OPEN ACCESS

Volume: 12

Special Issue: 1

Month: December

Year: 2023

P-ISSN: 2320-2645

E-ISSN: 2582-3531

Received: 19.10.2023

Accepted: 05.12.2023

Published: 14.12.2023

Citation:
Arul Dayanand, S., et al. "Digital Archives and Preservation Techniques for Revitalising Endangered Languages." *Shanlax International Journal of English*, vol. 12, no. S1, 2023, pp. 174–82.

DOI:
https://doi.org/10.34293/rtdh.v12iS1-Dec.133

Digital Archives and Preservation Techniques for Revitalising Endangered Languages

Dr. S. Arul Dayanand
Career Development Centre
SRM Institute of Science and Technology, Chennai, India

Dr. M. Uma Devi
Department of Computing Technologies
SRM Institute of Science and Technology, Chennai, India

Dr. Ramesh Kumar
English Language Department
HNU – ASU Joint International Tourism College, Hainan University, China

Abstract
In the field of linguistic preservation, the integration of digital archives with artificial intelligence (AI)-based tools presents a groundbreaking approach to revitalising endangered languages. Drawing upon the works of Bird et al. (2009) and Krafft and Kusters (2016), this study focuses on the development of a sophisticated digital archive that utilises AI to significantly enhance the preservation and revitalisation of these languages. This research involves creating an extensive digital repository to efficiently store, manage, and provide access to data related to endangered languages, thereby establishing a central hub for housing diverse linguistic materials, including audio, text, and video content. Crucially, this study incorporates AI-driven tools for comprehensive language analysis. These tools, which are essential for accurate language documentation, are adept at tokenizing and analysing texts in endangered languages, ensuring precise processing and preservation, as highlighted by Krauss (1992) and Lewis and Simons (2010). The methodology includes thorough data preparation for language model training, encompassing endangered language-specific tokens, regional dialects, and idiomatic expressions, as noted in research by Clyne (2003) and Bhuvaneswari (2022). The ultimate aim of this study was to preserve and revitalise endangered languages through accessible language learning resources and active community engagement. This study critically evaluates the impact of AI and digital archives on language revitalization by examining their effectiveness in promoting linguistic diversity and cultural preservation, as discussed by Robinson and Yip (2017) and Harrison (2020). This study underscores the potential of AI in enhancing and preserving linguistic and cultural diversity, offering a scalable and sustainable model that makes a significant contribution to the field of linguistic preservation.

Keywords: Endangered Languages, Digital Archives, Language Revitalisation, AI-based Tools in Linguistics and Language Preservation Techniques

Introduction

Contextualizing Endangered Languages

To reinvigorate endangered languages, which are fundamental to world cultural diversity, it is crucial to understand the complex

cultural, social, and historical factors that underpin their vitality (Bhuvaneswari, 2022). These languages are not merely communication systems; they are deeply intertwined with the identities, traditions, and worldviews of the communities speaking them. Addressing the pervasive influence of dominant languages and cultures is essential for revitalisation initiatives (Harrison, 2020). Such initiatives should promote the cultural relevance of the endangered language within the community and integrate it into traditional practices, cultural events, and educational settings to foster a sense of pride and continuity (Harrison, 2020). This cultural integration, in turn, helps counteract societal forces such as urbanisation, migration, and education that contribute to language endangerment (Lewis & Simons, 2010). By weaving endangered language into everyday social interactions, including education, media, and commerce, its relevance is revitalised, encouraging its use among younger generations (Lewis & Simons, 2010). Furthermore, delving into the historical trajectory of these languages and drawing on records, practices, and narratives can rekindle community engagement and strengthen a language's cultural significance (Krauss, 1992). This historical connection fosters a sense of ownership among community members, making contextualising endangered languages an academic exercise and a vital step towards effective revitalisation (Bhuvaneswari, 2022). Tailoring efforts to the specific needs and circumstances of the language community is key to preserving these linguistic treasures and safeguarding their rich cultural diversity (Bhuvaneswari, 2022).

Digital Archives: A New Frontier in Language Preservation

Digital archives have emerged as a powerful tool for language preservation, offering a comprehensive solution to safeguard the world's linguistic heritage. These cutting-edge repositories enable digitisation and detailed documentation of historical texts and facilitate intentional learning and cultural understanding. They play a crucial role in preserving endangered languages by efficiently managing large volumes of data. Furthermore, digital archives foster interdisciplinary collaboration by providing online access to language data and supporting various research and educational activities. However, adopting semantic web languages in these archives raises critical questions regarding data access and ownership, necessitating a nuanced approach to digital archiving.

