2025

Designing Boundary Objects

A COMMUNITY-CENTRED APPROACH TO ENDANGERED LANGUAGE REVITALISATION

PATRIK BUGÁR

AALBORG UNIVERSITY

Master's Thesis

Information Studies

Project title:	Designing for Revitalisation: A Community-Centred Approach to Endangered Language Engagement
Author:	Patrik Bugár
Supervisor:	Rikke Magnussen
Programme:	Information Studies
Semester:	4th
Project Period:	February – May 2025
Date of Delivery:	1st June 2025
Number of Characters:	176.665
Pages:	72

Abstract

This thesis explores how artificial intelligence (AI) can support the preservation and revitalisation of endangered languages through community-centred and ethically grounded design. Despite the growing interest in AI for language documentation and learning, many tools still fail to reflect the cultural complexity, dialectal diversity, and lived experiences of speaker communities. The project investigates how an AI-powered digital tool can be designed to bridge the gap between native speakers and language learners (heritage or second language) without imposing a uniform approach. Using mixed-methods research design, the study combines a systematic literature review, online surveys, and semi-structured interviews with stakeholders from various endangered language contexts. Quantitative data provides insight into engagement patterns, while qualitative data uncovers deeper tensions related to identity, trust in technology, and access to resources. The analysis is informed by Wenger's theory of Communities of Practice, the concept of boundary objects, and Gee & Hayes' framework of affinity spaces. Findings highlight a desire for digital tools that are modular, co-governed, and capable of supporting regional variation and cultural authenticity. Based on these insights, the study presents a conceptual design prototype of a digital platform aimed at supporting AI-assisted language revitalisation. The thesis contributes to current debates within information studies, human-computer interaction (HCI), and language technology by offering design directions that centre linguistic justice, community agency, and epistemic diversity.

KEYWORDS: endangered languages, artificial intelligence, co-design, language revitalisation, communities of practice

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Rikke Magnussen, for her guidance, encouragement, and thoughtful feedback throughout this project. I have deeply appreciated our discussions around design, participation, and sociotechnical systems, which have helped shape both this thesis and my broader academic development. I would also like to thank all the professors and staff I encountered during the Information Studies programme at Aalborg University, Copenhagen. The insights and knowledge gained over the course of my studies will serve me well in my future career. Special thanks go to the interview participants O, M, L, D, S, and A for generously sharing their time, experiences, and perspectives. Their contributions have greatly enriched this project.

Thank you!

Patrik Bugár

Copenhagen, 2025

Table of Contents

1. Introduction

1.1 Background

The importance of endangered language preservation

One of the fundamental aspects of human cultural and intellectual heritage is linguistic diversity, yet many languages worldwide risk becoming extinct. There are approximately 7,000 languages spoken today, of which nearly half are predicted to disappear within the next century if current trends continue (Zhong et al., 2024). Language loss often results not from natural linguistic evolution but from political, economic, and social pressures that marginalise minority communities and discourage intergenerational transmission (Liu et al., 2022).

Endangered languages encode local ecological knowledge, oral traditions, ancestral history, and culturally specific worldviews. Their extinction means not only the loss of a unique communication system but also the erasure of the epistemologies embedded within them (Low et al., 2022). Language is a cornerstone of cultural identity and social cohesion for many communities, and its preservation is intimately tied to broader issues of self-determination, historical justice, and indigenous rights (Mainzinger, 2024).

The loss of linguistic diversity also restricts the field of linguistics by reducing the range of structures and typologies available for studying human cognition, syntax, semantics, and phonology (Evans & Levinson, 2009). The preservation and revitalisation of endangered languages are therefore both a moral and a scientific necessity.

Current advancements in artificial intelligence (AI), particularly in natural language processing (NLP), machine learning (ML), and speech recognition, present new opportunities for language documentation and revitalisation (Liu et al., 2022). AI-driven tools can assist in transcribing oral histories, developing digital dictionaries, automating translation services, and creating interactive language-learning applications. These technologies have real potential in mitigating language loss, especially in cases where native speakers are scarce or geographically dispersed, by enabling learners to engage asynchronously, promote intergenerational interaction, and reduce the burden on fluent speakers by automating routine language tasks (Mager et al., 2023)

These advances, however, are not without critical challenges. Most AI systems rely on large volumes of high-quality training data, which is often something that many endangered languages fundamentally lack. This data scarcity, combined with dialectal diversity and under-resourced orthographies, often leads to unreliable or superficial AI outputs, sometimes referred to as “digital pollution” (Zhong et al., 2024). Furthermore, there are pressing concerns about data sovereignty, ethical governance, and the risk of technologies being designed without community involvement (Low et al., 2022; Liu et al., 2022).

These tensions highlight the need for more community-centred and ethically grounded design approaches that prioritise linguistic self-determination over technological novelty.

This thesis addresses a central problem, namely the gap between the needs of endangered language communities and the assumptions often embedded in AI-driven tools. While many technologies claim to support language revitalisation, they frequently fail to account for dialectal variations, cultural nuance, and community-defined success metrics. Moreover, speaker communities are not monolithic as tensions can and do arise, for instance, between native speakers and new learners regarding language standardisation, pronunciation norms, and ownership of linguistic resources.

In response, this thesis investigates the challenges that endangered language speakers and learners face and explores how an AI-powered, community-governed digital tool can serve as a boundary object (Star & Griesemer, 1989), by enabling collaboration across speaker roles without flattening cultural complexity. The thesis draws on both qualitative and quantitative data across multiple endangered language contexts, reframing preservation as a sociotechnical and co-designed process rather than a purely technical problem.

Rather than focusing on any one specific endangered language community, it addresses endangered language communities more broadly, recognising common patterns across diverse contexts, such as decentralised knowledge systems, fragmented digital resources, and concerns about technological overreach. The design proposed in this paper is not prescriptive but modular, adaptable, and grounded in the principles of mutual engagement, joint enterprise, and shared repertoire drawn from Wenger's (1998) theory of communities of practice.

Ultimately, this work argues that AI-driven tools must shift from treating languages as static data objects to viewing them as living practices embedded in dynamic, relational communities. Designing for such complexity requires both technical sensitivity and cultural humility, and a sustained commitment to community-led, ethical innovation.

1.2 Problem statement

To establish the focus of this research, the following problem statement was created:

What challenges do endangered language speakers face, and how can we design an AI-driven digital tool to support the preservation and revitalisation of endangered languages and their associated cultures through community engagement?

1.3 Research Questions

Three research questions were formed based on the problem statement to better explore the problem area. The first one seeks to explore the current state of AI-driven approaches employed in the field of endangered language preservation.

"What existing AI-driven approaches have been used for the preservation and revitalisation of endangered languages, and what are their advantages and limitations?"

The second research question examines the needs and challenges present in endangered language communities.

"What are the linguistic, cultural, and technological needs and challenges of endangered language communities?"(Initially framed in the context of Uralic endangered language communities, this was later broadened in scope – see Chapter 4)

The final research question aims to find out how an AI-driven digital tool can be designed to effectively contribute to endangered language preservation and revitalisation that addresses community needs.

"How can an AI-driven digital tool be designed to effectively support the preservation and revitalisation of endangered languages based on these needs?"

1.4 Structure of the Thesis

Figure 1 presents a visual overview of the structure of the report. The process begins with the introduction in Chapter 1, where the overall problem area is outlined, and the research questions are formulated to guide the investigation. This is followed by outlining key terms and existing digital tools relevant to the field of endangered language preservation, including both AI-powered applications and community-led platforms. This chapter provides the applied context necessary for understanding the problem field. Chapter 3 presents the literature review, situating the study within current academic research on community-centred design, language revitalisation, and ethical technology development. Moving on to Chapter 4, the methodology is described in detail, including the research design, and the collection, processing, and analysis of data. Later chapters introduce the theoretical framework employed in this project and its application in the analysis (Chapters 5 and 6). From the insights gained from the analysis, the prototype, with the corresponding design principles employed, will be presented (Chapter 7). In the penultimate chapter, the research questions are answered, and the problem statement is discussed. At last, in the conclusion, the methodological choices are reflected upon, and potential directions of future work are discussed.

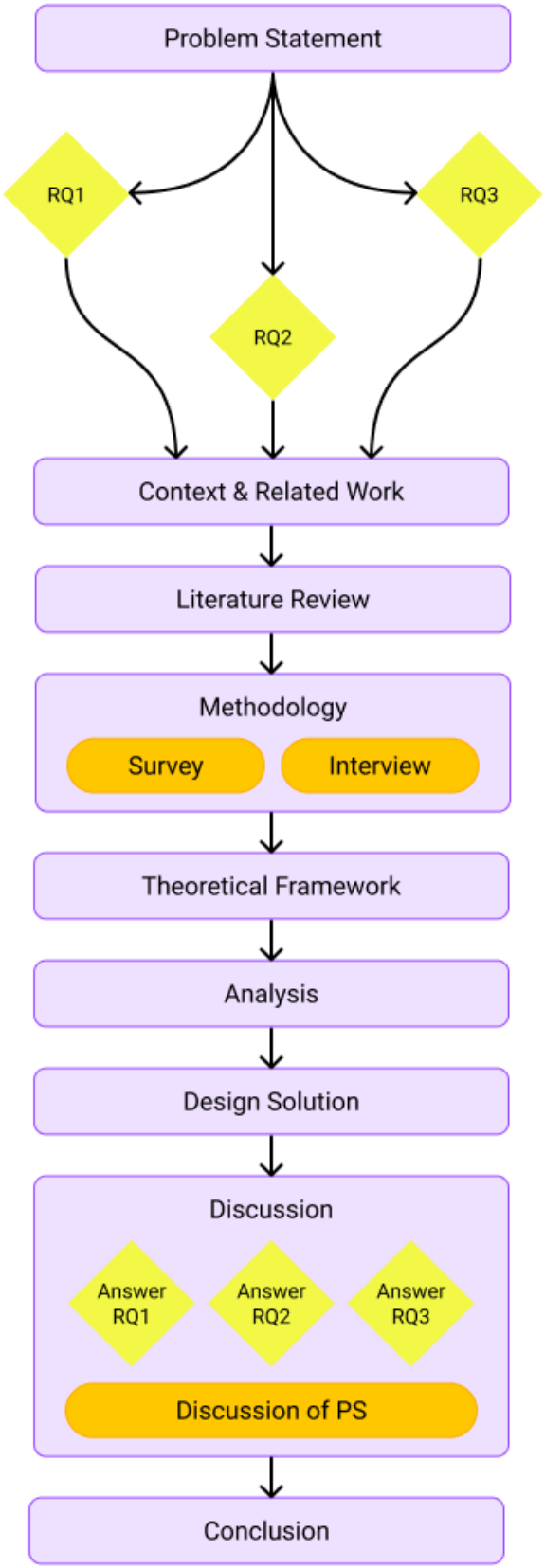

Figure 1: Visualisation of the project structure.

1.5 Significance & Limitations

This study contributes to the global effort to preserve endangered languages by exploring how artificial intelligence can serve as a scaffold for revitalisation efforts when guided by ethical, community-led design. It highlights the role of community agency in shaping digital tools, positioning language preservation not only as a technical challenge but as a socio-cultural and political one. Rather than relying on top-down models, the research emphasises co-creation, aiming to support community-defined goals, dialectal diversity, and culturally embedded practices. In doing so, it extends conversations in human-computer interaction (HCI), language technology, and critical design by proposing a boundary object that can navigate the complex terrain of multilingual, socially embedded language ecologies.

The project also demonstrates how emerging technologies such as AI can be sensitively applied in low-resource settings. It adds to the growing conversations surrounding the biases and shortcomings of mainstream language technologies and calls for tools that support fairness across languages and respect for different ways of understanding and using language.

However, the study is not without its limitations. The broad nature of the scope means that findings are necessarily broad rather than tied to a specific linguistic community. Resource constraints also limited the development and testing of a fully operational prototype. The design remains conceptual, informed by participatory research but not yet implemented at scale. Furthermore, while care was taken to foreground community voices, the diversity and complexity of endangered language contexts mean that not all perspectives could be captured. Ethical engagement remains an ongoing responsibility, particularly in navigating issues of data ownership, cultural sensitivity, and long-term impact.

2. Context & Related Work

2.1 Defining Endangered Languages and AI Terminology

A language is considered endangered when it is at risk of falling out of use as its speakers shift to using other, often dominant, languages. This is typically the result of social, political, or economic pressures such as urban migration, state-imposed language policies, or globalisation (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003). UNESCO classifies language endangerment on a spectrum ranging from vulnerable to critically endangered, with some languages nearing extinction due to a lack of intergenerational transmission. Closely related to this is the concept of *low-resource languages*, which refers to languages that are often less studied, resource-scarce, less computerised, less privileged, less commonly taught, or low-density (Liu et al., 2022). Most endangered languages are in this category, and it becomes challenging to incorporate them into AI-based linguistic tools.

Artificial intelligence (AI) is broadly defined as computer programs capable of simulating human mental processes such as learning, problem-solving, and reasoning (Blasi et al., 2022).

Natural language processing (NLP) is a branch of AI specifically addressing the capacity of computers to understand, interpret, and produce human language and is therefore of specific interest in speech recognition systems, chatbots, text-to-speech systems, and machine translation (Zariquiey et al., 2022). The development of *large language models* (LLMs), which are AI models trained on vast amounts of text data to generate human-like responses, has further advanced language technologies. However, while LLMs have demonstrated remarkable capabilities, they often struggle with low-resource languages due to limited training data, which can lead to biased or incomplete outputs (Blasi et al., 2022; Joshi et al., 2020). These challenges highlight the need for structured linguistic data and AI models that are specifically designed to support endangered and minority languages.

Deep learning, a subfield of machine learning, uses multi-layered neural networks to model complex data patterns and has significantly contributed to the evolution of LLMs. Its success, however, is heavily data-dependent, which poses challenges for endangered languages with limited corpora (Avetisyan et al., 2023). A *corpus* refers to a structured collection of linguistic data, such as recorded speech, written text, or transcribed narratives, used for language analysis or AI model training. For endangered languages, corpora are often scarce, fragmented, or difficult to standardise.

Another important distinction in this thesis concerns the terms native speaker and language learner. Native speakers are understood here as individuals who have acquired the language from early childhood and who use it fluently as part of their daily lives. Language learners, on the other hand, include not only second-language learners but also heritage speakers seeking to reclaim or strengthen their linguistic competence.

Finally, two key terms central to this thesis are language preservation and revitalisation. Preservation generally refers to documenting and archiving languages, especially those at risk of extinction, so that knowledge of them can be maintained (Grenoble & Whaley, 2005). Revitalisation, on the other hand, is the active effort to increase the use, teaching, and transmission of endangered languages, often through community programmes, formal education, or technological intervention (Low et al., 2022). In many cases, both preservation and revitalisation are interconnected and necessary for sustainable language survival.

2.2 Existing Digital Tools for Endangered Languages

Several digital tools have been developed to support endangered language preservation and revitalisation efforts, leveraging AI and digital technology to enhance linguistic accessibility. One such tool is *Neurotõlge*, an AI-powered machine translation system developed by the University of Tartu specifically for Uralic languages. The system uses neural machine translation techniques to facilitate translation between Uralic and other languages, demonstrating the potential of AI in addressing linguistic barriers for low-resource languages (University of Tartu, n.d.). The *Endangered Languages Project* (ELP) is another initiative that hosts an interactive map and an extensive repository of documentation resources. ELP serves as a collaborative platform for linguists, educators, and community members, facilitating the sharing of language data and tools to support preservation and awareness efforts (Endangered Languages Project, n.d).

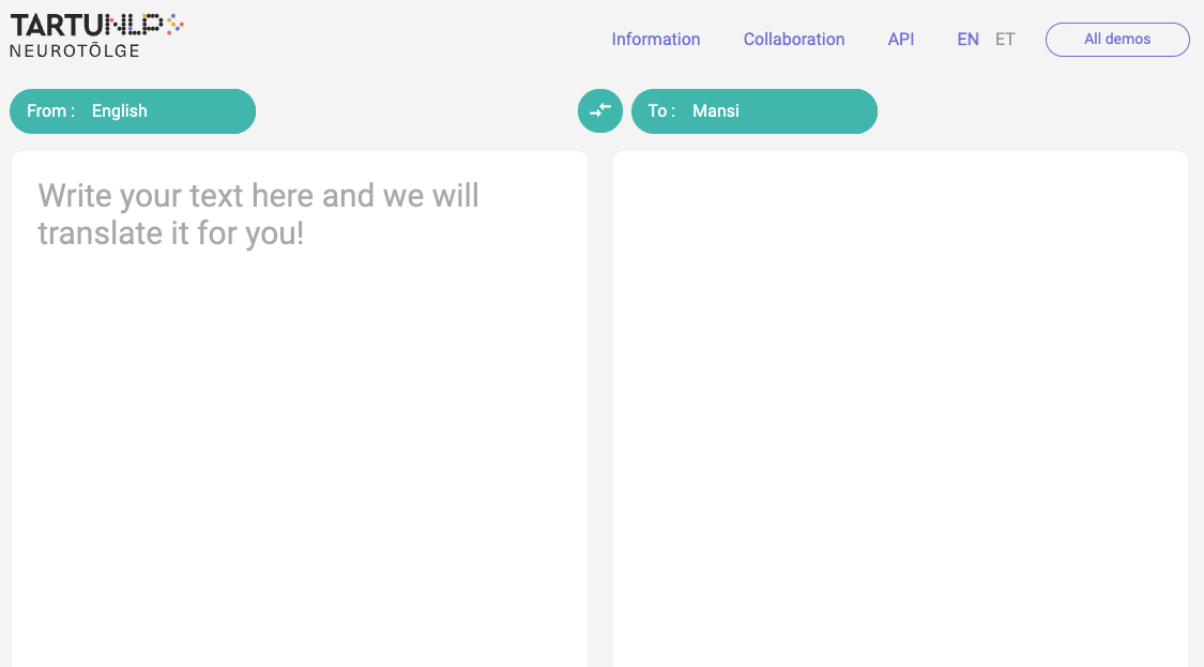

Figure 2. Neurotõlge interface, developed by the University of Tartu.

7000 Languages is a non-profit organisation dedicated to creating free online language-learning materials for endangered languages. Working directly with language communities ensures that revitalisation efforts align with cultural and linguistic authenticity, making digital tools more accessible to speakers and learners (7000 Languages, n.d.). Another notable example of an AI-powered educational tool is *LIVIU*, designed for Corsican. The platform provides conversation simulations, pronunciation assistance, and learning features that integrate cultural and linguistic knowledge into its design. It serves as an important case study for developing AI-driven solutions tailored to endangered languages, offering insights into user engagement and community-driven language preservation (Benoit, 2025).

The different approaches these tools take to endangered language documentation, learning, and revitalisation showcase how AI and digital solutions can address linguistic challenges while emphasising the importance of community engagement and ethical AI practices.

3. Literature Review

The preservation and revitalisation of endangered languages have long been key concerns in linguistics, cultural studies, and Indigenous Knowledge Systems (IKS) (Walter & Suina, 2018). With recent advancements in artificial intelligence (AI), new technological approaches have emerged, offering promising avenues for language documentation, learning, and revitalisation. AI-driven tools, such as natural language processing (NLP), speech recognition, and neural machine translation, have been explored as potential solutions to address the challenges of language loss. However, their effectiveness in the context of low-resource languages remains an open question.

This chapter provides a comprehensive review of existing literature on AI applications in endangered language preservation, focusing on six key themes:

1. **AI for Language Documentation, Archiving & Preservation** includes studies exploring how AI-powered tools support linguistic data collection, digital documentation, and automated transcription.
2. **AI & Language Learning in Revitalisation Efforts** focuses on research on AI-based learning tools, conversational AI models, and interactive educational platforms.
3. **Community-Driven AI Models for Language Revitalisation** presents studies highlighting the role of community involvement in AI development to ensure cultural and linguistic accuracy.
4. **Accessibility & Barriers in AI-Powered Language Tools** examines the challenges, such as digital accessibility, technological infrastructure, and data scarcity.
5. **Ethical Considerations in AI-Based Language Revitalisation:** Ethical concerns related to data sovereignty, AI biases, and the potential risks of AI replacing human speakers.
6. **Technical Challenges** surrounding language preservation and revitalisation in endangered language communities.

By critically analysing and mapping previous research with a focus on the design and testing of AI tools for language preservation, this chapter establishes a foundation for the theoretical framework and methodological approach presented in subsequent chapters.

3.1 Review Method

The search strategy and methodology section outlines the approach used to identify relevant academic sources, including search string formulation, database selection, and inclusion/exclusion criteria. To map and categorise existing literature by specific themes in order to identify gaps in research, a systematic mapping review method was used (Grant & Booth, 2009). This involved defining the scope of the review, as well as the inclusion and exclusion criteria; identifying potential studies through literature searches using keywords; screening the abstracts of studies to ensure they met the inclusion criteria; and finally, categorising the studies.

The thematic analysis then categorises findings from the reviewed studies, identifying recurring trends and gaps in the field. Finally, the research gaps section highlights the limitations of existing research and presents the rationale for this thesis, which aims to explore the design of AI-powered tools for endangered language revitalisation, grounded in community practices and ethical engagement across multiple linguistic contexts, including but not limited to Uralic languages.

