturally closer to Türkiye. This discrepancy highlights the urgent need for a systematic digital Turkology agenda in Türkiye—one that integrates lexicography, corpus linguistics, community participation, and AI-oriented language resource development.

Funding: This work has been supported by Anadolu University Scientific Research Projects Coordination Unit under grant number #SHD-2026-3339.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (GPT-5.5, OpenAI) for language editing and text refinement. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
BBAW	Berlin-Brandenburg Academy of Sciences and Humanities
ELAR	Endangered Languages Archive
IPA	International Phonetic Alphabet
MD	Maori Dictionary Project
NLP	Natural Language Processing
OPD	Ojibwe People's Dictionary
TATZET	Tatarstan Academy
TD	Talking Dictionaries

References

- Abbi, A. (2012). *Great andamanese dictionary*. Andamanese.org. Available online: <https://www.andamanese.org/great-andamanese/language/dictionary> (accessed on 29 June 2026).
- Abel, A., & Meyer, C. M. (2013). User contributions to online dictionaries. In *Electronic lexicography in the 21st century (eLex): Thinking outside the paper* (pp. 1–16). Available online: <https://chmeyer.de/documents/2013-10-18-meyer-talk-elex2013.pdf> (accessed on 6 July 2026).
- Adamou, E., Guillaume, S., & Michaud, A. (2025). The Pangloss collection: Opening up research data on endangered and underdocumented languages. *Language*, 101(1), e38–e59. [CrossRef]
- Altai–Khakas Dictionary. (2021). Available online: <https://speak.tatar/en/dict/alt-kjh/> (accessed on 5 May 2026).
- Anderson, G. D. S., Harrison, K. D., & Pucilowski, A. (2010). *Ho talking dictionary. Living tongues institute for endangered languages*. Available online: <http://ho.talkingdictionary.org> (accessed on 5 May 2026).
- Austin, P. K., & Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.
- Bartley, C., & Ragni, A. (2025). How I built ASR for endangered languages with a spoken dictionary. *arXiv*, arXiv:2510.04832.
- Barwick, L., & Thieberger, N. (2003). *Pacific and regional archive for digital sources in endangered cultures*. PARADISEC. Available online: <https://www.paradisec.org.au/> (accessed on 5 May 2026).
- Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79(3), 557–582. [CrossRef]
- Bowern, C. (2015). *Linguistic fieldwork: A practical guide*. Springer.
- Cablitz, G. H. (2011). Documenting cultural knowledge in dictionaries of endangered languages. *International Journal of Lexicography*, 24(4), 446–462. [CrossRef]
- Gasser, E. (2015). *Wamesa talking dictionary, version 1.0*. Swarthmore College. Available online: <http://www.talkingdictionary.org/wamesa> (accessed on 5 May 2026).
- Gatiatullin, A., Suleymanov, D., Prokopyev, N., & Abdurakhmonova, N. (2024). Turkic morpheme: From the portal to the linguistic platform. In *WCIS 2022. Lecture notes in networks and systems* (Vol. 912). Springer. [CrossRef]
- Grenoble, L. A., & Whaley, L. J. (2005). *Saving languages: An introduction to language revitalization*. Cambridge University Press.
- Guha, A., Saha, S., Ballav, D., Mitra, S., & Chakraborty, H. (2025). Integrating linguistics and AI: Morphological analysis and corpus development of endangered Toto language of West Bengal. *arXiv*, arXiv:2510.22629. [CrossRef]
- Harrison, K. D. (2007). *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

- Himmelman, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36(1), 161–195. [CrossRef]
- Hinton, L. (2013). *Bringing our languages home: Language revitalization for families*. Heyday.
- Hinton, L., Huss, L. M., & Roche, G. (Eds.). (2018). *The Routledge handbook of language revitalization*. Routledge.
- Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger, & M. Paquot (Eds.), *Electronic lexicography* (pp. 343–361). Oxford University Press.
- Livesay, N., & Nichols, J. D. (Eds.). (2012–2026). *Ojibwe people's dictionary*. Available online: <http://ojibwe.lib.umn.edu> (accessed on 5 May 2026).
- Moorfield, J. C. (2003–2026). *Te Aka Māori dictionary*. Available online: <https://maoridictionary.co.nz/> (accessed on 5 May 2026).
- Nathan, D. (2006). Thick interfaces: Mobilising language documentation with multimedia. In J. Gippert, N. P. Himmelman, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 363–379). Mouton de Gruyter.
- Ogilvie, S. (2011). Linguistics, lexicography, and the revitalization of endangered languages. *International Journal of Lexicography*, 24(4), 389–404. [CrossRef]
- Saykhunov, M. R., Khusainov, R. R., Ibragimov, T. I., Luutonen, J., Salimzyanov, I. F., Shaydullina, G. Y., & Khusainova, A. M. (2010). *Corpus of written tatar*. Available online: <https://www.corpus.tatar/en> (accessed on 5 May 2026).
- Seyfeddinipur, M. (2003). *Endangered languages archive*. Available online: <https://www.elararchive.org/> (accessed on 5 May 2026).
- Shi, W., Han, X., Zhou, C., Liang, W., Lin, X., Zettlemoyer, L., & Yu, L. (2026). LMFusion: Adapting pretrained language models for multimodal generation. *Advances in Neural Information Processing Systems*, 38, 9163–9182.
- Speak Tatar. (2021). Available online: <https://speak.tatar/en/> (accessed on 5 May 2026).
- UNESCO. (2003). *Language vitality and endangerment*. UNESCO.
- Vasile, A., Séverine, G., Mourad, A., & Alexis, M. (2020). Le digital object identifier, une impérieuse nécessité? L'exemple de l'attribution de DOI à la collection Pangloss, archive ouverte de langues en danger. *I2DInformation, Données & Documents*, 2(2), 155–175. Available online: https://shs.cairn.info/article/I2D_202_0155?tab=texte-integral (accessed on 6 July 2026). [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.