Digital archives are essential in preserving literary heritage and providing new opportunities for cultural and artistic experience. Although they offer several advantages, accessibility issues for some communities, such as metadata and data standardisation challenges, remain a concern. Collaborative projects such as the Cherokee Language Digital Archive emphasise the importance of community involvement in language revitalisation. University institutional repositories and blockchain systems offer innovative solutions for digital language archives, helping preserve them for future generations. These archives promote exploration and appreciation of the world's linguistic diversity.

Scope and Significance of Preservation Techniques

Preserving endangered languages is crucial for maintaining linguistic and cultural diversity, and the techniques for revitalising these languages extend beyond documentation. These methods involve recording and promoting the use of endangered languages in daily communication and preserving linguistic and cultural heritage for future generations. Preservation efforts include collecting and documenting audio-visual materials, transcriptions, and translations and actively promoting and teaching these languages through educational programs and initiatives. Digital archives and AI-based tools play a central role in preserving primary data. However, it is essential to note that specific digital archives may have limited roles in supporting revitalisation efforts.

Integrating technology, particularly AI-based language-learning approaches, is instrumental in fostering early exposure to and appreciation for endangered languages, and a needs-based approach is necessary to evaluate the most suitable technological tools for language revitalisation. Additionally, addressing challenges in the usability of digital collections for endangered languages is crucial for making these resources more practical for revitalisation. Collective intelligence-based systems for language revitalisation, as explored by Mirza (2017), present new opportunities for community engagement.

Aims and Objectives of the Study

The primary goal of this study was to develop a robust digital archive that employs artificial intelligence (AI) to revitalise endangered languages. This endeavour entailed the creation of an extensive digital repository to store, manage, and access language-related data. As a central hub, this archive houses linguistic materials, including audio, text, and video content, to preserve and revitalise endangered languages. Furthermore, the study incorporated AI-driven tools for in-depth language analysis to enhance the accuracy and completeness of language documentation. The ultimate objective is to preserve and revitalise endangered languages through accessible language-learning resources and community engagement. This study also evaluated the impact of AI and digital archives on language revitalisation and assessed their success in promoting linguistic diversity and cultural preservation. By demonstrating the potential of AI in enhancing and preserving linguistic and cultural diversity through a scalable and sustainable model, this study contributes to the field of linguistic preservation.

Review of Related Literature:

Historical Overview of Language Endangerment

Language endangerment, a significant concern of the 21st century, risks the irreversible loss of unique human experiences and cultural expressions associated with language extinction (Gorenflo et al., 2012; Roche & Tsomu, 2018). It is estimated that due to economic globalisation, urbanisation, and marginalisation of minority languages, the world may lose 50-90% of its languages by the end of this century (Maffi, 2005), leading to high rates of endangerment (Lee et al., 2022). This issue extends beyond linguistic aspects, impacting cultural and biological diversity (Hildebrandt, 2018; Gorenflo et al., 2012). Efforts to mitigate language endangerment have led to courses focusing on language endangerment, preservation, revitalisation, and the sociocultural, economic, and political factors exacerbating the depletion of linguistic diversity (Sharma, 2021). Additionally, computational modelling and artificial intelligence techniques are being explored to detect factors leading to language endangerment, emphasising the interdisciplinary nature of this issue (Koreinik 2011). The representation of agency in the discourse on language endangerment has been studied, highlighting the multifaceted approaches required to understand and mitigate language loss. In conclusion, language endangerment is a complex and multidimensional issue requiring interdisciplinary approaches to address the interconnectedness of linguistic, cultural, and biological diversity. The predictions of the disappearance of a significant percentage of the world's languages by the end of the century underscore the urgency of efforts to preserve and revitalise endangered languages.