Search String Formulation

To ensure a comprehensive and systematic review of relevant literature, a structured search string was developed to identify peer-reviewed research related to artificial intelligence and endangered language preservation. The search string was formulated using Boolean operators to combine keywords related to endangered languages, AI technologies, and language revitalisation efforts. The final search string was as follows:

	Language Type		AI Technologies		Preservation & Revitalisation
Keywords used	"endangered language" **OR** "minority language" **OR** "low-resource language" **OR** "Uralic languages"	**AND**	"AI" **OR** "artificial intelligence" **OR** "machine learning" **OR** "natural language processing"	**AND**	"language preservation" **OR** "digital documentation" **OR** "community-driven" **OR** "revitalisation"

Table 1. Search string with number of results.

This formulation ensured that the search captured interdisciplinary research from linguistics, AI, and computational language preservation, while minimising irrelevant results.

Databases used

A multi-database search strategy was employed to retrieve relevant studies from high-quality, peer-reviewed sources. The selected databases were chosen for their relevance to linguistics, artificial intelligence, and digital humanities research. The databases and their respective justification are summarised in Table 2.

Database	Why	Results
Scopus	Research across AI, linguistics, and digital humanities	30 results
Linguistics & Language Behavior Abstracts (LLBA)	Linguistic documentation and language preservation	23 results
ACL Digital Library	Research on artificial intelligence, machine learning, and computational linguistics	2 results
IEEE Xplore	Research on engineering, computer science, and technology, including artificial intelligence and natural language processing applications	2 results
Total no. of results		57

Table 2. Summary of selected databases with their respective justifications and search results.

Inclusion & Exclusion Criteria

To maintain relevance and quality, studies were screened based on predefined inclusion and exclusion criteria. These criteria ensured that only relevant, peer-reviewed, and methodologically sound studies were included in the review.

Criterion	Inclusion – included records should contain studies of…
Scope	Preservation, documentation, or revitalisation of endangered languages with the help of AI; community-driven AI tools (AI trained on community-generated data), machine learning, NLP (natural language processing)
Technology	AI tools
Users	Communities of endangered language speakers
Publication standard	Peer-reviewed research papers, journals, and academic books in English
Publication date	2020 – 2025 to capture recent developments in AI, machine learning, and NLP as they relate to low-resource language support

Table 3. Inclusion criteria for literature review.

Applying these criteria ensured methodological rigour and eliminated outdated or non-relevant studies, allowing for a focused thematic analysis of AI applications in endangered language preservation.

PRISMA Flow Diagram

A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram was utilised to document the selection process of studies included in the literature review. This approach ensures transparency and replicability in systematic literature reviews (Hartmann, 2017).

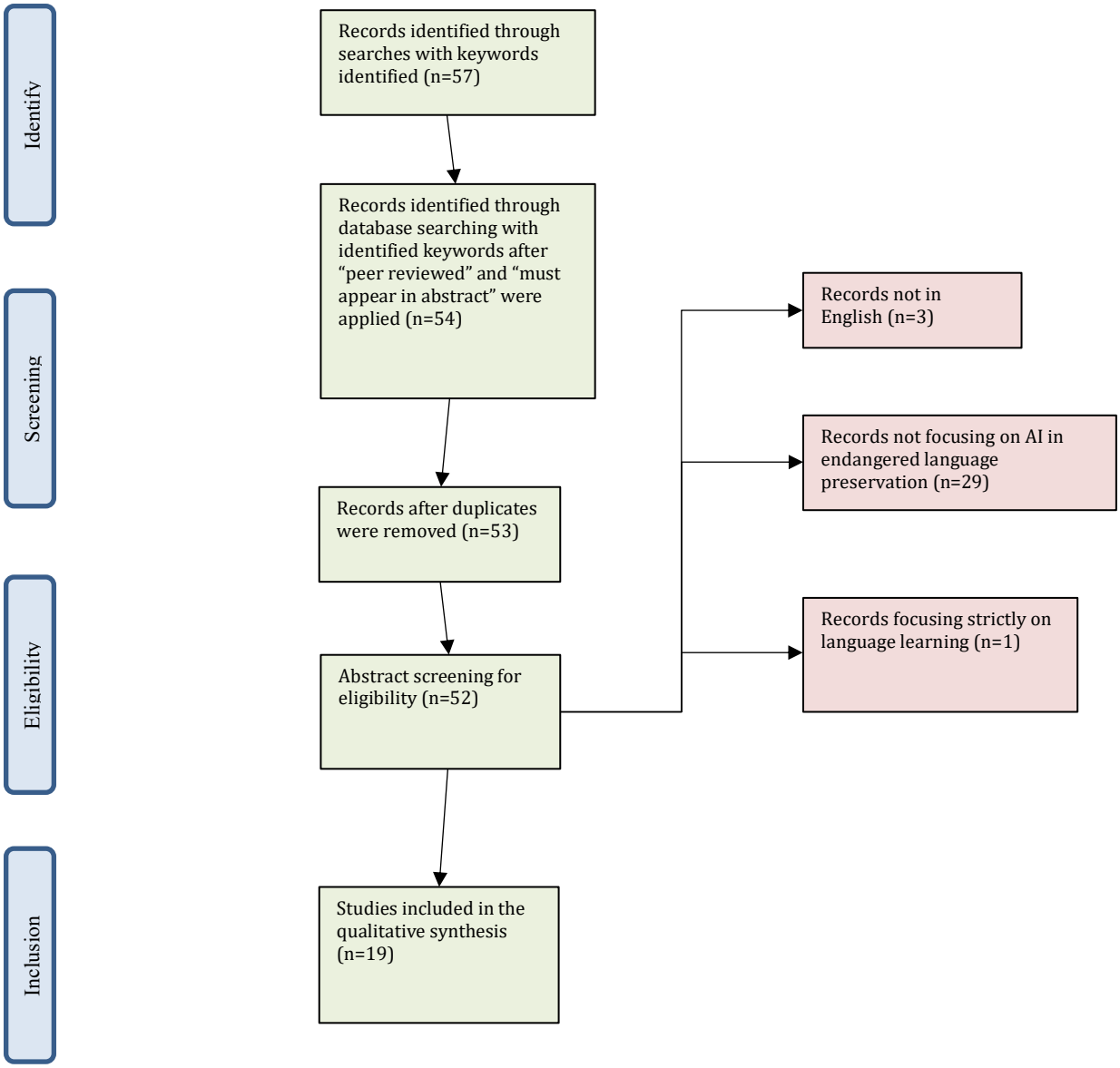

Figure 3: PRISMA flow diagram showcasing the literature selection process.

3.3 Results

The review of existing literature revealed six recurring themes in the application of AI for endangered language preservation and revitalisation. These themes emerged through a qualitative analysis of selected studies, categorising research based on its primary focus.

The thematic classification reflects key areas where AI is currently applied, as well as challenges that persist and unexplored opportunities. Table 3 provides an overview of the themes, listing the number of studies associated with each category and the key authors contributing to these discussions.

Theme	Authors	No. studies

AI for Language Documentation, Archiving & Preservation	(Soylu, 2024),(Miyagawa, 2023), (Zariquiey, 2022) (Hämäläinen, 2023), (Avetisyan, 2023) (Romero, 2024), (Chen, 2023);	7
AI & Language Learning in Revitalisation Efforts	(Orynycz, 2022) (Orynycz, 2023), (Elsner, 2023); (Dwivedi, 2020), (Chekole, 2024)	5
Community-Driven AI Models for Language Revitalisation	(Haidir, 2023), (Liu, 2022), (Mainzinger, 2024);	3
Accessibility & Barriers in AI-Powered Language Tools	(Vo, 2024), (Kochem, 2020), (Zhong, 2024)	3
Ethical Considerations	(Low, 2022), (Ondiba, 2025);	2
Technical Challenges†	(Zariquiey, 2022) (Miyagawa, 2023) (Hämäläinen, 2023) (Avetisyan, 2023) (Soylu, 2024) (Orynycz, 2023) (Elsner, 2023) (Liu, 2022) (Haidir, 2023) (Mainzinger, 2024) (Kochem, 2020) (Vo, 2024) (Dwivedi, 2020) (Low, 2022)	14

Table 3. Overview of themes emerging from the literature review.

The following sections present a detailed discussion of each theme, summarising the main findings and identifying gaps that inform the need for a community-driven platform supporting language revitalisation.

AI for Language Documentation, Archiving & Preservation

Several studies investigated how AI technology can be leveraged to aid in endangered language preservation, documentation, and archiving efforts.

Soylu & Şahin (2024) confirm that AI can support endangered language preservation through tailored learning tools, community engagement, and enhanced accessibility. They highlight the potential of speech recognition, digital storytelling, and translation tools in preserving and sharing indigenous narratives, while also emphasising the importance of community engagement. A cultural workshop organiser participating in their study highlighted that "community portals have become lifelines, where our language breathes and evolves."(Soylu & Şahin, 2024). They mention voice recognition technologies and interactive games as critical to aid in enhancing language acquisition. A linguist involved in the study noted that "translation tools and automated content generation have been game-changers in making our stories universally accessible" (Soylu & Şahin, 2024)

Zariquiey et al. (2022) emphasise the underutilisation of language documentation repositories for NLP and AI applications, a key concern for endangered language preservation. They

introduce the Computational Language Documentation and Development (CLD2) framework, highlighting the need for AI-friendly annotated data in future documentation projects (Zariquiey et al., 2022). Their framework encourages the integration of NLP into language documentation, which they deem fundamental for the development of AI tools designed to support endangered language preservation efforts.

Automatic speech recognition (ASR) has emerged as a crucial tool for documenting and preserving endangered languages by enabling transcription, pronunciation modelling, and language learning applications. Romero et al. (2024) explored tailoring ASR models for five indigenous languages of the Americas. They tested the tools with semi-fluent speakers whose key concerns were usability issues and the lack of dialectal variation. Their study emphasises the impact of dataset size, language complexity, and hyperparameter tuning on ASR accuracy. This shows that while some languages perform well, others struggle due to limited training data and linguistic complexity. In a similar study, Chen et al. (2023) explore the use of OpenAI's Whisper model for Hakka, another low-resource language. This highlights the importance of model fine-tuning in improving ASR performance for low-resource languages. They reported that users struggled with noise sensitivity and inconsistent output, suggesting the need for fine-tuned models not just linguistically but also environmentally. These findings align well with those of Zariquiey et al. (2022) on structured AI-compatible linguistic data.

Hämäläinen et al. (2023) further explore the development of infrastructure for documenting endangered Uralic languages, which includes tools for writing structured dictionaries in XML format, serving as the foundation for rule-based NLP tools (Hämäläinen, 2023). These are used for systematic documentation of Uralic languages, which could then be used to provide structured linguistic data for rule-based NLP tools (e.g., Finite State Transducers, treebanks). This has the potential to also serve as a dataset for AI-driven machine translation and speech recognition systems (Hämäläinen et al., 2023)

Finally, Avetisyan & Broneske (2023) delve into the challenges of applying NLP to Armenian. They identify gaps in the application of large language models (LLMs) to languages like Armenian and draw parallels to the challenges faced by Uralic languages. These challenges include the lack of structured, annotated corpora; poor performance on morphologically rich languages; and difficulties in adapting models trained on dominant language paradigms to endangered contexts. Furthermore, even where data exists, model adaptability is limited by the scarcity of user-centred evaluations, dialectal metadata, and cross-linguistic compatibility, all of which constrain the generalisability of current AI tools.

AI & Language Learning in Revitalisation Efforts

Orynycz (2023), in his paper, introduces the creation of an AI-powered neural machine translation (NMT) system for Lemko, a low-resource endangered language (Orynycz, 2023). Using transfer learning, a method that leverages high-resource language models to benefit low-resource languages, the system creates an online translation tool (LemkoTran.com) that applies AI for real-time language revitalisation. The study evaluates translation accuracy through BLEU score comparisons, a common metric in machine translation research, positioning AI as a tool to support language learning, revitalisation, and accessibility. However, user feedback

during testing of LemkoTran.com suggested difficulties in understanding translation confidence levels and contextual variation. The study highlights a need for transparent feedback mechanisms and culturally meaningful examples within the interface.

Another study evaluating GPT-3's ability to translate Inuktitut, another low-resource language, offers insights into the challenges faced when using AI for endangered languages (Elsner & Needle, 2023). The study explores alternative machine translation approaches and examines challenges such as AI-generated translation errors, by combining GPT-3 with human-readable dictionaries. This analysis highlights the importance of assessing AI's reliability in language preservation and ensuring that AI tools are suitable for the unique needs of endangered languages (Elsner & Needle, 2023).

Lastly, a study on neural machine translation for the endangered Khimtagne language of Ethiopia further highlights the role of AI in supporting language learning and preservation (Chekole et al., 2024). They implemented bidirectional translation between English and Khimtagne using deep learning models. While this confirms that machine translation can facilitate language accessibility and preservation, it also emphasises the challenges of limited linguistic datasets for endangered languages. This aligns with the findings of both Orynycz (2023) and Elsner & Needle (2023) on Lemko and Inuktitut, respectively, reinforcing the need for larger datasets and refined AI models to improve translation accuracy for endangered languages.

As a sidenote, AI applications have also been used to predict language endangerment (Dwivedi et al., 2020). By applying regression-based machine learning models, language vitality trends – the health and strength of a language - can be forecasted as well as language decline using census data. While the primary focus is on predicting extinction rather than revitalisation, these tools contribute to language documentation by identifying languages that are at risk and informing preservation strategies.

Community-Driven AI Models for Language Revitalisation

A handful of studies focused on endangered language revitalisation through the lens of community-driven approaches and integrating them with AI and digital tools. Haidir et al. (2023) emphasise that formal learning and cultural integration are key components to revitalisation efforts. Community-led initiatives such as dictionary development with the close involvement of native speakers, Malay Panai in this case, folklore preservation, and the use of social media are essential for preserving a language (Haidir et al., 2023). While not explicitly focusing on AI, the paper acknowledges that the use of digital tools aligns with community-driven language preservation efforts.

Liu et al. (2022) detail the importance of formal and informal learning efforts, with both being crucial for the success of AI-assisted language learning tools in real-world settings. They highlight the importance of prioritising community needs when developing language technology, and the challenges faced by low-resource language communities, including technological, cultural, and ethical – scarcity of skilled developers, lack of trust in external researchers by the community, and bias and misrepresentation - which must be considered when developing AI-powered tools for language revitalisation (Liu et al., 2022). In addition,

the paper stresses the importance of collaboration between researchers and language communities to ensure that technology is designed with input from those who will benefit most from it (Liu et al., 2022).

Further exploring community-driven language revitalisation, Mainzinger (2024) examines how NLP can support the revitalisation of the Mvskoke language. The paper presents AI as a complementary tool aiding indigenous efforts, emphasising technology's supportive role in human-led initiatives (Mainzinger, 2024). It assesses available language resources, supporting the building of AI-powered tools used in language revitalisation. Its focus on tribal engagement aligns with community-driven approaches to AI, reinforcing the notion that AI can serve as a tool to support, rather than replace.

Accessibility & Barriers in AI-Powered Language Tools

The revitalisation of low-resource languages poses their own unique challenges for AI, especially when it comes to machine translation, data scarcity, and accessibility. A study focusing on the Bahnar language of Vietnam stresses how translating from Vietnamese to Bahnar remains a significant challenge, despite government efforts, due to the extreme scarcity of linguistic resources, dictionaries, books, media, or this language (Vo et al., 2024). It emphasises how neural machine translation (NMT) has improved translation accuracy and fluency; however, the low-resource nature of the language presents significant hurdles. Vo et al. (2024) propose using transfer learning from pre-trained models, making use of the similarities between Vietnamese and Bahnar, to optimise translation quality. This approach was also validated by positive results on bilingual Vietnamese-Bahnar datasets, providing a promising solution that could be applied to other low-resource languages.

Kochem & Taylor (2020) explore the barriers indigenous and marginalised communities face when it comes to accessing digital language learning tools, labelling this digital divide a major challenge AI-based language preservation efforts face. Although the focus isn't solely on AI, the paper technological accessibility and funding issues in linguistics directly affect the effectiveness of AI tools in language preservation.

Another study by Zhong et al. (2024) evaluates how LLMs are used in linguistic research and historical documentation for underrepresented languages. They identify key challenges such as dataset availability, ethical concerns, and technical limitations. Similarly to the work done by Vo et al. (2024) and Kochem & Taylor (2020), they emphasise that data scarcity presents a fundamental challenge for AI applications in bilingual and code-switching environments, further restricting their effectiveness in endangered language preservation. These findings suggest that LLMs must be supplemented with customised, community-driven AI models that prioritise linguistic diversity and ethical data collection.

Ethical Considerations

The use of AI in language revitalisation raises ethical concerns, particularly regarding data sovereignty, cultural preservation, and the role of AI in endangered language contexts. In a study published by Low et al.(2022), they take a socio-cognitive approach to the intersection of language death, identity loss, and AI-driven preservation efforts. They explore the societal

impacts of AI and the ethical challenges that come with it, including concerns about AI replacing native speakers in documentation and the potential consequences for identity preservation (Low et al., 2022). The study frames AI as a transformative force in endangered language communities, addressing both its promise and risks within the broader societal context.

Additionally, the integration of AI-driven cybersecurity measures in endangered language preservation has been explored through a case study on the Suba language of Kenya (Ondiba, 2025).It underscores the importance of protecting linguistic data from cyber threats, particularly in corpus development, where sensitive cultural and linguistic information is stored.

Technical Challenges

AI has great potential for endangered language documentation and revitalisation. However, its effectiveness is often hindered by technical challenges that disproportionately affect low-resource languages. Many of these challenges stem from data scarcity, linguistic complexity, and model adaptability, making it difficult to integrate these languages into AI-driven language preservation efforts. This section outlines some of the key research gaps identified throughout the literature review and justifies the need for an AI-driven solution tailored towards endangered languages.

One of the main challenges of AI-driven language preservation is the lack of structured, optimised linguistic data for endangered languages. While AI tools have been successfully applied to high-resource languages, language documentation repositories for low-resource languages remain largely underutilised (Zariquiey et al., 2022). Many endangered languages lack digitalisation and structured linguistic datasets, making them difficult to incorporate into machine translation, speech recognition, and language learning applications (Miyagawa et al.). This issue is particularly relevant for Uralic languages, where structured lexicons and rule-based NLP tools are largely absent (Hämäläinen et al., 2023). Additionally, LLMs struggle with morphologically complex, low-resource languages such as Armenian, a challenge that extends to Uralic languages (Avetisyan & Broneske, 2023).

Even in cases where AI-powered tools exist for language learning and revitalisation, they often fail to integrate interactive, real-world applications that encourage sustained engagement (Soylu & Şahin, 2024). While machine translation and speech recognition systems, such as the Lemko NMT system (Orynycz, 2023) and GPT-3's evaluation for Inuktitut (Elsner & Needle, 2023), demonstrate progress, they frequently suffer from AI-generated translation errors, reducing their reliability. AI's role in education and language practice remains underdeveloped, highlighting the need for a community-driven, interactive solution that prioritises usability and accuracy.

Another key limitation is the lack of community participation in AI development. Studies stress that endangered language speakers should not be passive users but active co-creators of AI solutions (Haidir et al., 2023; Liu et al., 2022). While AI has the potential to support human-led language revitalisation efforts, it should not replace them (Mainzinger, 2024). Additionally,

ethical risks emerge when AI-powered machine translation is developed without native speaker involvement, as this can lead to linguistic misrepresentation and inaccuracies.

Beyond the technical limitations, accessibility and data scarcity remain significant challenges for AI-powered language tools. Indigenous and marginalised communities often face limited internet access and financial constraints, preventing them from engaging with digital language technologies (Taylor & Kochem, 2020) AI models also struggle to accommodate bilingual and code-switching environments, which are common in endangered language communities (Vo et al., 2024).

While AI has been used to predict language endangerment trends, existing models focus primarily on forecasting extinction rather than developing revitalisation strategies. For example, Dwivedi et al. (2020) developed machine learning models to predict language decline, but these models lack real-time community data, reducing their impact on targeted preservation efforts.

Finally, ethical concerns surrounding AI-based language revitalisation remain largely underexplored. There are concerns about AI replacing native speakers in documentation efforts, potentially leading to a loss of linguistic and cultural identity (Low et al., 2022) Without ethical oversight and meaningful community involvement, AI-based tools risk misrepresenting or erasing linguistic diversity rather than preserving it.

3.5 Summary of Literature Review Findings

This chapter has provided a comprehensive review of existing research on the role of AI in endangered language preservation and revitalisation. While prior studies highlight the potential of AI-driven tools, they also showcase critical gaps that must be addressed to create effective, community-driven solutions.

A key takeaway from the literature review is the lack of structured, AI-compatible linguistic data, such as annotated text corpora, speech datasets, or lexical datasets, particularly for morphologically complex, low-resource languages. This raises the question of how existing AI models can be adapted to function effectively in these linguistic contexts, which is further explored through interviews with language learners, community members, and other stakeholders involved in revitalisation efforts.

Another main issue is the limited real-world applications of AI tools for language learning. While conversational AI machine translation has been explored, concerns about accuracy, usability, and trustworthiness remain (Orynycz, 2023). Understanding how potential users, both language learners and native speakers, perceive and interact with AI-powered learning tools is crucial. This will be investigated through survey data and semi-structured interviews with native speakers and language learners.

The existing literature further stresses the importance of community involvement in AI-driven language preservation. While some studies highlight participatory models, there is little insight into how to implement such models effectively in practice.

Additionally, ethical concerns surrounding data sovereignty, AI bias, and the role of AI in Indigenous Knowledge Systems remain largely theoretical. The user studies will explore how these concerns are perceived by linguists, developers, and community members and whether they pose a barrier to adoption or engagement.

These areas will be explored in the next chapter through expert interviews and user surveys. The following chapter will introduce the methodological framework supporting the investigation.

4. Methodology

This chapter outlines the methodological approach employed in this study, detailing the research design, data collection methods, and analytical strategies. The study adopts a mixed-methods research design, combining both qualitative and quantitative approaches to gain a comprehensive understanding of engagement with endangered languages in online and offline spaces.

4.1 Research Design & Approach

Research design refers to a framework for the collection and analysis of data, shaping how a study addresses its research questions (Bryman, 2021). Taking into consideration the complexity of language endangerment, this study integrates quantitative and qualitative data collection methods: surveys and semi-structured interviews. This combination of methods provides a triangulated perspective, capturing both numerical trends and more in-depth insights. The following sections elaborate on the research design and approach before presenting the specific methodologies applied to each data collection method.

A quantitative approach through surveys enables the identification of patterns on a broader scale, while a qualitative approach through semi-structured interviews provides an in-depth insight into individual experiences and contextual factors. According to Bryman, a mixed-methods approach enhances validity by reducing biases inherent to single-method studies and allows for a more nuanced interpretation of findings (Bryman, 2021).

The study follows a convergent parallel design (Creswell & Plano Clark, 2018), where quantitative and qualitative data are collected simultaneously but analysed separately before being integrated in the discussion. This approach ensures that statistical trends identified in surveys can be contextualised with qualitative data, adding more depth to the analysis.

4.2 Data Collection Methods

The study employs two main data collection methods:

- **Surveys**: To quantify engagement patterns with endangered languages.
- **Semi-structured interviews**: To explore personal experiences, perceptions, and challenges related to endangered languages.

Each method is selected to contribute distinct yet complementary insights, aligning with the mixed-method approach. Initially, the study focused specifically on endangered Uralic language communities and resources. However, it quickly became clear that the participant pool was too limited to support meaningful analysis or design generalisability. After consultation with the thesis supervisor, the scope of the study was expanded to include endangered language communities more broadly. This methodological shift allowed for greater diversity of perspectives and more robust pattern identification. It also aligned with emerging themes in the literature review, suggesting that many endangered language communities, regardless of linguistic family, face similar sociotechnical and cultural challenges in the digital space.

4.2.1 Surveys

Surveys are particularly useful for identifying trends and patterns across a larger sample population (Goodman et al., 2012). In this paper, surveys are used to collect structured data on participants' language engagement habits, the challenges they face, and their attitudes towards the use of AI in endangered language preservation and revitalisation.

The survey consists of single-choice, multiple-choice, Likert-scale, and open-ended questions, allowing both quantitative measurements and some qualitative insights. The questionnaire was designed based on existing literature on language preservation and technology-assisted learning, ensuring relevance and validity.

The survey was distributed via social media channels (Reddit and Discord) with a focus on:

- Endangered language-speaking communities
- Endangered language learning communities
- Academic and activist networks focused on endangered languages.

The target population includes native speakers, heritage speakers, and language learners. The sample size aims to provide a representative overview of language engagement trends and attitudes towards AI in endangered language preservation and revitalisation across different user demographics.

Survey responses will be analysed to identify patterns in:

- Frequency and context of language use
- Challenges faced by speakers and learners
- Perceptions of AI in language preservation.

Likert-scale responses will be analysed using descriptive statistics, while open-ended responses will be coded thematically to identify key concerns and opinions.

4.2.2 Interviews

Semi-structured interviews were conducted to gain in-depth insights into personal experiences with endangered language engagement. This method allows for flexibility, enabling participants to expand on key topics while ensuring consistency across interviews (Bryman, 2021).

The interview guide (see Appendix B) was designed around the following themes:

- Language learning and community engagement
- Cultural practices and traditions linked to language
- Experiences using AI tools for language learning
- Challenges in accessing quality linguistic resources.

Participants for interviews were selected based on:

- Their engagement with endangered languages
- Their attitudes towards AI and digital tools related to endangered language preservation and revitalisation.

Recruitment was done through direct invitations via email. Interviews were conducted via video conferencing using Microsoft Teams and recorded with the consent of the participants. A thematic analysis was applied to code interview data, identifying recurring patterns and perspectives (Braun, 2006).

4.2.3 Ethical Considerations

Given the cultural sensitivity of endangered language contexts and the potential power dynamics between researchers and participants, ethical engagement was a key priority throughout this project. The study adhered to informed consent, anonymity, and data protection.

Surveys were conducted entirely anonymously. Participants were informed of the study's purpose and their right to withdraw at any point. Participation was entirely voluntary, and responses were stored securely and analysed only in aggregate form.

Semi-structured interviews were conducted with explicit, prior consent from each participant. Each interviewee was informed about the study's aim, how the data will be used, their right to withdraw, and the option to decline being recorded. All interviewees agreed to be recorded and understood the conditions of participation. Audio recordings and transcripts were stored securely, and all data were anonymised during transcription and reporting.

Throughout the analysis and reporting phases, particular care was taken to protect the identities of participants. Generic role descriptions (e.g., "heritage speaker", "language learner") were used to preserve anonymity while still offering contextual clarity. No data was shared with third parties or uploaded to cloud services without encryption.

The study also recognised the broader ethical importance of cultural and linguistic respect. Findings and design implications were grounded in community concerns, not imposed from external frameworks. This aligns with calls in the literature for ethical, community-led approaches to technical intervention in indigenous and minority language contexts (Low, 2022; Liu, 2022; Ondiba, 2025).

4.3 Data Analysis

The data analysis in this study follows the logic of a convergent parallel mixed-methods design, where quantitative and qualitative data are analysed separately and then brought together for the interpolation to generate integrated insights (Creswell, 2018). This approach supports a multidimensional understanding of endangered language engagement by enabling comparison between statistical patterns and individual lived experiences. Quantitative data collected through the survey were analysed using descriptive statistics to identify trends in language use, challenges, and perceptions of AI tools. Qualitative data from semi-structured interviews were subjected to thematic analysis, following the guidelines set out by Braun & Clarke (2006). The dual analysis not only reinforces the validity of the findings but also provides both breadth and depth in addressing the research questions.

The decision to analyse both data types separately before integrating them aligns with the study's emphasis on triangulation and supports the research aim of uncovering both general patterns and specific contextual meanings. In particular, while the survey provides a broader understanding of engagement trends across communities, the interviews offer insights into the nuanced practices, tensions, and needs that underlie those trends. Together, these analyses inform the development of a design response grounded in the lived realities of language

4.3.1 Thematic Analysis

Thematic analysis was selected as the primary method for analysing the qualitative data due to its flexibility and suitability for identifying patterns of meaning across participant narratives. Following the framework established by Braun & Clarke (2006), this study adopted a reflexive and semantic approach to thematic analysis, focusing on the explicit meanings within the data rather than seeking to interpret latent content. This choice is aligned with the study's goal of foregrounding participants' own conceptualisations of their linguistic experiences and challenges.

Braun & Clarke (2006) define thematic analysis as a "method for identifying, analysing, and reporting patterns (themes) within data" (p. 79). In this study, themes were derived inductively and iteratively through a process that involved familiarisation with the data, initial coding, theme development, and refinement. The process followed the six-phase model outlined by Braun & Clarke:

1. **Familiarisation**: Interview recordings were transcribed and read multiple times to become immersed in the content.
2. **Generating initial codes**: Codes were created manually, capturing specific actions, concepts, and participant expressions directly tied to the research questions.
3. **Searching for themes**: Related codes were clustered into potential themes, reflecting recurring patterns in how participants spoke about language use, identity, community, and digital engagement.
4. **Reviewing themes**: Themes were refined by re-reading transcripts and ensuring coherence within and between coded extracts.

5. **Defining and naming themes**: Each theme was defined based on its core meaning and was often phrased using participants' own terminology to preserve semantic integrity
6. **Producing the report**: Themes were presented alongside quotes to contextualise findings and ground the analysis in participant voices.

This process was guided by an inductive approach, allowing the data to speak for itself rather than imposing pre-determined theoretical constructs. The themes emerging from the thematic analysis were then contextualised through the lens of Wenger's (1998) communities of practice framework.

By employing a thematic analysis grounded in the explicit language of participants, the study maintains methodological transparency and ensures that the themes generated remain closely tied to the lived experiences and priorities of endangered language community members.

4.3.2 Descriptive Statistics

Descriptive statistics were used in this study to summarise and interpret the survey data, allowing for the identification of patterns, trends, and variations across the sample population. This approach provides a foundational understanding of the broader landscape of engagement with endangered languages and informs the contextual interpretation of qualitative themes. Descriptive analysis is particularly valuable in mixed-methods research for its capacity to distil large datasets into accessible summaries (Bryman, 2021). Moreover, by incorporating visual representations of statistical findings, the study aligns with best practices in information visualisation, enhancing both interpretability and transparency (Fekete et al., 2008).

To carry out the quantitative analysis, the statistical programming language R was employed within the RStudio environment (R Core Team, 2023). Data were handled and transformed using tools from the *tidyverse* ecosystem, which provides coherent and intuitive syntax for data manipulation and cleaning (Wickham, 2023). This process was particularly important given the structure of the study's datasets, which consisted of two distinct survey collections, one initially focused on Uralic languages (`uralic_data`) and a broader survey targeting endangered languages globally (`global_data`). These datasets were ultimately merged into a unified dataset, referred to as `combined_data`, allowing for a consolidated analysis that reflected the full scope of participant responses.

Data Import and Initial Cleaning

The raw survey data were exported from Microsoft Forms as CSV files and imported into R using the `read_csv()` function from the *readr* package, which preserved the column names and encoding (Wickham & Girlich, 2023). Initial preprocessing involved verifying data types, inspecting missing values, and cleaning up irregular formatting such as stray quotation marks, HTML artefacts, and line breaks. These steps ensured that the dataset could be reliably used for both statistical summaries and graphical representations. To see the full code used in the analysis see Appendix K.

Variable Renaming and Code Readability

To facilitate cleaner code and enhance readability, several survey question variables were renamed using the rename() function from the *dplyr* package (Wickham et al., 2023). For example:

- what_is_your_biggest_concern_about_using_ai_for_endangered_language_preservation_select_all_that_apply → biggest_concerns_about_ai
- how_trustworthy_do_you_find_ai_powered_tools_used_for_endangered_language_preservation → trust_in_ai
- would_you_be_interested_in_contributing_to_an_ai_powered_language_preservation_project → contribute_to_ai

Renaming variables helped simplify the script and made later interpretation of visualisations and summaries more transparent (Hinton, 2014).

Handling Multiple-Choice Responses

Several survey items allowed for multiple selections, which were stored as character strings containing multiple selections. These responses required transformation into a long format for analysis. Using str_extract_all() from the *stringr* package (Wickham, 2022), each selected item was extracted from the string and converted into individual rows using the unnest() function from the *tidyr* package. This allowed for frequency analysis of each unique response. For instance, to analyse the variable biggest_concern_about_ai, each respondent's selections were broken down into discrete rows, stripped of quotation marks, and filtered for completeness using filter().

Categorisation and Manual Re-coding

Qualitative responses to open-ended questions were categorised using pattern matching and manual review using the str_detect() function. Responses were classified into themes, including:

- Inaccurate or low-quality output
- Lack of cultural/linguistic nuance
- Data privacy & ethics
- AI replacing humans
- Environmental concerns

Responses that did not initially fit into any category were labelled "Other / Unclassified" and then manually reviewed. Two responses were recoded into existing categories based on content relevance, ensuring consistent thematic grouping. To implement the re-coding:

1. Unclassified responses were filtered with filter / category == ("Other / Unclassified")
2. Relevant entries were reassigned using a new mutate() with case_when().
3. These updated entries were merged back into the main dataset using bind_rows() after excluding the originals, resulting in biggest_concerns_categorised_final.

A small number of long-form responses that were highly specific, off-topic, or anecdotal in nature were excluded from quantitative categorisation to avoid skewing the results. These were filtered using filter(!x %in% y) and documented separately in the appendix.

Single-Choice Responses and Visualisation

For Likert-scale and single-choice variables such as levels of trust in AI tools (trust_in_ai) and interest in contribution to an AI-powered preservation project (contribute_to_ai), count() and mutate() were used to calculate proportions and create basic statistical summaries. Visualisation techniques included donut charts for proportionally distributed data and lollipop charts for frequency-based comparisons, in accordance with recommendations for effective visual communication (Fekete, 2008).

Final Dataset Preparation

All cleaned and categorised variables were stored in finalised data frames, such as biggest_concerns_categorised_final, and were stored for analysis and figure generation. All scripts were version-controlled and documented to ensure reproducibility and transparency.

4.4 Human-Centred Design

This project adopts a Human-Centred Design (HCD) methodology as its guiding design approach. HCD prioritises the needs, contexts, and lived experiences of intended users throughout the design process, with a focus on inclusivity, empathy, and iterative feedback (Norman, 2013). In the context of endangered languages, this means designing not merely for usability, but for cultural alignment, trust, and community sovereignty.

As pointed out by Mainzinger (2024), AI systems can be powerful tools in supporting revitalisation goals, but only when integrated into broader human-led efforts that reflect community-defined success. This resonates with broader critiques of AI in low-resource settings, which often caution against top-down, universalist solutions that overlook dialectal complexity, oral-first knowledge systems, and digital asymmetries (Liu et al., 2022; Taylor & Kochem, 2020)

Rather than designing a tool that assumes shared expectations of learning, success, or technological readiness, this project draws from community insights to propose a modular infrastructure. This allows users to activate only the features they deem culturally and linguistically appropriate. This orientation is consistent with participatory and community-informed design ethics, emphasising collaborative authorship, consent, and flexibility.

Moreover, user personas and qualitative themes derived from surveys and interviews were used to shape the interface architecture. These personas reflect distinct epistemic positions between native speakers and language learners, making asymmetries in access, legitimacy, and goals. This aligns with Star and Griesemer's (1989) concept of boundary objects, wherein shared tools can mediate between differing user needs without enforcing a single definition of success or fluency.

Although time constraints limited the scope for full co-design workshops, community voices were foregrounded in the analytical synthesis of needs, ensuring that the design responds to real-world frictions rather than abstract assumptions. Ultimately, HCD in this context acts not just as a technical methodology but as an ethical imperative: to centre the knowledge, priorities, and governance structures of endangered language communities in shaping the digital tools intended to support them.

This chapter has outlined the study's mixed-method approach, integrating surveys and semi-structured interviews to examine endangered language engagement and attitudes towards AI-driven tools used in endangered language preservation and revitalisation. The combination of qualitative and quantitative data ensures a holistic understanding of endangered language use, challenges faced by endangered language communities, and the potential role of AI in preservation and revitalisation efforts.

5. Theoretical Framework

Chapter 5 outlines the theoretical lenses that inform the analysis and design components of this study. Drawing on socio-cultural and participatory perspectives, the study is grounded in three interconnected frameworks: Communities of Practice (Wenger, 1998), the concept of boundary objects (Star & Griesemer, 1989), and Affinity Spaces (Gee & Hayes, 2012).These perspectives collectively support an understanding of how language engagement and revitalisation efforts operate across overlapping, and sometimes conflicting, communities. They also provide critical foundations for the design of a digital tool that aims to foster ethical and inclusive collaboration across speaker groups.

5.1 Communities of Practice (Wenger, 1998)

At the core of this study is Wenger's (1998) theory of Communities of Practice (CoP), which views learning as a social process embedded in everyday participation. A community of practice is defined by three essential elements: mutual engagement, a joint enterprise, and a shared repertoire. Mutual engagement refers to the interactions and relationships that bind members together; joint enterprises reflect the collective goals and commitments pursued by the group; and shared repertoire encompasses the symbolic and material resources (e.g. tools, languages, routines, narratives) that support participation.

In the context of endangered language communities, Wenger's framework is useful for understanding how native speakers and language learners (including heritage speakers) form overlapping social groups that engage in various practices of language use, learning, and preservation. These groups may not always have shared goals or mutual access to the same resources, but they remain interdependent in shaping the trajectory of language revitalisation.

Wenger (2002) later extended this theory in his work on social learning systems, emphasising the role of CoPs as part of broader networks of learning that evolve through boundary interactions, negotiation of meaning, and identity formation. This notion is particularly relevant to the analysis in this paper, as it highlights how participation is not only about acquiring

knowledge but about becoming recognised as a legitimate member of a community. For heritage speakers or learners, this process of identity negotiation can be complex and fraught with exclusion.

In later sections, this framework is used not only to structure the analysis of interview and survey data but also to inform the design of a digital platform that supports social learning across linguistic and cultural divides. Wenger et al.'s Seven Principles for Cultivating Communities of Practice (2002) may also inform the design component, particularly with regard to nurturing participation, enabling community stewardship, and developing a rhythm for interaction, although these are more closely examined in the discussion of the boundary object (see Chapter 6.5).

5.2 Boundary Objects

The concept of boundary objects was introduced by Star and Griesemer (1989) to describe artefacts that inhabit multiple social worlds and satisfy the informational needs of each without requiring consensus. Boundary objects are "plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites" (Star & Griesemer, 1989, p. 393). They function as mediators between communities, enabling collaboration without full alignment.

In this paper, the concept of a boundary object is used to understand the design of a digital infrastructure that aims to bring together different user groups, native speakers and language learners, who hold varying ideas of authenticity, fluency, and revitalisation success. The proposed design is therefore conceptualised not as a uniform solution, but as a flexible, co-governed space that enables coordination without consensus. It offers shared affordances while remaining adaptable to different dialects, levels of fluency, and cultural values.

The notion of boundary objects complements the CoP framework by acknowledging the practical tensions that arise when distinct communities attempt to collaborate. It also provides a design rationale for features such as modular content, regional tagging, and community-defined permissions, which allow the platform to serve different needs without imposing a single definition of what endangered language engagement "should" look like.

5.3 Affinity Spaces

While Communities of Practice are grounded in sustained relationships, mutual accountability, and identity formation, Affinity Spaces offer a different but complementary model for analysing online participation. Developed by Gee & Hayes (2012), Affinity Spaces refer to informal, interest-driven spaces in which people gather around a shared passion or goal, which is often mediated through digital platforms. Unlike traditional communities, affinity spaces do not require strong social ties, stable group membership, or collective identity. Instead, they are characterised by fluid participation, distributed expertise, and low barriers to entry.

This framework is particularly relevant for understanding how many language learners, especially heritage speakers or those outside the geographic reach of native-speaking

communities, engage with endangered languages online. Participation in Discord channels, Reddit forums, or WhatsApp groups often resembles affinity spaces more than fully-fledged communities of practice. These spaces provide access to knowledge, peer support, and cultural content without necessarily conferring membership status or legitimacy.

By incorporating the concept of affinity spaces, this study accounts for the plurality of participation modes present in the data and avoids romanticising the idea of "community". Not all learners are seeking full membership or cultural belonging; some simply want access to learning tools or casual opportunities for practice. Recognising this distinction helps ensure that the design solution does not assume a one-size-fits-all model of engagement and remains open to different levels of commitment and identity alignment.

In sum, the three theoretical frameworks employed in this study, Communities of Practice, boundary objects, and Affinity Spaces, provide complementary lenses for analysing and responding to the complex dynamics of endangered language engagement. They shape not only the analytical strategy but also the conceptual underpinnings of the design solution, which seeks to build bridges across speaker communities while respecting their differences.

6. Analysis

This chapter presents the analysis of both quantitative and qualitative data gathered through an online survey and six semi-structured interviews, in line with the mixed methods approach outlined in Chapter 4. The analysis adopts an inductive perspective, allowing empirical findings to emerge from the data while drawing interpretive depth from the Communities of Practice (CoP) framework (1998), supported by the notion of boundary objects (Star & Griesemer, 1989) and Affinity Spaces (Gee & Hayes, 2012). This theoretical lens positions endangered languages revitalisation as a socially situated learning process shaped by mutual engagement, shared repertoires, and differentiated participation.

The chapter is structured in two parts. First, Section 6.1 presents the quantitative findings from the survey data, offering a descriptive overview of participant demographics, engagement patterns, and perceptions of technology use in endangered language contexts. This section provides a broad empirical backdrop that contextualises the more nuanced themes explored in the subsequent qualitative analysis.

Second, Sections 6.2 to 6.6 present the thematic analysis of qualitative interview data, which was coded using a semantic, inductive approach following the guidelines of Braun and Clarke (2014). Five core themes – Communities, Learning, Versions, Technology, and Identity – emerged from the data, each comprising multiple sub-themes. These themes are analysed in light of Wenger's theoretical constructs as well as through the lens of Affinity Spaces and are illustrated with direct participant quotations to preserve the situated perspectives of speakers, learners, and linguists.

This chapter provides the foundation for the design exploration that follows. By examining the shared and divergent practices, values, and struggles across participants, it sets the stage for

identifying opportunities for ethical, community-aligned design interventions that can serve as boundary objects to connect different groups in the endangered language landscape.