Development and Evolution of Digital Archiving

Digital archives and preservation techniques are pivotal in language revitalisation efforts. Evaluating digital language archives development platforms provides insights into selecting suitable tools for building these archives, ensuring their effectiveness in preserving endangered

languages (Bharti & Singh, 2022). Community-centred archives, such as the Cherokee language archive, highlight the importance of community collaboration in language preservation (Snead, 2023). Open-source solutions like Archives Space contribute to the evolution of digital archives by offering innovative platforms for preserving linguistic heritage (Sarkar & Biswas, 2020). The conceptualisation of digital preservation strategies in archival institutions underscores the need for continuous adaptation to ensure long-term accessibility of digital information (Ismail & Affandy, 2018). A literature review on sustaining accessibility through digital preservation emphasises the ongoing efforts to develop effective strategies for preserving digital information, which is crucial for the sustained accessibility of language archives (D Burda& Teuteberg, 2013). In conclusion, the development and evolution of digital archiving for language revitalisation are shaped by continuous evaluation of development platforms, community-centred approaches, open-source solutions, and the ongoing refinement of digital preservation strategies, all essential for the effective preservation and revitalisation of endangered languages.

Preservation Techniques: Past and Present

Digital archives and preservation techniques have been instrumental in revitalising endangered languages. Evaluation and analysis of digital language archives development platforms provide insights into selecting suitable tools for building these archives, ensuring their effectiveness in preserving endangered languages (Bharti & Singh, 2022). SiDHELA, India's first endangered language archive, reflects ongoing efforts to develop practical solutions for preserving and promoting endangered languages (Narayanan, 2020). The need for trusted repositories has been emphasised, highlighting the importance of reliable platforms for preserving linguistic diversity (Ferreira et al., 2021). Best practices for information architecture, organisation, and retrieval in digital language archives within university institutional repositories have been identified as a viable collaborative solution for researchers (Vann, 2021). Collaboration with language community members to enrich ethnographic descriptions in a language archive has been recognised as a valuable approach to capturing and preserving indigenous knowledge and language heritage (Burke, 2021). Comparative analysis of free-text metadata in language archives has shed light on the richness of information available, contributing to the preservation and documentation of endangered languages (Burke & Zavalina, 2020). The exploration of information organisation in language archives has highlighted the challenges and complexities oforganising and retrieving archived data, emphasising the need for effective organisational strategies (Burke & Zavalina, 2019). Attitudes towards endangered languages have been studied extensively, providing valuable insights that can inform current revitalisation efforts and policies (Heinrich, 2015). Integrating documentation and revitalisation through innovative approaches, such as language apps, has been identified as promising for language preservation and revitalisation (Little, 2017). These references demonstrate the multifaceted and evolving nature of digital archives and preservation techniques, underscoring their continued significance in the past and present for revitalising endangered languages.

Gap Analysis in Current Research

While significant strides have been made in research on digital archives and preservation techniques for revitalising endangered languages, gaps still need to be found. More research is needed on effective strategies for community involvement in developing and maintaining digital archives (Burke, 2021). The accessibility and usability of digital archives for non-expert users must be improved (Burke & Zavalina, 2019). The potential of innovative approaches, such as language apps, for language preservation and revitalisation needs further exploration (Little,

2017). More research is required on long-term digital preservation strategies (Ferreira et al., 2021) and on policy and funding issues related to developing and maintaining digital archives for endangered languages. Addressing these gaps could enhance the effectiveness of digital archives and preservation techniques in revitalising endangered languages.

Methodological Framework:

Design and Strategy of the Study

This study employed a qualitative research methodology with an interpretive perspective, utilising participatory observation techniques to gather linguistic data within indigenous communities, as shown in Figure 1. Direct community involvement, combined with audio recording and linguistic software tools such as ELAN and FLEx, ensures accurate data annotation, segmentation, and storage. As emphasised by Ellery et al. (2018), the implementation of structured approaches, including training and post-training assessments, is essential to ensuring data accuracy and reliability ("Using Community Members to Collect Observational Data: Observer Training and Data Quality Assessment," 2018). Smylie, Kaplan-Myrth, and McShane (2009) also highlight the effectiveness of focus groups and key informant interviews in collecting comprehensive linguistic data ("Indigenous Knowledge Translation: Baseline Findings in a Qualitative Study of the Pathways of Health Knowledge in Three Indigenous Communities in Canada," 2009). Whiteside (2013) stressed the importance of incorporating transnationalism concepts into participatory research methods ("Research on transnational Yucatec Maya-speakers negotiating multilingual California", 2013), while Evans et al. (2009) advocated for the integration of indigenous methodologies with participatory action research ("Common Insights, Differing Methodologies", 2009). These methodologies underline the importance of community engagement, structured processes, and advanced linguistic tools for effectively collecting linguistic data in indigenous settings.