6.1 Quantitative Findings: Survey Results

This section presents the descriptive statistics derived from the online survey conducted with individuals engaged in endangered language contexts. The survey explored participants' demographics and linguistic backgrounds, roles within language communities, and experiences with digital tools for language learning and communication. The goal of this section is to provide a high-level overview of general trends that will later be contextualised and deepened through the qualitative analysis.

The findings are presented in two parts:

- Descriptive statistics, covering participant demographics, language affiliations, and community roles.
- Engagement with technology, exploring patterns of digital tool usage, perceived affordances and barriers, and community dynamics.

The quantitative results serve as a backdrop to the qualitative interview themes, highlighting common patterns and divergences across stakeholder groups. Refer to Appendix C for the full list of data visualisations.

6.1.1 Descriptive Statistics

The survey captured a diverse cross-section of individuals involved in endangered language revitalisation efforts, providing valuable insights into their demographic distribution, roles, and linguistic affiliations.

In terms of age distribution, the largest group of participants fell within the 25-34 years category, accounting for 40.7% of the total sample. Participants aged 18-24 represented the second most populous group at 25.9%, while those aged 35-44 accounted for 22.2%. Respondents aged 55 years and above made up 7.4%, and the smallest group, those under 18, accounted for 3.7%. This distribution indicates a relatively diverse range of age groups participating in the survey, although young adults are in the majority.

Survey participants were geographically diverse, though concentrated in North America and Europe. Approximately one-third of respondents (29.6%) reported residing in the United States, followed by Ireland and Canada at 14.8% and 11.1%, respectively. Hungary and Finland each contributed 7.4%, while Croatia, France, Germany, Norway, Poland, the Netherlands, Spain, and the United Kingdom each made up 3.7% of the sample. However, no respondents reported being from Asia, Africa, South America, or Oceania, which may limit the global generalisability of the findings.

Participants reported a variety of roles within endangered language contexts. Approximately one-third (36.4%) identified as active learners of an endangered language, making it the most populous group. Native speakers, language activists or educators, and heritage speakers each represented 15.9% of the sample, while linguists or researchers accounted for 13.6%. Fluent

second language speakers made up the smallest group, standing at 2.3%. These figures highlight the varied, yet complementary roles participants hold within endangered language communities.

Finally, when asked which endangered language they were most connected to, respondents cited a broad range of languages. Irish was the most frequently mentioned language at 20.8%, followed by Tlingit at 8.3%. The remaining languages, each representing 4.2% of responses, reflected a wide range of linguistic diversity. Languages mentioned included Chamorro, Mansi, Udmurt, Cornish, Nahuatl, among others. This illustrates a wide distribution of linguistic affiliations and highlights the survey's relevance across varied revitalisation contexts.

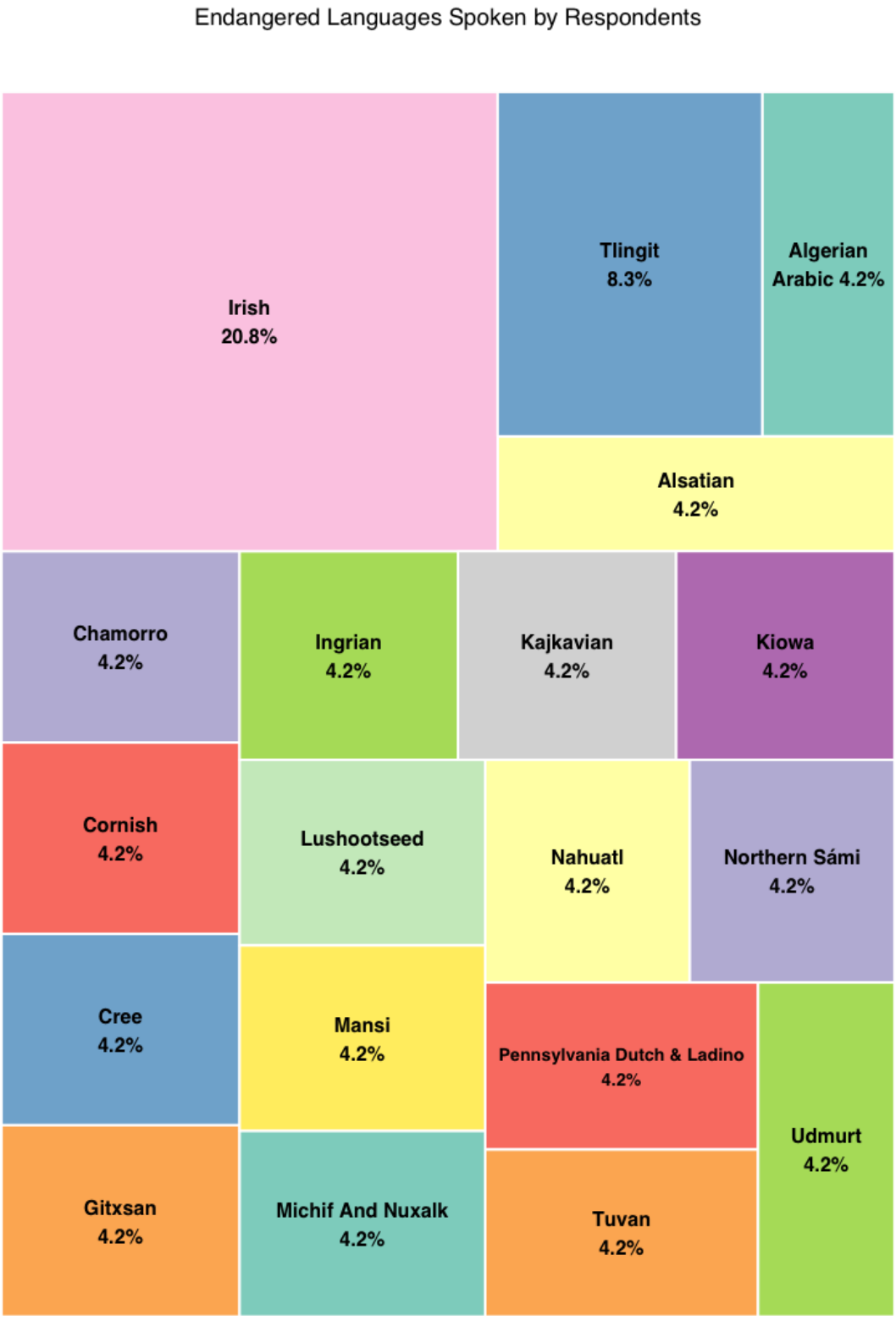

Figure 4. Endangered languages spoken by participants.

6.1.2 Engagement with Technology and Community Practices

Beyond demographic and linguistic profiles, the survey also explored patterns of engagement, perceptions of technology, and structural barriers within endangered language communities. The responses offer a nuanced view of both the diversity and fragmentation of contemporary practices in language revitalisation.

Modes and Frequency of Engagement

Approximately 66.7% of respondents reported daily engagement with their endangered language, while 18.5% engaged a few times per week. This high frequency suggests strong personal investment and routine integration into daily life for many participants. Modes of engagement were notably multimodal: everyday communication (20.2%), writing or blogging (18.3%), use of digital tools (16.3%), and participation in educational settings (15.4%) all featured prominently. These findings suggest that language use is embedded across social, cultural, and academic contexts.

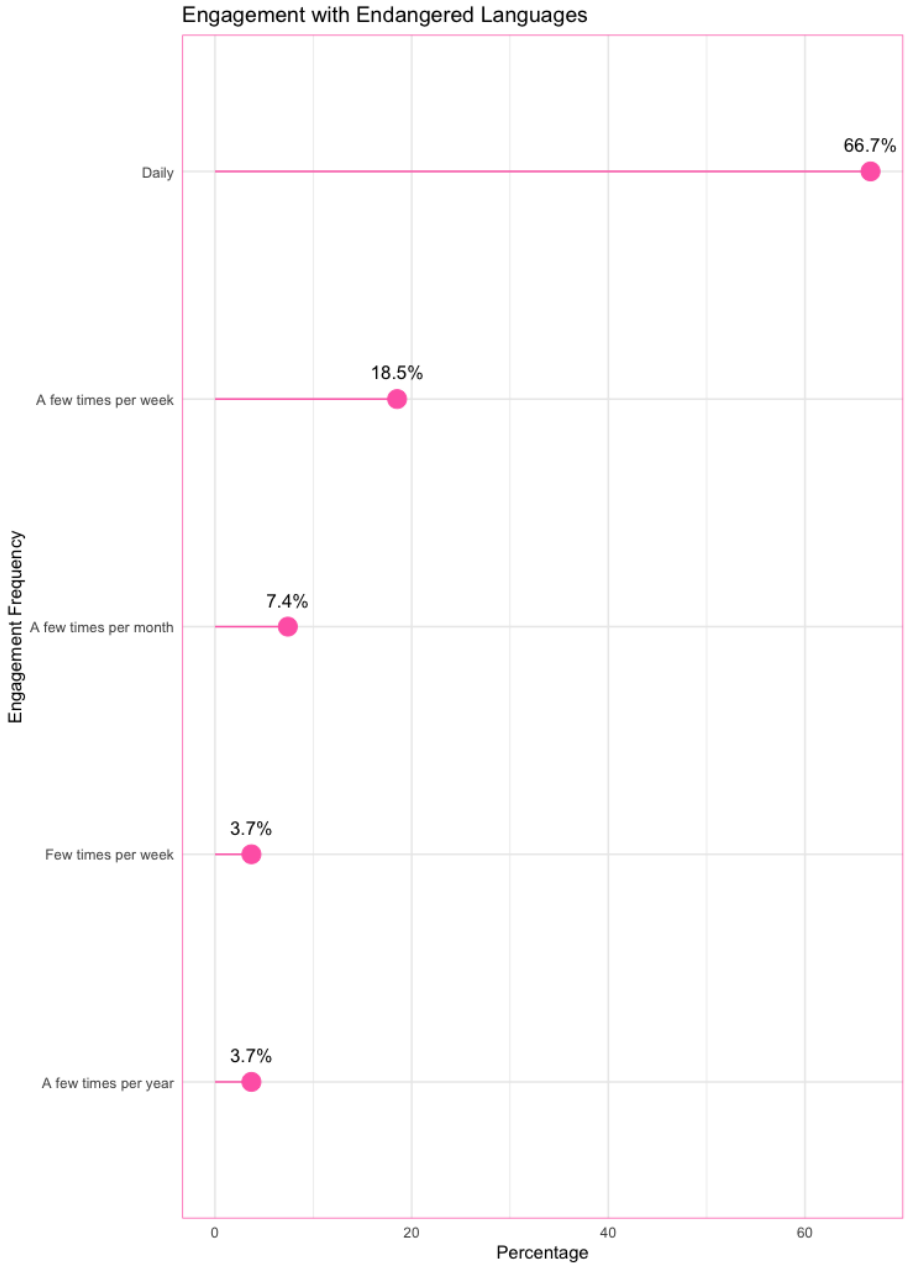

Figure 5. Frequency of engagement with endangered languages.

However, not all respondents reported such structured access to language use. 11.1% indicated they lacked any regular way to engage with their language community, highlighting a significant challenge: while some communities of practice are cohesive and well-developed, others remain fragmented, raising the question of whether they can even be referred to as communities of practice. This disparity underscores the importance of community-specific design solutions that can support the emergence or reinforcement of participatory structures. This, however, assumes everyone wants to learn an endangered language in a CoP setting, which may not always be the case. It also does not account for variations in individual learning styles.

Community Participation and AI Initiatives

When asked about participation in community-driven or AI-enhanced language initiatives, respondents expressed mixed views. One-third (33.3%) were willing to contribute, but 37% were uncertain, and 29.6% expressed unwillingness. These figures suggest a degree of ambivalence, perhaps reflective of past experiences, ethical concerns, or a lack of trust in external technologies. As will be explored in later chapters, trust, community governance, and transparency are key prerequisites for meaningful technological engagement.

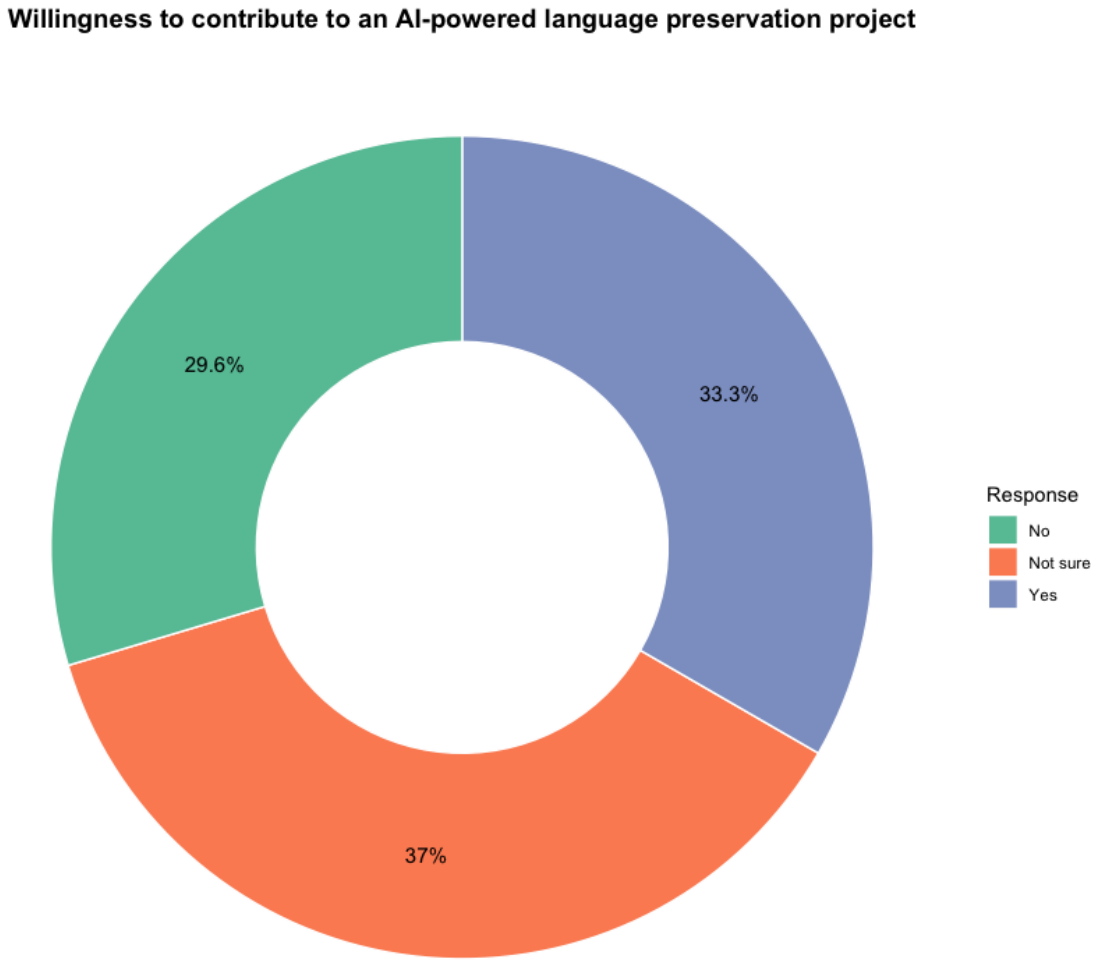

Figure 6. Participants' willingness to contribute to AI-powered language preservation projects.

Barriers to Engagement and Learning

Quantitative responses further identified key barriers to effective engagement. The most reported obstacle was a lack of resources (44%), followed by limited availability of fluent speakers or mentors (20.2%), and policy or institutional constraints (13.1%). Additional challenges included insufficient community support (10.7%) and linguistic complexity (9.5%).

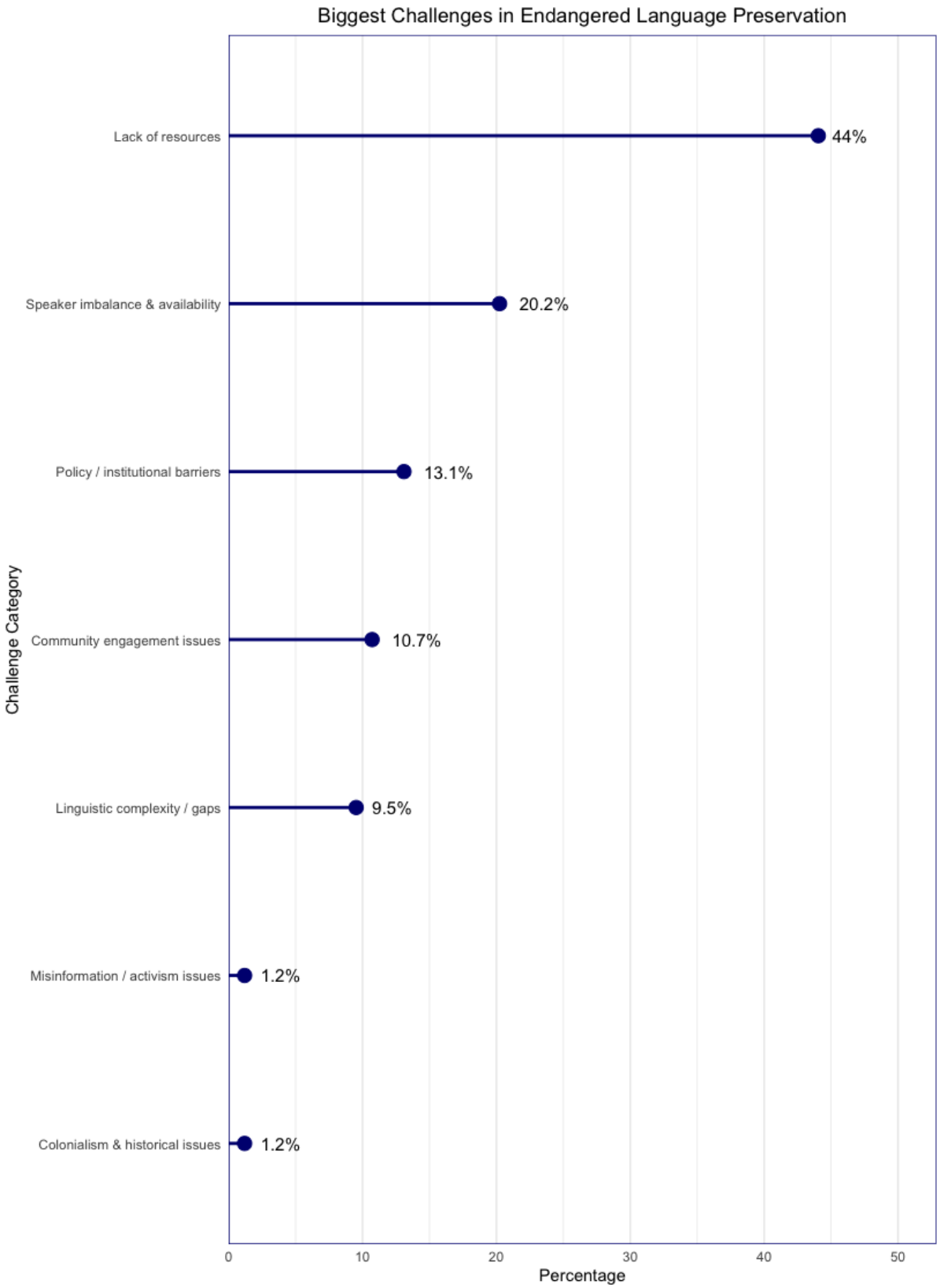

Figure 7. Biggest challenges in endangered language preservation perceived by participants.

These responses were reinforced by open-ended comments in the survey, which mirrored findings from the interview data detailed in the subsequent chapters. Several participants pointed to issues such as orthographic inconsistency, tensions between native speakers and

language learners, and concerns about activists lacking linguistic expertise, echoing the challenge of negotiating meaning, authenticity, and legitimacy in revitalisation contexts.

Availability of Learning Resources

When asked whether sufficient resources were available for their endangered language, only 14.3% of respondents agreed. In contrast, 42.9% reported that existing resources were insufficient, and another 42.9% said very few resources existed at all. These responses point not only to infrastructural gaps, but also to competing definitions of what constitutes a 'legitimate' or useful resource, a theme taken up in the qualitative analysis.