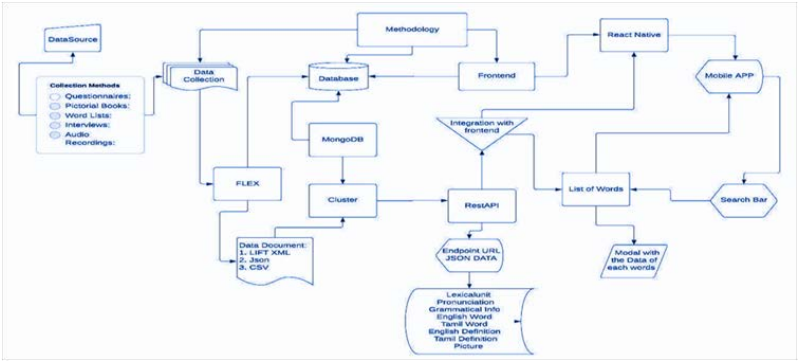

Figure 1 Unified Model Language (UML) Diagram

Data Acquisition and Processing

Integrating advanced Natural Language Processing (NLP) technologies and AI-based tools has become increasingly essential for preserving endangered languages. Our comprehensive NLP-driven language preservation system, equipped with the newly introduced NLP Model, is critical for safeguarding linguistic treasures, as shown in Figure 2. This model excels in tokenizing and analysing texts in endangered languages, which is vital for the accurate processing and preservation of these languages (Bird et al., 2009; Krafft & Kusters, 2016). Central to this system is a meticulous data preparation step that processes specific tokens crucial for language model training. These tokens include endangered language-specific tokens, which capture unique linguistic features, and regional dialect-specific tokens that account for local variations (Lewis & Simons, 2010; Fishman,

1991). Additionally, idiomatic phrases and expressions are vital for reflecting cultural nuances and are processed (Harrison, 2020).

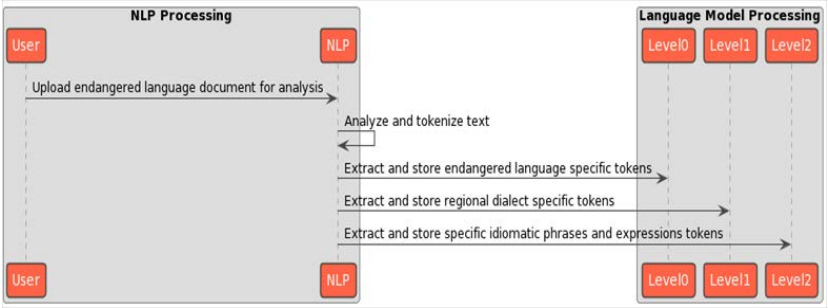

Figure 2 Language Model Processing

Following data preparation, the language model training process commenced. Each level of the language model – Level 0 (Foundation Model), Level 1 (Regional Dialect Model), and Level 2 (Idiomatic Phrase and Expression Model) – is carefully trained to capture the respective features of the endangered language, its regional dialects, and idiomatic expressions (Clyne, 2003; Bhuvaneswari, 2022). This hierarchical training approach ensures a comprehensive understanding and preservation of diverse forms of language.The system is essential for updating AI models, which are critical for optimising language models. This step integrates new findings or insights from text analysis and processing stages, ensuring that AI models remain current and effective in language preservation efforts (Robinson & Yip, 2017).

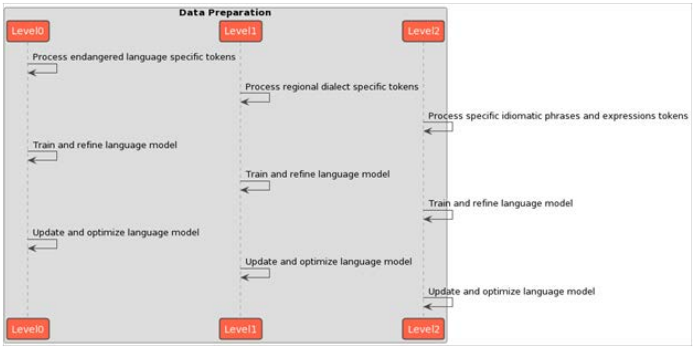

Figure 3 Data Processing

Tools and Techniques in Digital Archiving

Digital archiving involves the implementation of various tools and methodologies to secure and manage digital data. These tools and methods guarantee the lasting accessibility, dependability, and validity of digital information while protecting it from potential risks, such as data loss, corruption, and outdated technology.