Use and Perception of Digital Tools

Among digital tools currently used for language learning and communication, the most common were online dictionaries and databases (44.4%), followed by social media groups and forums (17.1%), and language learning apps (15.6%). More advanced technologies, such as automated transcription or AI-enhanced platforms, were used by only a small minority. Even where such tools were present, they were often viewed with scepticism or concern. As one participant noted, *"AI-produced content in endangered languages is extremely low-quality and not trustworthy at all."*

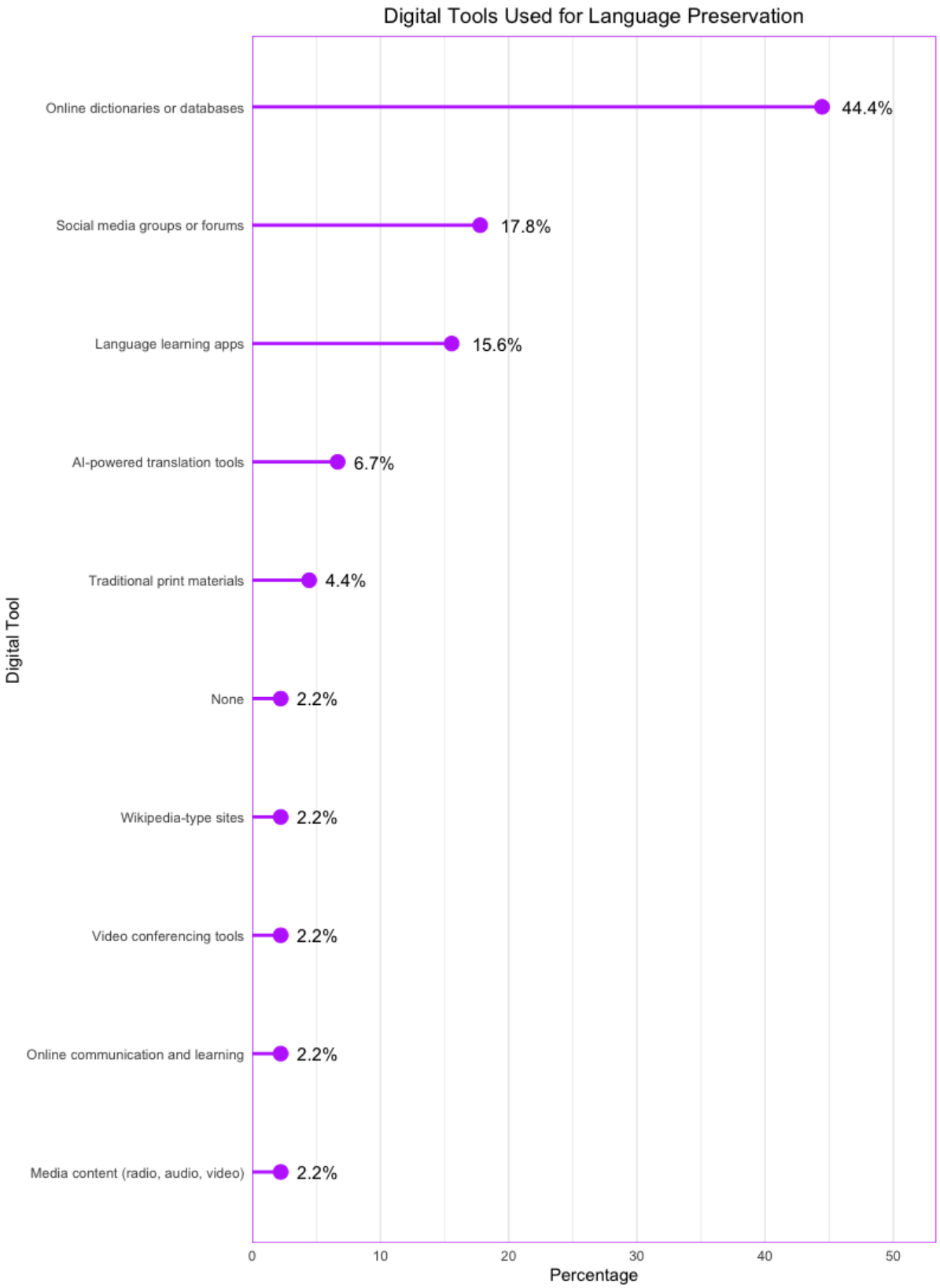

Figure 8. Types of digital tools used for language preservation.

Despite this mistrust, many participants expressed cautious optimism about the potential of well-designed tools. Open-ended responses highlighted interest in tools that could reduce cognitive and emotional labour for fluent speakers, particularly by automating routine tasks

such as transcription, grammar correction, or dictionary compilation. One participant envisioned tools that could *"automate simple tasks to free up fluent speakers for more nuanced work,"* while another emphasised the importance of *"accurate phonetics based on the speech of the strongest remaining speakers… in-depth knowledge of the grammar… then vocabulary."*

Desired Features in Language Technology

When asked about priorities for AI-powered or digital learning tools, respondents frequently mentioned:

- Accurate pronunciation
- Support for regional and dialectal variation
- Culturally appropriate grammar and phrasing
- Community control over data and tool governance
- Vetting by native speakers

While these responses indicate a real desire for technological support, they also stress the necessity of community involvement, linguistic accuracy, and cultural sensitivity in any future design efforts.

Trust in AI and Automated Tools

Trust emerged as a critical issue. When asked about the trustworthiness of AI-based tools in endangered language contexts, 48.1% of respondents rated them as "not at all trustworthy", with another 41.9% expressing reservations about their lack of nuance and contextual understanding. Several respondents raised fears of data misuse, cultural appropriation, or the erasure of indigenous ownership. One participant simply stated: "To not be used," in response to whether AI tools should be involved in language preservation, emphasising how deeply ethical concerns can run.

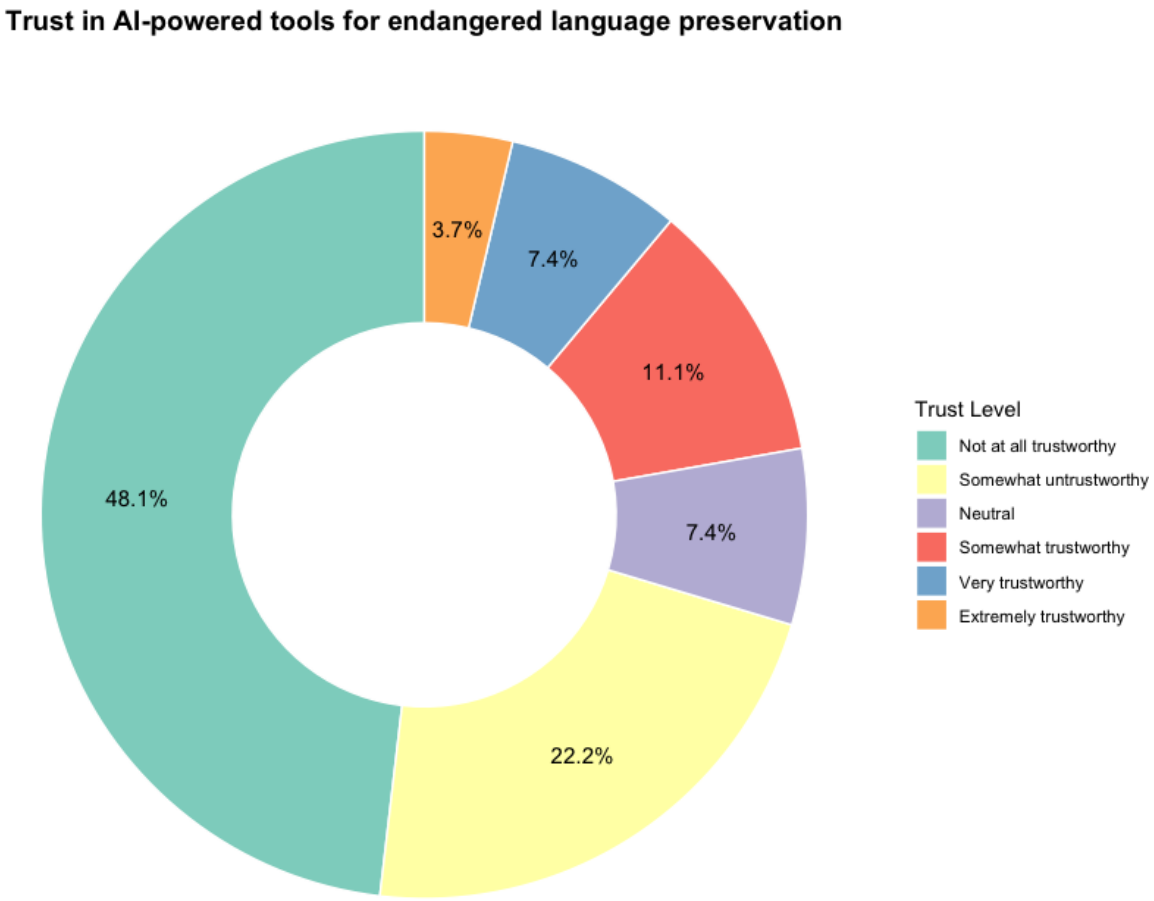

Figure 9. Perceived trustworthiness of AI-powered tools used for endangered language preservation.

This section has outlined key quantitative trends in engagement, barriers, and attitudes toward technology among endangered language communities. While daily engagement is high among many respondents, there are persistent challenges related to access, resources, and trust in technical solutions. These findings provide a backdrop for the following qualitative themes, where individual experiences and contextual complexities are examined in depth.

6.2 Thematic Analysis of Qualitative Data

This section presents the analysis of six semi-structured qualitative interviews conducted with individuals engaged in the preservation and revitalisation of endangered languages. The aim was to explore participants' lived experiences, motivations, and challenges in navigating linguistic, cultural, and technological landscapes. This portion of the analysis adopts an inductive, constructivist approach, grounded in thematic analysis (Braun & Clarke, 2014), allowing themes to emerge directly from the data rather than imposing predetermined categories.

The analysis process followed Braun and Clarke's six-phase framework, involving familiarisation with the data, initial coding, theme development, review, definition, and reporting. Coding was carried out manually, focusing on semantic-level patterns across

transcripts. Through iterative comparison and refinement, five overarching themes were developed: *Communities*, *Learning*, *Versions*, *Technology*, and *Identity*. These themes reflect how participants construct, maintain, and experience endangered language spaces, both online and offline.

This thematic structure is presented in Table 4 below, which outlines each main theme along with its corresponding sub-themes and illustrative quotations. These themes do not represent mutually exclusive categories but rather overlapping dimensions of practice and participation within endangered language contexts.

Themes	Sub-themes	Example Quotes
Communities	• Online community • Families • Groups don't interact • Etiquette • Native speakers stopped passing it on • Disconnect between communities • Death of local culture	"Some people were like me. They didn't hear it or speak it. Some people were somewhere in between, so we just formed an **online community** together and we have been gathering weekly online." – *Chamorro speaker* "[Native speakers] can deal with the grammar mistakes better than they can deal with pronunciation, because a lot of them won't have exposure to learners of Irish. The two groups don't really interact. What I think a lot of English speakers don't realise is how much exposure we have to foreign accents. That's a skill in and of itself. And like **the native speaking communities would not really interact with the learning communities** in Dublin and Belfast." – *Irish speaker (US)*
Identity	• Excluded by natives • Legitimacy concerns	"When you go outside of learning spaces [attitudes] can range. I know a lot of [learners] have encountered people, just like, doing their best to stop us. Ripping us down, even getting to identity, things like **saying that we aren't real Chamorros**, or making fun of our accents" – *Chamorro speaker* "The myths… So they've been retold so many times. It's kind of just like, it's not something that, I mean, it's something that we, consider, you know, Irish, but it's at the same time I wouldn't say we have a lot of ownership over it, you know. Because it's already out there. I personally don't have that much of an issue with [training AI models with this data]. **But then again, I'm not, like, a native speaker anyway, so I don't really feel like I can claim ownership**. – *Irish speaker (heritage)*
Learning	• Pursuing fluency • Not normalised • Old-fashioned • Tailored materials • Hard for beginners • Ways of practicing	"Another way to keep up on my learning, to help me to make sure I'm like stretching and growing, right? 'Cause I never want to be content of like, "Oh yeah, I've reached it." My partner calls it: **Forever pursuing the horizon of fluency.** Because we are second language learners. We know this, we know that we will always be growing and learning." – *Chamorro speaker* Part of that is just kind of school thing where you're forced to learn [Irish] in school and, you know, people don't want to. It's also kind of like there's a lot of criticism of how you learn it and it's kind of boring… **It's kind of old-fashioned**. A lot of it is focused on poetry and prose. It's not, like, communication-focused like other languages would be." – *Irish heritage speaker*

Versions	• No centralised version • Different spellings • Fight over what words to use	"Three million [Nahuatl] speakers are actually covering like a vast quantity of different dialects, multiple of which are so distinct from each other that they're not mutually intelligible…. This language was recorded in, well, in Western script in like 5 different ways. Initially, during the Spanish invasion and since then has had multiple revisions. For every word, there's like **12 different spellings**, minimum, for each of these different versions. So, **there is no centralised version**… This goes for teaching them with AI as well, like your training data is going to be finding 12 different spellings of every word. So, which version is it going to pick up?" – *Nahuatl learner* "But even they fight over what word to use, for example, I don't know, government… So they would rather use the Russian word because most of their speakers are Mansi-Russian bilinguals… I actually got in contact with one of the editors and she told the story about these problems that they face daily and how they, not fight over, but bicker over whatever is better and at the end they don't choose any of them, just go with Russian. That's easier. Why "bother" getting new vocabulary if the language is dying out?" – *Mansi learner*
AI and Technology	• Help with pronunciation • Nothing in the training data • Learners dominate • Companies mining data • Generating incorrect information • Be involved • Ownership doubts	"I think there's the most work that could be done with AI is on accent trainers, on pronunciation trainers, because that is where most people are failing. Like grammar, they make some mistakes… but in general it's the accent… I really do think AI could help the most is like **pronunciation** tools focused on the Gaeltacht speech. So you get people from Galway, Donegal. It's like how do we turn this? How do we make the learners sound like Bill?" – *Irish speaker (US)* "Our native speakers, they're elders, so they don't have a lot of familiarity with these technologies. Other than "Oh, this might be a good idea", they may not be aware of, like other considerations. **Should we be concerned about outside companies mining our data**, making an algorithm and then trying to sell their translation services back to us?" – *Chamorro speaker* "I know this is kind of veering, but it's like making sure that when we do have technology, like it has an appropriate place amongst everything else, right? **It shouldn't replace your grandmother**. Don't treat [tech] as if we have no native speakers left." – *Chamorro speaker*

Table 4. Overview of themes and sub-themes identified in thematic analysis.

Although the themes emerged inductively from participant narratives, they are subsequently interpreted through the lens of Communities of Practice theory (1998). This framework offers conceptual tools such as mutual engagement, joint enterprise, shared repertoire, trajectories of participation, and modes of belonging, which enable a deeper understanding of how participants negotiate access, legitimacy, and identity within their language communities.

By using Wenger's theory not as a prescriptive model but as an interpretive lens, the analysis highlights how different forms of participation, expertise, and cultural knowledge are situated within specific social and technological environments. The concepts of boundary objects (Star & Griesemer, 1989) and affinity spaces (Gee & Hayes, 2012) are also drawn upon to examine

how participants traverse fragmented or overlapping linguistic communities, particularly where traditional structures of learning and authority are absent.

The five themes are analysed in the sections that follow, each beginning with empirical findings and supported by direct quotations. These thematic chapters build toward a design solution that recognises the complexity of endangered language ecologies while identifying shared values and challenges across different speaker communities.

6.3 Communities

This section explores how communities form and function within the context of endangered language engagement. Drawing on Wenger's (1998) theory, it focuses on the concept of mutual engagement – the shared social practices, commitments, and routines that bind members together in a joint enterprise. The subchapter begins by identifying the shared passions and goals that fuel these communities, and later on examines how mutual engagement is facilitated, negotiated, or, in some cases, disrupted. While some groups reflect Wenger's more conventional model of CoPs, others align more closely with the affinity spaces described by Gee & Hayes (2012), particularly in digital and informal learning settings.

Across the dataset, participants described multiple types of language-focused communities, some grounded in heritage identity, others in online peer learning, and still others in intergenerational family contexts. Despite differences in structure, these groups shared a core interest in revitalising, reclaiming, or re-engaging with endangered languages, and were characterised by shared routines, learning norms, and emotional investments. The analysis that follows considers both how mutual engagement is expressed in these groups and what barriers exist to sustaining it over time.

Participants spoke extensively about the importance of informal, peer-led online communities in sustaining engagement with endangered languages. These communities often form organically, without institutional oversight, and were characterised by shared interests, flexible participation, and mutual support, qualities closely aligned with affinity spaces (Gee & Hayes, 2012). While these digital spaces may lack the long-term stability or institutional anchoring of conventional communities of practice, they serve an important function in fostering motivation, connection, and identity among learners.

"Some people were like me. They didn't hear it or speak it. Some people were somewhere in between, so we just formed an online community together, and we have been gathering weekly online." – (see Appendix E, Chamorro heritage speaker transcript)

"And yeah, the kind of community I got [in the Irish learner group] through like well, through, like just a general Irish group [in Barcelona]. Not, Irish speaking. Just Irish people group and then from that there was another guy who said there was an Irish speaking, like, WhatsApp group. So I joined that. And then they organized, you know? Just like. Meet up tonight." – (see Appendix F, Irish heritage speaker transcript)

Such digitally mediated groups mirror Wenger's notion of mutual engagement, albeit in a dispersed and often decentralised format. Participants described using platforms such as

WhatsApp and Discord to organise practice sessions, share resources, and support one another's learning. These spaces operate more as affinity spaces, interest-driven and non-hierarchical, than conventional CoPs, suggesting that traditional models of community must be adapted to reflect the realities of language revitalisation in the digital age.

Participants also highlighted the diminishing role of families as spaces of language transmission and community participation.

"It's really hard for us to use it with the families we have. We have maybe two family members like one, one family member that we interact with semi regularly, who is actually comfortable enough to just, like, speak. They either are too uncomfortable, because they just automatically switch to English. More often than not, they just actually cannot converse right. They can't say more than a couple of sentences before they just switch to English."– (see Appendix E, Chamorro heritage speaker transcript)

In Wenger's (1998) terms, many families no longer function as communities of practice due to the erosion of shared repertoires and opportunities for mutual engagement. Although language may still symbolically represent identity within family structures, it is often not actively used in daily interaction.

This breakdown in familial CoPs undermines intergenerational continuity and places the burden of revitalisation on alternative social structures, particularly peer-based and digital communities. The absence of consistent language use at home further reinforces the need for supplementary spaces of engagement, spaces where learners can build competence, practice fluency, and negotiate identity.

Another recurring theme in the data was the fragmented nature of endangered communities. Participants frequently described a disconnect between native speakers and learners, with limited interaction, differing expectations, and, in some cases, outright tensions. While both groups may be engaged in the joint enterprise of language revitalisation, they often do so from divergent perspectives, resulting in fragmented or parallel practices: *"The two groups don't really interact... native speaking communities would not really interact with the learning communities in Dublin and Belfast."* – (see Appendix G, Irish fluent speaker transcript)

Wenger (1998) notes that CoP are bounded yet permeable, and successful cross-community interaction often relies on brokers – individuals or tools that can translate meaning across social worlds. In the absence of such brokers or bridging mechanisms, misunderstanding and hierarchical dynamics emerge. Native speakers may question the legitimacy of learners, while learners may feel excluded or undervalued.

Some communities have responded by teaching "etiquette" explicitly to learners, aiming to reduce friction and improve mutual understanding:

"What we also try to do with the learners in our group we try to teach them, for lack of a better word, etiquette around how to approach the community of speakers. How to interact. How to be respectful. Just things like, OK, don't go to your grandma and say "Grandma, how do I put this in infinitive?" – (see Appendix E, Chamorro heritage speaker transcript)

This practice can be interpreted as an attempt to scaffold legitimate peripheral participation (Lave & Wenger, 1991), where newcomers are gradually socialised into community norms through low-risk, observational, or guided activities. In this case, learners are not expected to participate fluently from the outset but are instead introduced to unspoken cultural expectations and social etiquette that govern interaction within the speech community. Teaching learners how to "approach" native speakers, not only linguistically, but socially, serves to mediate potential friction and builds the kind of mutual trust needed for deeper engagement. By explicitly framing these interactions, communities help learners position themselves not as outsiders seeking knowledge, but as legitimate members-in-the-making, thereby replacing the epistemic tension that can arise between fluent speakers and novices. Over time, this can foster stronger alignment around shared goals and reinforce the learner's sense of belonging, even in communities where access to fluent speakers is limited or highly mediated.

Finally, participants lamented the erosion of traditional physical and social infrastructure that once supported language use and community gathering. The loss of communal spaces, whether due to policy shifts, urbanisation, or generational change, has reduced the opportunities for everyday language use in informal, intergenerational settings.

"[Ireland] introduced a zero-tolerance drunk driving law… that caused the death of a lot of local culture. I'm not gonna say it was a bad thing, but it means that a lot of the local pubs, the old men couldn't get in anymore. You have to go further away and then [people] don't get out. There's nowhere to meet, nowhere to bring the kids to do stuff like that. It's totally starting to come back, but I know, at least [on the Airen Islands] it's very difficult to get the people out now." – (see Appendix G, Irish fluent speaker transcript)

These observations align with Wenger's argument that mutual engagement is sustained not only through repeated, routine contact. The disappearance of pubs, community centres, and neighbourhood traditions represents a breakdown in the physical ecology that once supported vibrant communities of practice. Community members, especially the younger generation, have fewer options to engage with one another in their native languages, which, combined with media consumption habits reported on by one Irish speaker, *"Pretty much everything outside of school would be through English. Even a lot of their interactions with each other, just 'cause that's the language of social media,"* (see Appendix G, Irish fluent speaker transcript) actively contributes to the aforementioned breakdown of the physical ecology, but to the erosion of the language as a whole.

6.3.1 Interpretation and Design Implications

Taken together, the data suggest that while not all groups described by participants would qualify as full communities of practice under Wenger's strict definition, many shared enough features, including mutual engagement, shared repertoires, and joint enterprise, to be understood as *emergent* or *partial* CoPs. Instead, learners often organise in ways that reflect affinity spaces, peer-led, flexible, and mediated by shared passion rather than formal membership. Native speakers, meanwhile, may remain embedded in geographically bounded or more traditional community structures.