This Framework is Described in Detail below

- Data Collection and Storage: This component serves as the foundation for the archives. This involves collecting and managing multimedia data related to endangered languages, including audio recordings, videos, and text documents.
- Digital Language Archive: Central to the framework, this component is a secure long-term repository for language documentation collections. It stores the data collected from the Data Collection and Storage components.

- Accessibility: The archive was designed to be freely accessible to researchers, communities, and the public. This ensures that the data are available for future generations, broadening the scope of research and community engagement.
- Usage and Access Policy: This component outlines the terms of use for the archive's materials and ensures the ethical and appropriate use of the data.
- Language Documentation: This area supports language documentation efforts, including automatic transcription and analysis and the creation of written records and language resources.
- Collaboration with Local Communities: The framework emphasises collaboration between local communities and researchers in data collection, ensuring culturally sensitive and respectful preservation efforts.
- Technological Advancements: Integrate AI and machine learning technologies to enhance the efficiency and effectiveness of language preservation and revitalisation efforts.
- Training and Support: This component provides the necessary training and support to depositors and users, facilitating collection creation, preservation, and data discovery and access.

Digital archives consist of several interdependent components crucial for proper functioning and success. The process begins with Data Collection and Storage, followed by secure preservation in the Digital Language Archive. Accessibility, Usage, and Access Policy are instrumental in ensuring ethical and widespread utilisation of the archive, whereas Language Documentation and Community

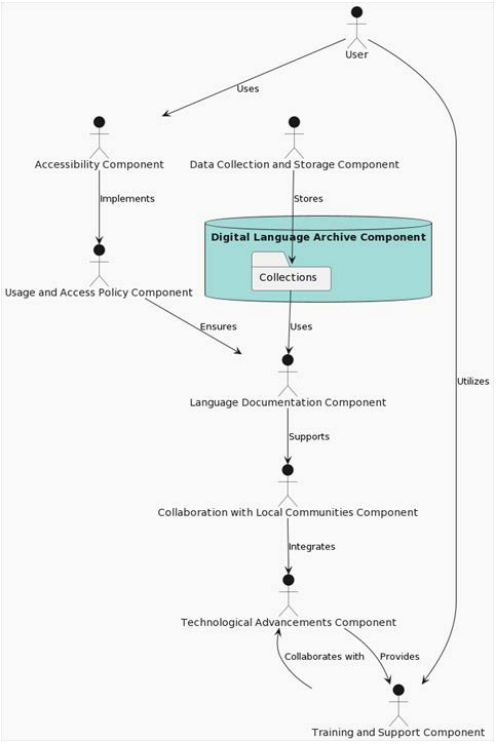

Figure 4 Digital Archive Framework

Collaboration contribute to its cultural relevance and sensitivity. The role of Technological Advancements is to enhance the archive's capabilities, and Training and Support are provided to ensure effective utilisation and sustainability.

Conclusions

This research represents a momentous breakthrough in the field of linguistic preservation. This demonstrated the transformative potential of integrating digital archives with AI-based tools. No longer mere repositories of language data, digital archives have evolved into platforms that significantly enhance comprehension and appreciation of linguistic diversity. The application of Natural Language Processing (NLP) and Artificial Intelligence (AI) technologies has opened new avenues for the analysis, processing, and revitalisation of endangered languages, ensuring their accessibility to both future generations and researchers. Thus, the importance of continuous innovation and collaboration in this domain cannot be overstated. The seamless integration of cutting-edge technology with traditional archival practices offers a promising pathway for safeguarding linguistic diversity and cultural heritage. However, it is imperative to address challenges related to usability, accessibility, and ethical considerations to fully realise the potential of digital archives and AI in language preservation. This perspective aligns with the insights of Cushing & Osti (2022) and Borgman, Scharnhorst, & Golshan (2018). This study significantly contributes to the field of linguistic preservation by providing invaluable insights and suggesting directions for future research on digital archiving and AI-based language revitalisation. This underscores the necessity of adopting a balanced approach that combines technological advancements with a thorough understanding of the complexities inherent in language preservation, ensuring that these efforts meaningfully contribute to protecting our global linguistic heritage.