This fragmentation indicates a need for boundary objects, shared artefacts or platforms that can facilitate interaction across communities with different levels of expertise, legitimacy, and practice. A well-designed digital infrastructure could serve this purpose by:

- Enabling respectful and structured interaction between native speakers and learners.
- Supporting etiquette and norm-setting in peer-led spaces.
- Encouraging co-created repertoires through collaborative tools.

In doing so, such a platform would not impose a singular model of community but instead support plural forms of engagement and belonging, consistent with both Wenger's theory of social learning and Gee et al.'s concept of distributed, interest-based affinity spaces.

6.4 Identity

The second theme centres around questions of identity, legitimacy, and belonging, particularly as they emerge in the fraught space between native speakers and language learners. Participants across interviews described how their engagement with endangered languages was not only shaped by logistical or pedagogical barriers, but by deeper concerns about how they are perceived, by others and by themselves, as legitimate members of a linguistic community. Drawing from Wenger's concepts of community membership, modes of belonging, and non-participation, this section explores how identity becomes a site of tension and negotiation in the revitalisation landscape. Affinity spaces further help contextualise how learners, who may not be fully recognised within traditional CoPs, nonetheless carve out new forms of identification and participation.

A strong undercurrent in participant narratives was the feeling of being excluded or delegitimised by native-speaking communities. This dynamic was most clearly articulated by a Chamorro heritage speaker:

"When you go outside of learning spaces [attitudes towards learners] can range. I know a lot of [learners] have encountered people, just like, doing their best to stop us. Ripping us down, even getting to identity, things like saying that we aren't real Chamorros, or making fun of our accents." – (see Appendix E, Chamorro heritage speaker transcript)

This statement highlights a recurring form of non-participation, not simply an absence of engagement, but an exclusionary process whereby learners are actively denied full membership within the community (Wenger, 1998, p. 165) Accent, often viewed as a marker of inauthenticity, becomes a boundary marker that reinforces insider-outsider distinctions, despite the learner's best effort to belong.

Similarly, an Irish heritage speaker reflected on their uncertain standing within the community, particularly in relation to data ownership and cultural authority:

"I, personally, don't have that much of an issue with [training AI models with this data]. But then again, I'm not, like, a native speaker anyway, so I don't really feel like I can claim ownership." – (see Appendix F, Irish heritage speaker transcript)

This sentiment reflects a form of disidentification, an internalised sense of partial belonging, wherein individuals hesitate to assert rights or roles associated with full membership. These ambiguities illustrate Wenger's notion of modes of belonging, particularly the tensions between engagement (active participation), imagination (projecting oneself into a future identity), and alignment (adopting shared goals and values). When legitimacy is questioned by others or by oneself, the path to fuller engagement is interrupted or foreclosed.

While CoPs offer a useful model for understanding how identity is shaped through shared practice, they may also be too rigid to accommodate learners whose access to the community is partial, peripheral, or digitally mediated. In contrast, Gee & Hayes' concept of affinity spaces provides an alternative lens through which to view learner engagement. Affinity spaces are interest-driven, low-barrier environments where participants cluster over shared passion and goals, rather than a unified identity (Gee & Hayes, 2012). Spaces like these value contribution over credentials, and legitimacy is often earned through participation rather than conferred through status or heritage.

Many of the online forums, peer groups, and informal WhatsApp circles described by participants more closely resemble affinity spaces than formal CoPs. These digital communities offered learners a sense of inclusion and progress even in the absence of native-speaker validation. As one participant shared, *"we're here for the language, we're here to support each other, but we also have a sense of we know how to laugh at ourselves good-naturedly."* (see Appendix E, Chamorro heritage speaker transcript). Here, participation is framed not around credentials or fluency, but shared passion and consistent contribution. These spaces value contribution over status, allowing learners to develop confidence, practice skills, and co-construct meaning in a low-stakes, supportive environment. While they may not confer full legitimacy in the eyes of native speakers, they offer a valuable scaffolded mode of belonging, a way to move from peripheral interest to deeper identification on learners' own terms.

6.4.1 Interpretation and Design Implications

The theme of Identity reveals that revitalisation is not only about linguistic knowledge but also recognition, both internal and external. Learners often navigate a liminal space in which their legitimacy is questioned, their contributions undervalued, and their roles uncertain. Wenger's (1998) framework helps articulate how these dynamics are structurally produced within communities, while affinity spaces (Gee & Hayes, 2012) illustrate how new forms of belonging can still emerge despite these barriers.

For design, these findings suggest the importance of creating boundary objects that do not simply reinforce existing hierarchies but allow for flexible, respectful entry points into community engagement. By supporting multiple forms of participation and legitimacy, from peripheral involvement to deep cultural stewardship, such tools can help learners move toward fuller membership while ensuring the native speakers retain agency over their linguistic heritage.

6.5 Learning

The theme of *Learning* emerged as a central concern across all interviews, with participants consistently reflecting on how they engage with endangered languages, the challenges they face in doing so, and the socio-cultural constraints that shape these practices. This section explores learning as a socially situated process, rather than a purely cognitive or individualistic one. Learning here is framed as participation in socially meaningful practices, shaped by one's position within a community, access to resources, and sense of belonging. Additional insights drawn from affinity spaces help illuminate how informal digital environments provide alternative learning structures when full community membership is difficult or inaccessible.

Participants described learning endangered languages as an ongoing, open-ended process, often pursued in the absence of clear paths to fluency or access to formal support structures. One Chamorro heritage speaker captured this sentiment by stating:

"My partner calls it: Forever pursuing the horizon of fluency. Because we are second language learners. We know this, we know that we will always be growing and learning." – (see Appendix E, Chamorro heritage speaker transcript)

This reflects Wenger's concept of trajectories, which describes the evolving participation of individuals in a community (Wenger, 1998). Learners in this context are aware of their peripheral position and work persistently toward fuller participation, even when access to native speakers or immersion environments is limited. Their learning, then, is shaped by imagination, a mode of belonging where individuals envision themselves as part of a community, they are not yet fully integrated into (Wenger, 1998).

A recurring theme in participants' reflections was a sense of social disconnection or marginalisation, especially within their immediate environments. Learners often described feeling like outliers in spaces where language use is uncommon or unrecognised:

"It's still kind of weird to, like, wanna try to use the language and learn it in our culture. It's not normalised at all. So, like, we're all the odd ones out, like usually those of us who are learning together. We're the only ones who either can speak it at this point, or want to use it." – (see Appendix E, Chamorro heritage speaker transcript)

This indicates that learning endangered languages often lacks normative support or social validation, especially within the learner's immediate environment. Without normalisation or local reinforcement, learners must rely on external or self-created networks, often in the form of online communities.

Participants frequently criticised the dominant modes of endangered language education as being outdated, inaccessible, or misaligned with communicative goals. One Irish heritage speaker remarked, *"It's kind of old-fashioned. A lot of it is focused on poetry and prose. It's not, like, communication-focused like other languages would be."* (see Appendix F, Irish heritage speaker transcript). Such statements highlight a disconnect between reification, the production and use of artefacts like textbooks or institutional curricula, and participation, the lived experience of using a language (Wenger, 1998). When reified tools fail to support meaningful participation, they risk becoming obstacles rather than aids.

This mismatch was especially noticeable for learners with varying levels of proficiency or different linguistic backgrounds. As one Irish speaker noted:

> *"If half the class is beginners from a foreign country, another half are Irish people wanting to come back and get into the language again, they're gonna be completely different. You need a set of 'Here's what we do at this level.'"* – (see Appendix G, Irish fluent speaker transcript)

These challenges illustrate the need for tailored materials that support differentiated learner trajectories, accommodating heritage speakers, diaspora learners, and complete beginners alike. Without such responsiveness, many learners experience friction in their attempts to engage meaningfully.

Despite the limitations of formal pedagogical settings, many participants described active, self-directed efforts to incorporate language learning into their lives. These practices reflect a commitment to engagement, one of Wenger's (1998) core modes of belonging. A Mansi learner shared, *"I mostly try every day. I'm doing something around the house and just trying to speak."* (see Appendix D, Mansi learner transcript). In this context, informal, low-threshold platforms such as WhatsApp groups or language-focused Discord servers become key spaces for learning. This is further emphasised by a Chamorro heritage speaker, stating, *"So we interact. We text each other in Chamorro on WhatsApp."* (see Appendix E, Chamorro heritage speaker transcript). These digital spaces do not necessarily constitute CoPs in the traditional sense, but they function as affinity spaces – environments where individuals come together around a shared interest regardless of their background or level of expertise. These spaces allow learners to exchange resources, practice vocabulary, and form peer relationships, often filling the gaps left by institutional or familial structures.

Such affinity spaces support what Wenger (1998) calls multimembership, where individuals participate in multiple communities or spaces simultaneously. One participant noted, *"I actually got into the language community in my university",* and *"I'm part of Uralics of Russia[1], yes."* (see Appendix D, Mansi learner transcript). For endangered language learners, this might mean engaging in heritage networks, academic forums, and digital affinity groups, each of which supports a different aspect of their learning journey.

6.5.1 Interpretation and Design Implications

The findings presented here suggest that learning endangered languages is deeply tied to issues of access, representation, and identity. Many learners operate on the periphery of formal CoPs and rely on informal digital spaces that mirror the structure of affinity spaces. Their learning is motivated not by institutional mandates but by cultural connection, personal identity, and the pursuit of fluency on their own terms.

These dynamics carry important implications for the design of digital tools intended to support language revitalisation. Tools must support non-linear learning paths, accommodate diverse learner identities, and foster community-building through features like peer matching,

[1] Uralics of Russia is an online community hosted on Discord. It focuses mainly on Uralic languages found in Russia, and to some extent the Baltics and Finland.

customisable content streams, and asynchronous participation. Rather than replicating the limitations of institutional curricula, such platforms should scaffold community participation through flexible, learner-centred architectures that respond to real-world challenges of access, relevance, and belonging.

6.6 Versions

A key tension that emerged across participant interviews relates to the absence of standardisation and the resulting complexity in navigating dialectal and orthographic variation within endangered languages. The theme *Versions* illustrates how language learners and speakers negotiate the boundaries of "correctness," authenticity, and usability in real time, deciding not only what form of the language to use, but whose usage is validated, taught, or encoded into tools. In Wenger's (1998) framework, these negotiations reflect the dynamic interplay between participation (how people speak, write, and relate to the language in practice) and reification (how meanings become formalised through dictionaries, corpora, or educational materials). In the context of language shift or fragmentation, these negotiations are not merely linguistic but deeply cultural and intergenerational, shaping questions of legitimacy, authority, and representation.

Participants described these tensions as highly practical: Which spelling should I use? Which dialect should I teach? Who decides what counts as correct? One Nahuatl learner articulated this complexity:

> *"Three million [Nahuatl] speakers are actually covering like a vast quantity of different dialects, multiple of which are so distinct from each other that they're not mutually intelligible… For every word, there's like 12 different spellings, minimum, for each of these different versions. So, there is no centralised version… This goes for teaching them with AI as well, like your training data is going to be finding 12 different spellings of every word. So, which version is it going to pick up?"* – (see Appendix H, Nahuatl learner transcript)

This quote underscores how the fragmentation of orthography, dialect, and linguistic register makes both community learning and technological design incredibly complex. Learners are not just trying to "learn a language" in the abstract, but are often forced to choose, or have imposed upon them, particular versions that may be perceived as more "legitimate," modern, rural, or institutionally acceptable.

In Wenger's terms, this reflects a breakdown in reification, when a community lacks stable artefacts such as dictionaries, curricula, or corpora that reflect collective consensus, reified tools cannot support learning, participation, or identity-building. Instead, variation creates friction across generations and roles, leaving both learners and fluent speakers uncertain about what constitutes "correct" or appropriate usage.

Beyond orthographic variation, participants also described lexical negotiation, particularly when new or politically charged concepts had to be expressed. A Mansi learner shared:

> *"But even they fight over what word to use, for example, I don't know, government…… So they would rather use the Russian word, because most of their speakers are Mansi-Russian*

bilinguals… I actually got in contact with one of the editors, and she told the story about these problems that they face daily and how they… bicker over whatever is better and at the end they don't choose any of them, just go with Russian. That's easier. Why 'bother' getting new vocabulary if the language is dying out?" – (see Appendix D, Mansi learner transcript)

This kind of negotiation isn't purely linguistic; it also reflects competing ideologies of language survival. Some speakers may resist coining new terms, viewing the language as already endangered and not worth expanding; others may advocate for adaptation and modernisation. In this way, word choice becomes a proxy for deeper disagreement about the value of revitalisation, the future of the language, and the legitimacy of different speaker roles.

6.6.1 Interpretation and Design Implications

The fragmented nature of endangered language versions poses a major challenge to AI-supported tools and digital infrastructure. Language technologies typically rely on consistent, structured input to produce usable outputs. But in the context of endangered languages, inconsistency is not an error, but a reflection of lived linguistic reality.

From a design perspective, this calls for tools that accommodate, rather than eliminate, variations. Interfaces must support multiple orthographies, dialect tagging, and user-generated metadata that capture regional or social variation. Rather than enforcing standardisation, tools should scaffold negotiation, allowing communities to document competing forms, explain lexical choices, and build collective consensus where possible.

At the same time, such tools must avoid collapsing complexity into misleading simplicity. As Wenger (1998) argues, meaning is not a static property of words or systems, but a dynamic product of social practice. Technologies that fail to account for this may inadvertently privilege some users over others, particularly those whose dialects, spellings, or terms are already overrepresented in digital spaces.

6.7 AI and Technology

The final theme explores the affordances, limitations, and contested roles of technology, particularly AI, in the context of endangered language revitalisation. Participants expressed both cautious optimism and significant concern regarding current and emerging tools, especially those driven by machine learning and large language models. This section situates technology not as a neutral medium, but as an artefact embedded in social practice, one that can shape, disrupt, or support community learning depending on how it is developed and adopted. Wenger's (1998) concepts of learning architectures, imagination, and alignment are particularly relevant in this context.

Several participants recognised specific areas where technology, particularly AI, could offer meaningful assistance in endangered language learning and preservation. A frequently cited example was pronunciation support, which is often underrepresented in learning tools but crucial to community acceptance. As one Irish speaker noted:

"I really do think AI could help the most is like pronunciation tools focused on the Gaeltacht speech. So, you get people from Galway, Donegal. It's like, how do we turn this? How do we make the learners sound like [a native speaker]?" – (see Appendix G, Irish fluent speaker transcript)

This reflects a desire for learning tools that respect regional linguistic authenticity and support learners' trajectories toward fuller participation. In Wenger's terms, such tools could be seen as part of a learning architecture, a configuration of tools and relationships that facilitate legitimate peripheral participation in a CoP (Wenger, 1998, p. 229). However, for these architectures to be effective, they must align with the needs and expectations of both learners and native speakers. These needs were articulated in various ways by participants: some wanted pronunciation models grounded in specific dialects *("How do we make the learners sound like Bill from Donegal?")* (see Appendix G, Irish fluent speaker transcript), while others prioritised connecting language learners and native speakers *("My idea is really to help learners and speakers connect better")* (see Appendix E, Chamorro heritage speaker transcript). These perspectives point to a broader desire for tools that are dialect-aware, emotionally supportive, and grounded in everyday use and not just accuracy-focused or corpus-based.

Despite the perceived promise of AI, participants expressed concern that current technologies often reflect the priorities of institutions or developers, rather than those of the communities they aim to serve. This disconnect was articulated by an Irish speaker:

"Forefronting the communities and what their needs are. And that is kind of my issue, like, I worked for the group that does a lot of digitisation and digital efforts for Irish, and they're like "Well, French has this, X has this, Y has this. We need this too!" But is that the most important thing that's facing Irish? Is a new corpus the most important thing, or would text-to-speech be much more important? Who thinks that having these [tools] is the most important step forward to Irish? Is that what the Irish communities want? Or is that what the academics want?" – (see Appendix G, Irish fluent speaker transcript)

Here, alignment, the negotiation between local practice and external structures, is lacking (Wenger, 1998). Tools may be reified into polished products without adequate participation from those who will use or be affected by them. This results in digital solutions that fail to support meaningful engagement, even as they are marketed as revitalisation initiatives.

One example of this misalignment is the issue of training data. AI systems require large amounts of structured input, but endangered languages are typically underrepresented or misrepresented in such data. One Nahuatl learner said, *"With Nahuatl I gave [ChatGPT] a very simple list of vocabulary, but within two lines ChatGPT was telling me, with utmost confidence, incorrect information."* (see Appendix H, Nahuatl learner transcript). This experience demonstrates a breakdown in imagination, the CoP mode of belonging that enables individuals to construct identities and practices through models and projections (Wenger, 1998, p. 176). When AI-generated content confidently delivers false information, it can mislead learners who lack access to fluent speakers or corrective feedback. As a result, learners may internalise incorrect patterns that go unchallenged, entrenching linguistic error and potentially reinforcing inauthentic norms within digital ecosystems. Over time, these patterns may become difficult to dislodge, particularly if they are encoded into future training data. In this way, the technology

not only disrupts learners' ability to imagine themselves as competent users of the language but also risks solidifying misrepresentations that can further distance revitalisation efforts from community norms.

Participants also voiced significant ethical and epistemological concerns about how data is sourced, who controls it, and how it is repurposed. A Chamorro speaker reflected, *"Should we be concerned about outside companies mining our data, making an algorithm and then trying to sell their translation services back to us?"* (see Appendix E, Chamorro heritage speaker transcript). Such concerns reflect not only a lack of transparency but a broader sense of dispossession where communities are not only excluded from tool development but also risk having their cultural resources commodified without consent. This undermines engagement, another mode of belonging in Wenger's model, by severing the relationship between cultural practice and community ownership.

These reflections speak to a broader fragmentation in the shared repertoire that underpins endangered language engagement. Drawing on Wenger's (1998) concept of shared repertoire, the communal resources, tools, stories, and language norms that support participation, it becomes clear that current infrastructures are both scattered and unevenly governed. Table 5 below maps key tools and practices to their primary users, highlighting significant gaps in accessibility, cultural alignment, and governance. It functions as both a diagnostic tool and the design rationale for the platform envisioned in this thesis.

Tool/Practice	Primary Users	Function	Gaps Identified
Oral traditions (songs, stories)	Native speakers	Transmission of values, identity, immersion	Often inaccessible to learners, not digitised
Digital dictionaries/wordlists	Learners	Vocabulary acquisition, reference	Lack of dialectal variation. Low contextual examples
Online groups	Learners (peer-led groups)	Mutual support, social learning	Unarchived, lacks permanence for community memory
Grammar or etiquette instruction	Learners	Guidance o respectful use, interaction norms	Not standardised, depends on group knowledge
Cultural idioms/expressions	Native speakers	Rooted knowledge, worldview expression	Rarely taught, not well-integrated into formal learning materials
AI-generated translations/tools	Learners	Supplementary practice tools	Mistrust, inaccuracy, lack of linguistic and cultural nuance

Community-produced resources	Both	Trusted materials with contextual relevance	Limited distribution, inconsistent quality or documentation
Academic/institutional tools	Learners	Linguistic corpora, grammar analyses, Bible translations	Often inaccessible, overly formal, or disconnected from user needs

Table 5. Elements of the shared repertoire and identified gaps.

As the table illustrates, reification has occurred unevenly across contexts, with some tools like dictionaries being highly developed but lacking nuance, while oral traditions remain rich but inaccessible. From a design perspective, this fragmentation calls for boundary objects that do not homogenise or erase difference, but scaffold alignment between communities, platforms, and use cases. Such a system must acknowledge the distributed nature of knowledge, the asymmetries of access, and the diverse trajectories of learners and native speakers alike.

In several interviews, participants questioned whether learners, particularly those without deep cultural ties, should be allowed to contribute data or shape technological tools. One Irish speaker explained:

> *"The learners dominate and that's a huge issue. If you go to, say ChatGPT, or even the Microsoft AI voice models, they are based on learners. But to those learners, they also have the idea 'I'm Irish, therefore Irish is my native language.' But it's not their native language in the same sense."* – (see Appendix G, Irish fluent speaker transcript)

This reflects a deeper identity tension that digital tools may potentially favour some learners while alienating others. Technologies that treat all data points as equal can inadvertently elevate learner-produced content as normative, marginalising native speaker knowledge and reinforcing misalignment in the community.

Despite these concerns, participants were not universally dismissive of AI. Rather, they called for participatory design and ethical frameworks that reflect community priorities. As one linguist articulated, *"If we really want to see positive speech technologies come into fruition for some of these languages, then the people need to be involved from the ground up."* (see Appendix I, Linguist transcript) This reflects a call for technologies that support community imagination, not just individual learning goals, but collective visions for language futures. It also underscores Wenger's (1998) argument that learning systems must be embedded within communities' own structures of meaning, rather than imposed from above.

6.7.2 Interpretation and Design Implications

The theme Technology reveals a complex and often contradictory set of perspectives. On one hand, there is a clear potential for AI to support pronunciation, automate repetitive tasks, and expand access to learning. On the other hand, current implementations often reproduce existing power asymmetries, overlook cultural nuance, and exclude key community voices. Wenger's concepts of learning architecture, alignment, and imagination offer a useful lens through which

to evaluate not just whether technologies work, but whether they enable deeper participation, respect diverse identities, and support meaningful revitalisation trajectories.

Designers and researchers must therefore reframe technological development as a community-centred learning process, not as a solution delivered to communities, but as an infrastructure co-developed with them. Only then can AI tools become enablers of engagement, rather than barriers to it.

6.8 Conclusion

The preceding analysis has surfaced a number of recurring tensions that cut across participant roles, technological practices, and community values within endangered language contexts. While individuals and groups demonstrate high levels of engagement and commitment, their efforts often remain fragmented, misaligned, or unsupported by the current digital infrastructure. This concluding section synthesises the qualitative and quantitative findings to inform the design rationale introduced in Chapter 7. Drawing on the theoretical frameworks of Communities of Practice (Wenger, 1998), boundary objects (Star & Griesemer, 1989), and affinity spaces (Gee & Hayes, 2012) it articulates the sociotechnical conditions that shape language engagement today and frames the proposed design not as a singular solution, but as a flexible, co-governed infrastructure for bridging divides.

6.8.1 Thematic Cross-Synthesis

Three cross-cutting tensions emerged across the thematic analysis, reflecting the competing demands placed on speakers, learners, and the tools they employ:

- **Participation vs. legitimacy**: Learners seek pathways to fluency and cultural engagement, yet often encounter legitimacy barriers from native speakers, particularly around pronunciation, grammar, and perceived authenticity. This dynamic produces a gatekeeping effect that discourages informal practice and limits opportunities for mutual learning.
- **Cultural richness vs. accessibility**: Oral traditions and idiomatic speech, both central to many endangered languages, are rich, expressive, and culturally embedded. However, they are also difficult to digitise or replicate through AI-driven tools, which tend to prioritise textual and standardised data. As a result, digital resources often feel superficial or incomplete.
- **Engagement vs. fragmentation**: While digital platforms like WhatsApp and Discord have enabled routine interaction and community formation, these spaces are often siloed by geography, ideology, or linguistic role. Learners and native speakers rarely overlap in meaningful ways, limiting opportunities for intergenerational or intercultural collaboration.

These frictions underscore the need for design approaches that acknowledge epistemic difference and asymmetry, not by seeking consensus, but by enabling coordination through shared, negotiated infrastructures.

6.8.2 Theoretical Reflection

Viewed through the lens of Wenger's (1998) Communities of Practice framework, the data reflect partial and uneven enactments of mutual engagement, joint enterprise, and shared repertoire. Mutual engagement was evident in peer-led WhatsApp groups, Discord servers, and Zoom study sessions (see 6.3), which offered social support and casual practice opportunities. However, many of these communities lacked mechanisms for relational continuity or shared accountability, often depending on irregular schedules or peer initiatives. The joint enterprise of revitalisation was similarly fragmented, with participants prioritising different aims, such as fluency, cultural reconnection, linguistic documentation, or AI development (see 6.4 and 6.7), often in isolation from one another. Shared repertoire, while present, was often contested: learners relied on tools like dictionaries or AI models, which lacked dialectal nuance (6.6), while native speakers often drew from oral traditions and informal norms that were poorly represented in digital formats. These frictions illustrate the structural limits of participation across roles and generations.

Certain tools, particularly online dictionaries and Discord channels could function as boundary objects in the Star and Griesemer (1989) sense. They enabled interaction across user groups while preserving epistemic distance. Yet, these tools often lacked the co-governance or adaptability required to truly mediate between divergent community needs. Learners adapted them for vocabulary acquisition, while native speakers used them for preservation or resisted them entirely. The absence of negotiated governance structures meant that such tools facilitated parallel use rather than collaborative meaning-making.

In contrast, many digital environments resembled affinity spaces (Gee & Hayes, 2012), where informal, interest-driven participation enabled flexible entry but did not necessarily lead to mutual recognition or long-term commitment. These spaces were crucial for early engagement, particularly for isolated learners or diaspora participants, but did not always foster the deeper relationships or sustained interaction that typify robust CoPs. The coexistence of fragmented CoPs, loosely structured affinity spaces, and underdeveloped boundary objects suggests a need for hybrid infrastructures: designs that accommodate multiple user roles and epistemologies without enforcing standardisation. Rather than expecting consensus, these systems must scaffold interoperability and co-governance, enabling distinct communities to collaborate without collapsing their differences. This pluralistic approach offers a path forward to inclusive, culturally responsive digital tools in endangered language contexts.

These findings inform the design direction in the next chapter: not a singular platform or fixed architecture, but a modular, co-created infrastructure for bridging linguistic, cultural, and technological divides.

7. Design Solution

The previous chapters have explored how engagement with endangered languages manifests across a diverse range of community contexts. From online learner networks and diaspora-led revitalisation groups to heritage speakers and native speaker elders, the analysis revealed a highly heterogeneous landscape of language identities, practices, and priorities. Each group

interacts with language in distinct ways, shaped by historical, geographic, and socio-political conditions. Yet, despite these differences, the data also revealed a set of shared structural challenges that span across communities. These are fragmentation, resource scarcity, mistrust of digital tools, and limited opportunities for ethical, mutual engagement.

This chapter introduces a proposed design response to these intersecting realities: a shared, modular platform conceptualised as a boundary object (Star & Griesemer, 1989), capable of supporting diverse endangered language communities without imposing uniformity. Rather than designing separate tools for each community, the platform is intended as a flexible infrastructure that can be adapted, configured, and governed locally, while still providing shared scaffolding for challenges that are common across contexts.

Crucially, the rationale for a common design emerges from the thematic analysis itself. As shown in Chapter 6, endangered language stakeholders face different challenges in content, power dynamics, and goals, but experience remarkably similar barriers in the design and use of digital tools:

- Native speakers express concern over cultural misrepresentation, loss of control, and digital surveillance (Liu et al., 2022)
- Learners struggle with access to trustworthy resources, pronunciation feedback, and cultural gatekeeping.
- Both groups experience siloed engagement, limited interaction, and frustration with rigid or poorly tailored learning infrastructures (Gee & Hayes, 2012).

What links these experiences is not a shared identity, but a shared socio-technical condition, which is a lack of an ethical, usable, community-governed platform that supports revitalisation without flattening cultural nuance (Dantec & Disalvo, 2013). These overlapping and diverging needs are further synthesised in Table 6 below, which outlines key priorities, barriers, and design considerations across two central stakeholder groups: native speakers and language learners. This mapping builds on the thematic findings from Chapter 6 and grounds the design objectives that follow.

Community Group	Key Needs	Barriers Faced	Desired Tool Attributes
Native speakers	Cultural identity, respectful use	Data misuse; lack of digital access	Dialectal control; oral-first input; secure data sharing
Language learners	Accessible practice environments	Accent policing, limited resources	Scaffolded tools; community etiquette cues; pronunciation help

Table 6. Overview of user needs, barriers, and design implications across communities.

This table reinforces a central insight of this thesis: a one-size-fits-all solution is neither realistic nor desirable. Language revitalisation is not simply a shared mission; it is an overlapping set of community practices shaped by divergent experiences and asymmetrical power relations. As such, design infrastructures must not only support multiple pathways but also enable

coordination without collapsing under the differences. The platform proposed in this chapter responds directly to this landscape by offering modularity, co-governance, and dialect-sensitive tooling, supporting both shared scaffolding and local autonomy.

7.1 Design Objective

The primary objective of the proposed design is to create a modular digital infrastructure that supports the preservation and revitalisation of endangered languages by enabling participation, trust, and collaboration across different speaker communities. The platform is conceptualised as a boundary object, a shared artefact that can be interpreted and used differently by distinct groups, while still facilitating coordination and communication (Star & Griesemer, 1989).

The rationale for this shared platform stems from a core insight in the data, that being while each language community has its own identity and challenges, many of their infrastructural needs overlap (Liu et al., 2022). Rather than building isolated, community-specific tools from scratch, the proposed solution creates a shared framework that is:

- Flexible enough to support different linguistic traditions and governance models
- Specific enough to handle community-authored content and dialectal variations
- Scaffolded enough to enable interaction between users without enforcing standardisation.

The platform is composed of two integrated components:

1. A participatory archive, where users can upload, annotate, and curate linguistic and cultural materials, including audio stories, regional idioms, and pronunciation examples.
2. A social learning layer, which enables forums, mentorship, dialect-specific spaces, and etiquette sharing, encouraging mutual engagement without erasing linguistic complexity.

These components are designed to address the core themes identified in Chapter 6:

- **Mutual Engagement**: The platform facilitates dialogue between speakers and learners through peer-led forums and mentorship systems, helping bridge generational and geographic gaps.
- **Joint Enterprise**: By allowing different revitalisation goals to coexist, the design respects each community's definition of success, whether it is fluent usage, cultural knowledge transmission, or digital documentation.
- **Shared Repertoire**: The archive supports the creation of trustworthy, community-vetted resources that reflect local norms and resist generic, one-size-fits-all language content.

While each community is unique, the revitalisation landscape is increasingly networked, making it both logical and ethical to develop tools that can be reused, localised, and governed differently across contexts (Mainzinger, 2024).

Aligned with a community-informed human-centred design process, the design prioritises:

- **Co-governance**: Community-defined permissions, content moderation, and participatory decision making.
- **Transparency and trust**: Provenance tracking, opt-in data sharing, and annotated contributions.
- **Cultural specificity**: Orthographic variation support, dialect tagging, and multimedia content upload.
- **Bridge-building**: Features that support ethical contact, mutual learning, and cross-role collaboration.

The solution does not presume to solve language endangerment. Instead, it offers a modular foundation that communities can build upon to pursue revitalisation on their own terms, whether that means tightly controlled storytelling circles, open learning spaces, or dialectal repositories. In doing so, it responds directly to the third research question, proposing a design that is both technically responsive and socially respectful, grounded in the lived realities of those working to keep languages alive.

7.2 Concepts and Principles

This section outlines the key conceptual foundations and guiding principles that shape the proposed design. The principles articulated here emerge from the synthesis of the empirical findings presented in Chapter 6 and the theoretical frameworks established in Chapter 5. Specifically, the design responds to the dynamics of mutual engagement, joint enterprise, and shared repertoire (Wenger, 1998); the interpretive flexibility of boundary objects (Star & Griesemer, 1989); and the distributed, interest-driven nature of affinity spaces (Gee & Hayes, 2012). Rather than proposing a static or universal solution, the platform is conceptualised as a modular infrastructure that acknowledges both the distinctiveness of individual language communities and the cross-cutting challenges they share (Mainzinger, 2024). The design accommodates divergent practices, epistemologies, and revitalisation goals, while providing a shared foundation that communities can adapt according to their own terms.

7.2.1 Core Design Concepts

- **Boundary Object as Infrastructure**: The platform is designed as a boundary object, an artefact that different communities can interpret and use in different ways, while still enabling coordination and mutual recognition. This approach responds directly to the fragmentation identified in Chapter 6, offering a space where learners, native speakers, and other potential stakeholders can engage.
- **Plural Pathways to Revitalisation**: As demonstrated in the analysis, language communities do not share a single definition of fluency, legitimacy, or revitalisation success. The platform reflects this reality by supporting multiple forms of participation, ranging from cultural documentation to conversational learning, without privileging one path over another (Haidir et al., 2023).
- **Situated Repertoire**: Tools such as pronunciation aids, story archives, or discussion forums are not presented as generic or decontextualised. Instead, they are embedded in

a system that supports dialect-specific tagging, user annotations, and local governance, allowing communities to build a repertoire that is both shared and situated.

- **Trust as a Design Priority**: Given widespread concerns about data misuse and cultural misrepresentation, the platform embeds trust-building mechanisms, such as transparent moderation practices and opt-in sharing protocols that foreground consent at all levels(Ondiba, 2025; Zhong et al., 2024).

7.2.2 Community-Informed Design Principles

The following design principles were derived from recurring patterns in the data and grounded in the theoretical framing of sociotechnical systems. They reflect both the challenges and aspirations articulated by participants from different roles and language contexts.

Principle	Rationale
Support Co-Governance	Participants voiced concerns about loss of control over content and rules. The platform enables communities to define permissions, moderate content, and govern usage norms.
Preserve Dialectal and Cultural Nuance	Many were frustrated by tools that enforced standardisation. The design supports multiple orthographies, dialect tagging, and regional variations.
Ensure Transparency and Consent	Mistrust of data extraction and AI misuse emerged repeatedly. The platform includes clear content provenance, opt-in data sharing, and visible annotations
Facilitate Ethical Engagement	Learners and speakers often lacked structured, respectful ways to interact. Features such as etiquette guides, discussion norms, and role-sensitive forums help scaffold mutual engagement
Avoid Standardisation Bias	Many users rejected a one-size-fits-all tools. The platform supports diverse revitalisation goals without enforcing fixed models of language learning or fluency
Enable Modularity and Adaptability	Communities in needs and capacity. The platform is modular by design, allowing communities to activate only the features they find meaningful or appropriate

Table 7. Community-informed design principles based on empirical themes and sociotechnical theory.

7.2.3 Boundary Object as CoP Infrastructure

The proposed platform is not a fixed toolset but a boundary infrastructure, an adaptable, co-governed system that facilitates respectful participation and distributed ownership. Drawing from Wenger's (1998) dimensions of mutual engagement, joint enterprise, and shared repertoire, this section outlines how key components of the design directly address the tensions surfaced in Chapter 6.

The platform integrates two core components:

1. A participatory archive, supporting the curation, tagging, and annotation of community-sourced linguistic and cultural material.
2. A social learning network, enabling peer-to-peer support, language etiquette sharing, and collective practice.

These features operationalise the theoretical concepts explored earlier, addressing fragmentation, trust deficit, and asymmetrical legitimacy across stakeholder groups.

CoP Element	Platform Application
Mutual Engagement	Forums, dialect-specific spaces, language practice circles, mentorship opportunities
Joint Enterprise	Community-defined goals; co-governance tools; flexible participation paths
Shared Repertoire	Curated archive of oral histories, idioms, pronunciation aids, and dialect-tagged tools

Table 8. Mapping platform features to Wenger's Communities of Practice dimensions.

These mappings ensure the platform is responsive not only to technical constraints but to the relational and epistemic structures that define community life in endangered language contexts. Crucially, "effectiveness" here is not defined by engagement metrics or scalability, but by trust, dialectal fidelity, and participatory ethics.

Native speakers	Shared Priorities	Language Learners
Daily use rooted in cultural and intergenerational practices	Community inclusion and respectful engagement	Motivated by heritage, identity, or personal reconnection
Emphasis on oral transmission and context-driven interaction	Preservation and revitalisation of linguistic and cultural assets	Reliance on peers, digital tools, and informal spaces
Concern over language solutionism and extractive technologies	Ethical development and co-governance of digital tools	Need for structured resources, scaffolding, and feedback
Frustration with intergenerational discontinuity	Collaborative platforms for knowledge exchange	Difficulty finding mentorship and native speaker input
Scepticism toward outsiders or institutions	Creation of culturally sensitive, trusted learning environments	Regular exposure to accent policing and legitimacy challenges
Strong community gatekeeping norms and dialectal boundaries	Shared digital spaces that allow for diversity without dilution	Desire for pronunciation tools, dialectal clarity, and guidance

Figure 10. Summary of divergent needs and overlapping goals between native speakers and language learners.

These co-designed components position the platform not as a fixed solution, but as an adaptive infrastructure that scaffolds meaningful interaction across differences. The following section

introduces two representative personas, developed from empirical data, that illustrate how these platform elements can support distinct user needs and trajectories.

7.3 User Personas

In order to ground the proposed design in the lived experiences of its intended users, this section introduces two representative user personas developed from the qualitative data gathered. Personas are commonly used in human-centred design to synthesise key patterns in user behaviour, goals, and frustrations (Miaskiewicz & Kozar, 2011) enabling designers to empathise with diverse perspectives and make informed, user-aligned decisions.

While the broader language revitalisation ecosystem comprises a wide range of roles, including native and heritage speakers, language learners, linguists, educators, and community activists, this section focuses on the two primary groups most frequently discussed and represented in the dataset: native speakers and language learners. These groups reflect not only distinct identities but also divergent experiences with digital tools, contrasting expectations of revitalisation success, and at times conflicting definitions of linguistic legitimacy (Low et al., 2022).

Despite these differences, both groups also share overlapping challenges, such as a lack of trust in existing tools, difficulty navigating online spaces, and limited access to culturally situated learning or preservation environments (Taylor & Kochem, 2020; Zhong et al., 2024). The personas presented below reflect these tensions and commonalities, offering concrete reference points for design decisions made in the platform prototype.

Figure 11. Persona A representing the native speaker community.

Figure 12. Persona B representing the language learner community.

These personas are representative syntheses based on real participants' experiences. They reflect the dual imperative of this platform, to preserve and honour native speaker knowledge, while also supporting learners in ways that are ethical, context sensitive, and relationally grounded. The interactions between the personas form the relational fabric that the boundary object is designed to facilitate, not by enforcing shared norms, but by enabling mutual recognition and respectful contact across differences.

7.4 Prototype

The proposed platform is designed to function as a boundary object, providing a shared digital infrastructure that supports different language communities without imposing uniform standards or workflows (Star & Griesemer, 1989) Central to this approach is the implementation of role-sensitive interfaces. While all users interact within the same underlying system, the platform dynamically adapts its features, interface framing, and interaction flows according to the user's role.

Two core perspectives are supported:

- **Native Speaker Perspective**: Tailored for elders or fluent speakers focused on storytelling, preservation and cultural stewardship.
- **Learner Perspective**: Designed for heritage learners or second-language learners who seek guidance, practice, and ethical forms of engagement.

While these roles often reflect different levels of digital literacy, language authority, and expectations, the platform scaffolds ethical contact and mutual recognition through carefully

tailored features. Importantly, it also allows for proxy use (Dantec & Disalvo, 2013), recognising that native speakers, especially elders, may require the assistance of family members or facilitators to engage with digital tools. The sections below outline the user experience for each role, starting with the perspective of the native speaker.

Native Speaker Perspective

This interface prioritises oral-first workflows, simplicity of navigation, and full content control. It anticipates assisted use, enabling helpers to upload and manage content while preserving speaker authorship and decision-making authority.

Upon logging in, the user enters the home page, offering four primary actions: (1) recording a new story, (2) uploading a pre-recorded clip, (3) accessing the story archive, and (4) managing permissions. The interface emphasises accessibility, with large buttons, clear labels, and a voice-assisted guide for non-literate or elderly users

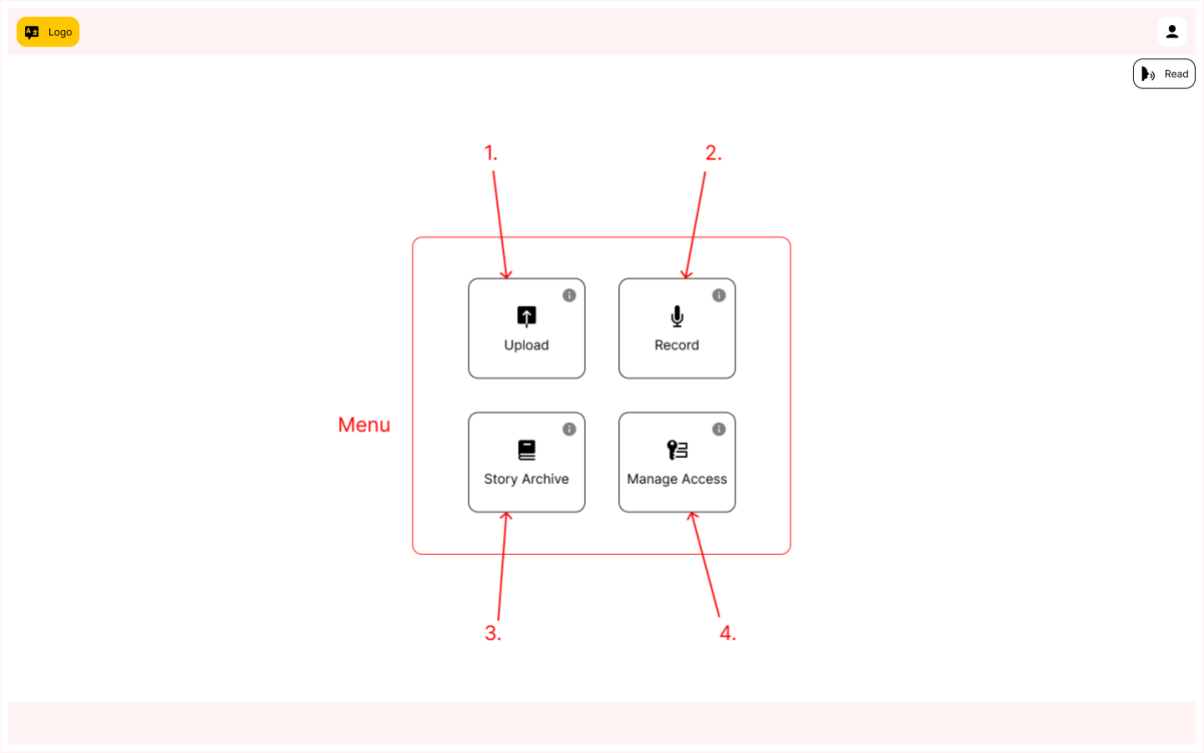

Figure 13. Native speaker dashboard, displaying primary actions.