References

1. Baggio, S. I Phonogrammarchive di Berlino e Vienna: Un Banco di Prova per i Linguisti. *Lingua E Stile*, 54, 2019, 95-118.
2. Berez, A. L. The Digital Archiving of Endangered Language Oral Traditions: Kaipuleohone at the University of Hawai‘i and C’ek’aedi Hwnax in Alaska. *Oral Tradition*, 28, 2013.
3. Bird, S., Simons, G., & Huang, C-R. The Open Language Archives Community and Asian Language Resources. *ArXiv*, 2001.
4. Brooks, J. D. On Training in Language Documentation and Capacity Building in Papua New Guinea: A Response to Bird et al. *Language Documentation & Conservation*, 9, 2015, 1-9.
5. Burke, M., Zavalina, O. L., Phillips, M. E., & Chelliah, S. Organization of Knowledge and Information in Digital Archives of Language Materials. *Journal of Library Metadata*, 20, 2020, 185-217.
6. Burke, M., & Zavalina, O. L. Identifying Challenges for Information Organization in Language Archives: Preliminary Findings. *Proceedings of the Association for Information Science and Technology*, 2020, 622-629.
7. Burke, M. Collaborating with Language Community Members to Enrich Ethnographic Descriptions in a Language Archive. *Proceedings of the International Workshop on Digital Language Archives*, 2021.
8. Gorenflo, L., Romaine, S., Mittermeier, R., & Walker-Painemilla, K. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*, 2012.
9. Hall, T. A. Syllable Structure and Syllable-Related Processes in German. *Language*, 69, 1992.
10. Hildebrandt, K. Teaching about endangered languages in the undergraduate curriculum. *Language and Linguistics Compass*, 12(7), 2018.
11. Jacobson, M., Michailovsky, B., & Lowe, J. B. Linguistic Documents Synchronizing Sound and Text. *Speech Communication*, 33, 2001, 79-96.
12. Khait, I., Lukschy, L., & Seyfeddinipur, M. Linguistic Archives and Language Communities Questionnaire. Proceedings of the International Workshop on Digital Language Archives, 2021.

13. Koreinik, K. Agency lost in the discourse of language endangerment: Nominalisation in discourse about south Estonian. Estonian Papers in Applied Linguistics, 7, 2011, 77-94.
14. Lee, N., Siew, C., & Ng, N. The network nature of language endangerment hotspots. *Scientific Reports*, 12(1), 2022.
15. Maffi, L. Linguistic, cultural, and biological diversity. *Annual Review of Anthropology*, 34, 2005, 599-617.
16. Powell, T. W., Müller, N., & Ball, M. Electronic Publishing: Opportunities and Challenges for Clinical Linguistics and Phonetics. *Clinical Linguistics & Phonetics*, 17, 2003, 421-426.
17. Rajagopalan, K. Prescription, Language Politics and the Field of Applied Linguistics: A Tribute to Prof. Alan Davies. *Language & Communication*, 57, 2017, 22-28.
18. Roche, G., & Tsomu, Y. Tibet's invisible languages and China's language endangerment crisis: Lessons from the gochang language of western sichuan. *The China Quarterly*, 2018, 186-210.
19. Sharma, D. Early detection of factors, including pandemics and disasters, leading to language endangerment: thinking statistically. *Iars International Research Journal*, 11(1), 2021, 31-35.
20. Shnukal, A. A Selected Bibliography of the Traditional Languages of Torres Strait. *Australian Aboriginal Studies*, 1998.
21. Tumbe, C. Corpus Linguistics, Newspaper Archives and Historical Research Methods. *Journal of Management History*, 2019.
22. Victoria, S. Psycholinguistic Analysis of Lexical-Semantic Structure in Linguistic Consciousness of Russian, English and German Native Speakers. *Training, Language and Culture*, 2017, 54-70.
23. Weber, T. Conceptualising Language Archives through Legacy Materials. *The Electronic Library*, 40, 2022, 525-538.
24. Wells, R. S. Archiving and Language Typology. *International Journal of American Linguistics*, 20, 1954, 101-107.