When selecting *Record*, users encounter an audio interface with record, playback, and save functions. Alternatively, the *Upload* option allows users or their assistants to submit existing audio files (e.g., from mobile phones or community events). After recording or uploading, the system prompts for metadata entry, including language, dialect, speaker attribution, optional location, and cultural context (e.g., ceremonial or seasonal).

This tagging process is flexible by design, avoiding over-standardisation while supporting discoverability and dialectal fidelity (Hämäläinen et al., 2023)

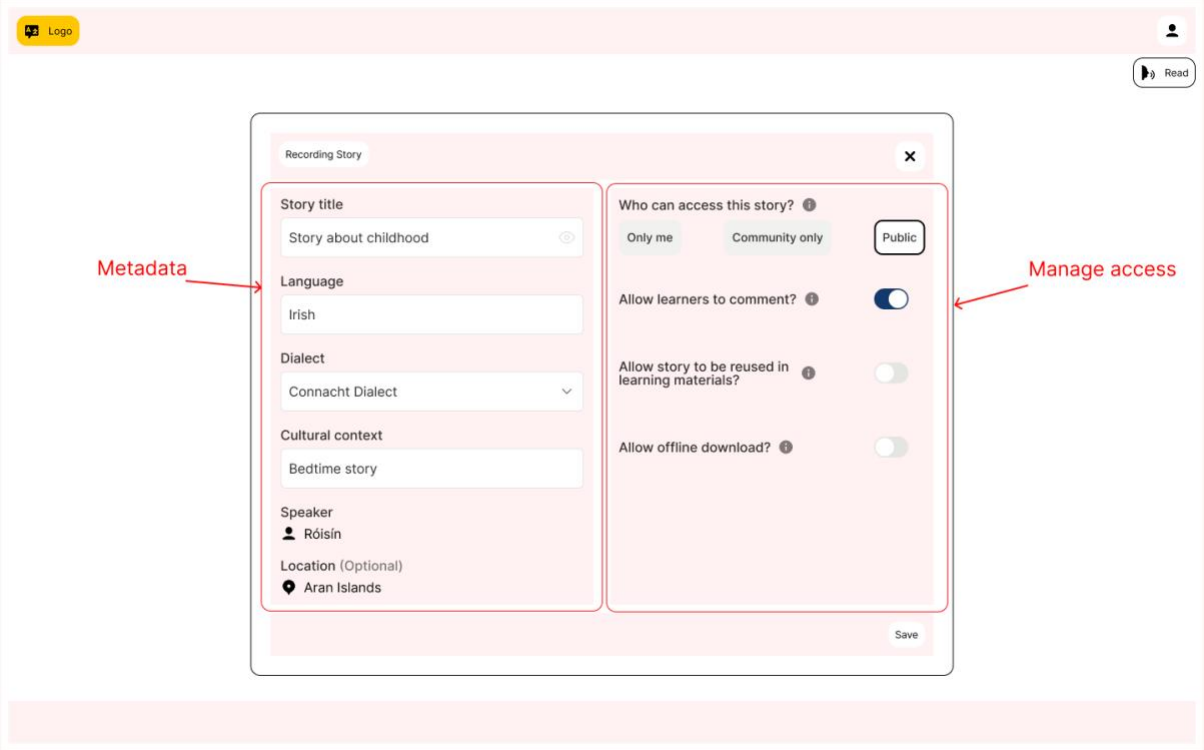

Figure 14: Metadata entry screen supporting dialect tagging and cultural context annotation.

The *Manage Access* interface foregrounds user consent, offering granular privacy settings: content can be restricted to the speaker, shared with the community, or made public. Additional permissions include whether learners may comment, download, or use the material in learning paths. Defaults lean conservative, with guidance text explaining each option's implications.

Once submitted, stories populate the *Story Archive*, a visual library sortable by title, tags, and permissions. Each entry displays playback controls, length, language, dialect, cultural tags, permissions, and user feedback depending on permission levels (e.g., "3 thank-yous"). The archive supports editing and deletion, with a layout optimised for low digital literacy.

To preserve speaker autonomy, public comment threads are disabled. Learners engage indirectly via a notification system, which speakers may ignore or respond to. Sample notifications include: "This story was added to a learner's practice path," or "A learner left a thank-you note." These communications are screened and optional, with no direct contact permitted without explicit consent.

Importantly, the system is designed to validate assisted participation. A granddaughter may record her grandmother's story, while a cultural facilitator may tag or upload on behalf of an elder, all while preserving the elder's attribution and control. This approach embraces existing relational infrastructures within oral communities and positions the tool as an extension of these practices rather than a substitute for them.

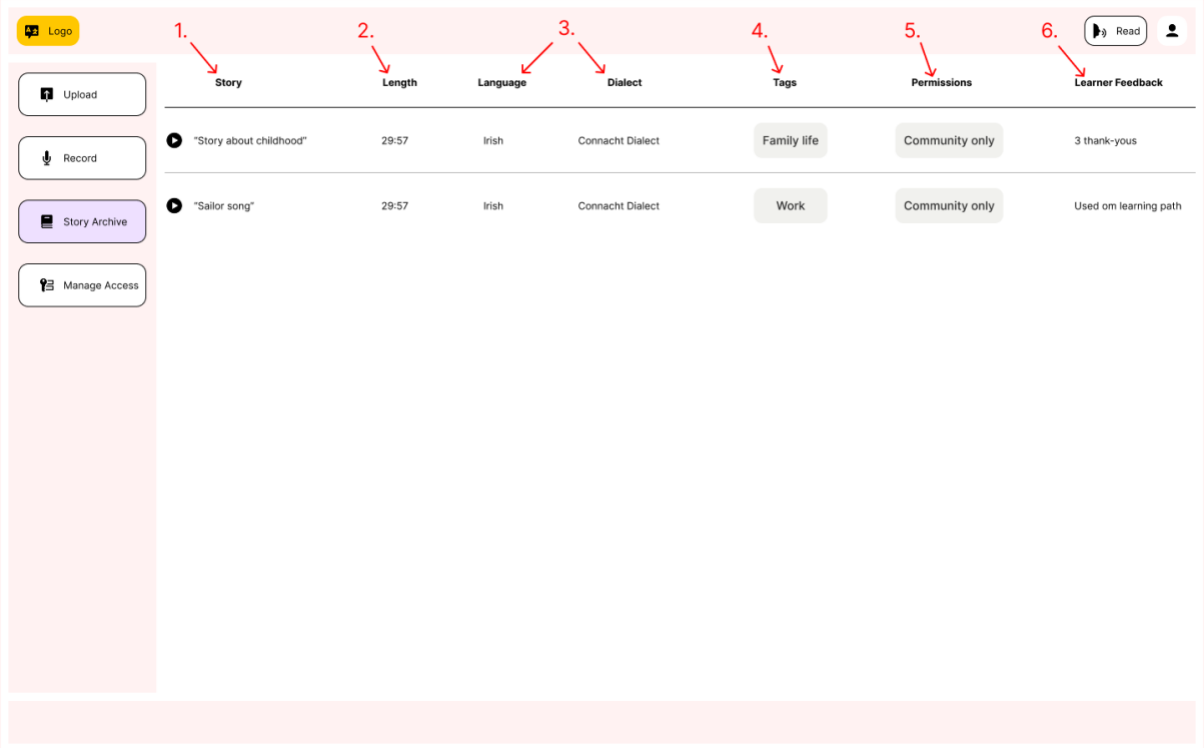

Figure 15. Story archive showcasing a list of previously uploaded stories.

Learner Perspective

The learner interface supports self-directed learning, relational ethics, and contextualised content access. It directly addresses challenges surfaced in the empirical data, including accent policing, legitimacy anxiety, and the need for scaffolded, respectful engagement (Gee & Hayes, 2012).

Upon login, learners enter a dashboard with three clear sections: *Learn*, *Library*, and *Community*. This structure allows for both exploration and structured progression.

In the *Learn* section, users define a personalised learning journey. They select a language or dialect, set goals (e.g., conversational fluency, ceremonial expression, or pronunciation refinement), and receive suggested materials accordingly. Metrics such as the number of stories listened to or exercises completed provide light scaffolding without gamified pressure.

The *Library* houses curated materials submitted or vetted by communities, including pronunciation clips, contextual dictionaries, translated stories with cultural annotations, and visual phrase guides. Users can filter by dialect, difficulty, and resource type. To foster trust, each entry includes visibility markers (e.g., “Community-approved”, “Speaker-submitted”).

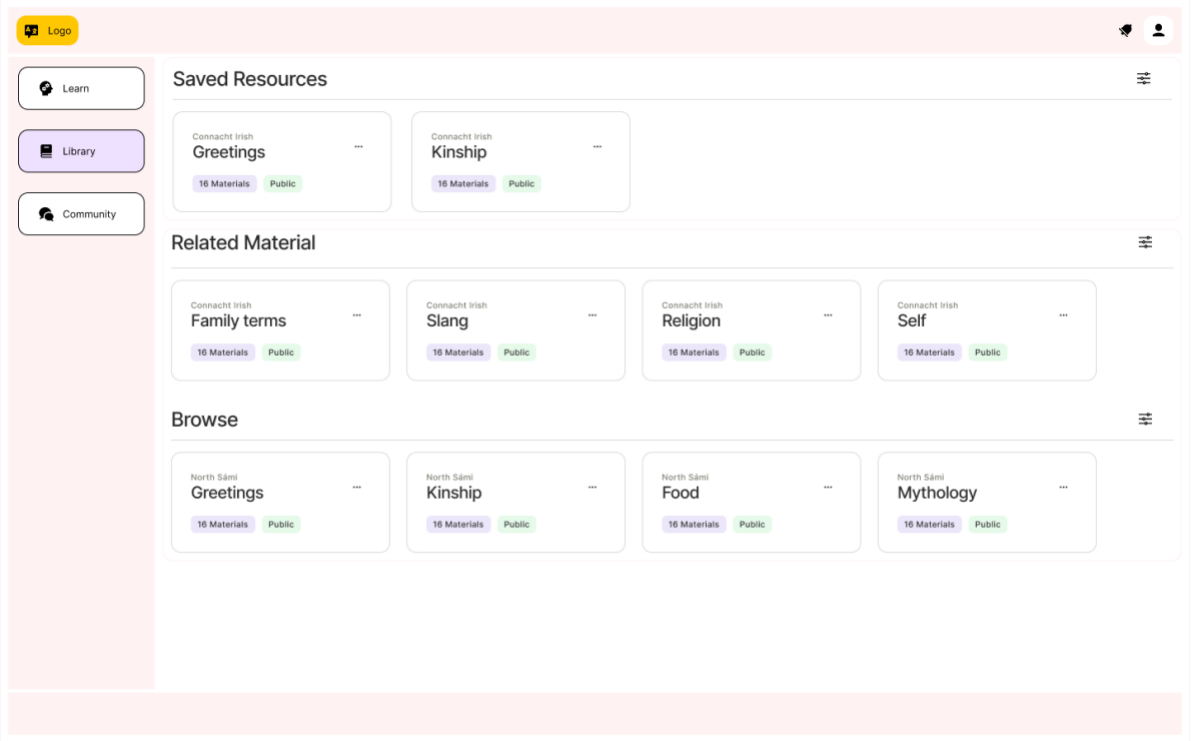

Figure 16. Library interface showcasing dialect-specific content with trust indicators.

The Community section functions as a semi-public forum, offering spaces that resemble moderated Discord Channels or learning hubs. These may be open, restricted by request, or require agreement to community norms. Inside, learners can access:

- Dialect etiquette guides and pronunciation protocols
- Non-public storytelling sessions or resources
- Forums and Q&A threads moderated by cultural facilitators

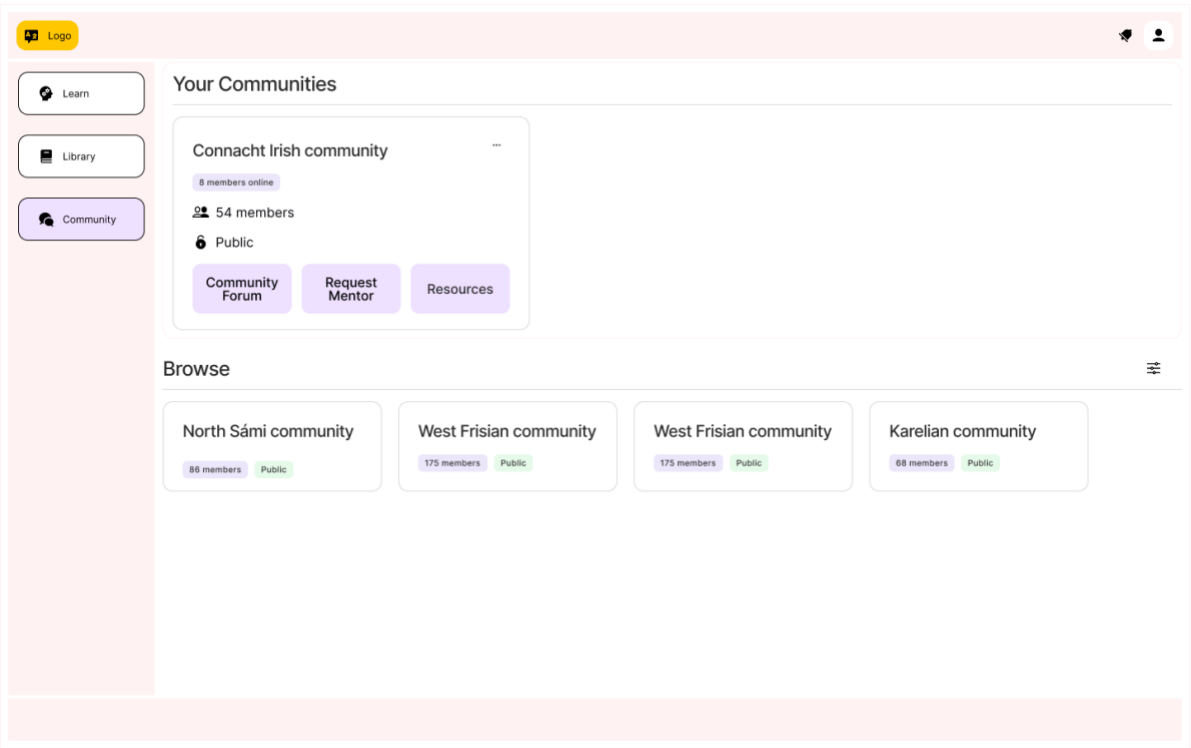

Figure 17. Community interface showcasing language communities the user is part of.

This approach supports differentiated access by respecting the boundaries around sacred or sensitive knowledge and offers learners the opportunity to build cultural fluency alongside linguistic skills. Before entering any community, users must accept a code of conduct, content norms, and relational expectations.

Rather than being passive recipients of content, learners thus become participants in a network of co-governed practices, where access is earned through recognition of community-defined etiquette and cultural legitimacy (Gee & Hayes, 2012; Wenger, et al., 2002).

7.5 Reflections

This chapter has outlined a role-sensitive, research-informed prototype developed in response to the needs, tensions, and aspirations identified through the analysis. At its core, the proposed design is not a solution to language endangerment, but a scaffold. It serves as a boundary object supporting ethical, relational, and culturally grounded forms of digital engagement between distinct user groups.

The prototype is deliberately modest in scope and low fidelity. Developed within the constraints of time, access, and research ethics, it is intended as a conceptual and functional sketch rather than a polished or immediately deployable product. Its modular structure, role-sensitive interface, and community-informed design principles serve to demonstrate how digital tools might better reflect the social realities of language revitalisation work, without collapsing epistemic difference into a universal model.

Crucially, the design recognises that many speakers, particularly elders, may not engage directly with digital infrastructure. As such, the prototype builds in support for assisted use, proxy authorship, and fine-grained permission settings, positioning the platform not as a replacement for relational language transmission but as a tool that can extend and support those relationships where appropriate.

The platform was shaped by a synthesis of empirical data and theory. The structure and feature set draw directly on insights from interviews and surveys, particularly the persistent fragmentation across communities of practices, asymmetrical distribution of authority and access, and the lack of trusted, culturally embedded learning environments. In turn, the design incorporates key elements of human-centred design, boundary object theory, and the affordances of affinity spaces, offering multiple entry points for interaction, contribution, and co-governance.

At the same time, the prototype's limitations are significant and intentional. It does not attempt to automate fluency, prescribe revitalisation paths, or replace community decision-making with top-down technical design. Instead, it offers a platform architecture that could evolve under community leadership, flexible enough to accommodate dialectal nuance, cultural boundaries, and plural definitions of success.

The next chapter turns to a broader discussion of what this design process reveals about digital toolmaking for endangered language communities. It critically evaluates the implications of

the prototype, its alignment with the original research questions, and its potential and limits as a model for ethical, socially embedded design in this space.

8. Discussion

This chapter will dive into the reflections, choice of methods, and the results in relation to the research questions and the problem statement.

8.1 Addressing the Research Questions

This section returns to the three research questions that guided the project. Each is addressed through synthesis of literature, empirical data, and design outcomes, providing a cohesive narrative of how the study advanced from exploration to intervention.

RQ1: "What existing AI-driven approaches have been used for the preservation and revitalisation of endangered languages, and what are their advantages and limitations?"

The literature review in Chapter 2 highlighted the increasing application of AI technologies to endangered language documentation and revitalisation, including tools based on automatic speech recognition (ASR), neural machine translation, and text-to-speech synthesis (Bird, 2020). These tools offer technical affordances such as speed, scalability, and broad accessibility, which have been celebrated in many NLP and digital linguistic initiatives (Zariquiey et al., 2022)(Hämäläinen et al., 2023).

However, a recurring critique across the literature is that these systems tend to adopt a top-down, standardising logic, often disregarding dialectal richness, speaker consent, or cultural boundaries (Bird, 2020). AI-driven approaches typically prioritise data over quantity over epistemic validity, leading to outcomes that may be linguistically accurate but culturally alienating or ethically problematic (Zhong et al., 2024).

This was echoed in the empirical findings. Several participants expressed concerns around inauthentic pronunciation, algorithmic flattening of dialects, and the loss of control over how their language is represented online. These critiques align with broader concerns about algorithmic neutrality and how AI design often reflects dominant knowledge systems, rather than community-led priorities (Avetisyan & Broneske, 2023) In endangered language contexts, especially those with strong oral traditions or decentralised authority structures, these tensions become particularly prominent.

RQ2: "What are the linguistic, cultural, and technological needs and challenges of endangered language communities?"

The analysis identified a complex constellation of needs and challenges. Linguistically, participants pointed to the lack of dialect-sensitive tools, limited access to accurate pronunciation models, and the absence of resources that reflect oral-first knowledge systems (Hämäläinen et al., 2023) Culturally, there were concerns about legitimacy policing, particularly towards heritage learners and second-language users, as well as broader anxieties about surveillance, platform ownership, and the appropriation of culturally embedded content (Ondiba, 2025)

Technologically, access remains uneven. Elders and rural speakers often lack the devices or digital literacy required to engage with most existing platforms, creating a reliance on intermediaries or cultural facilitators (Taylor & Kochem, 2020) At the same time, many participants noted the lack of peer-led, low-pressure learning environments, spaces where cultural etiquette, trust, and co-presence could guide engagement more than grammar drills or gamified fluency scores (Mainzinger, 2024)

These insights directly inform the design decisions presented in Chapter 7. The prototype incorporated:

- Audio-first controls and oral storytelling workflows, in line with CoP-informed participation patterns
- Access and privacy settings that support cultural gatekeeping (where deemed appropriate by community members) and co-governance
- Dialect tagging, learner-speaker separation, and community rules, echoing the boundary object approach used to support asymmetrical knowledge exchange
- Acknowledgement of proxy users and intermediaries, based on real-world practices observed in digitally mediated communities

The design was not a resolution to these challenges, but a response structured around flexibility and community control, in contrast to standardised language apps or automated AI interfaces.

RQ3: "How can an AI-driven digital tool be designed to effectively support the preservation and revitalisation of endangered languages based on these needs?"

The prototype developed in Chapter 7 prioritised human-centred design and community-informed values over automation (Haidir et al., 2023). While AI features such as pronunciation feedback or dialect-aware prompts were conceptually explored, they were deliberately constrained in favour of a role-sensitive, socially scaffolded interface. The underlying principle was that tools should follow social context, not precede or reshape it.

This aligns with critiques of solutionist approaches in tech design, particularly when applied to endangered language contexts (Low et al., 2022) Rather than presenting AI as an inevitable or optimal fix, the prototype proposes a relational infrastructure, a modular system where AI could be layered in only if communities choose to integrate it on their own terms.

It is important to note that the prototype was a low-fidelity one, not a production-ready platform. Time constraints limited the ability to conduct iterative testing or co-design workshops. Still, the design is grounded in strong ethical alignment and research-driven logic. Its value lies in illustrating an alternative to dominant, extractive AI models, one that centres community sovereignty, access, and relational design.

8.3 Theoretical, Methodological, and Design Reflections

This project was shaped by a combination of methodological pragmatism and ethical sensitivity. Initially rooted in an endangered Uralic language context, the scope was broadened to encompass endangered languages more generally due to limited data access and a recognition that many endangered language communities face overlapping structural

challenges. This decision, while methodologically necessary, required balancing specificity with transferability. While the design does not claim to represent all endangered language contexts equally, it draws on shared concerns, such as dialectal erasure, lack of community control, and oral-first traditions, that emerged consistently throughout the data.

From a methodological perspective, the decision to adopt a human-centred design (HCD) approach proved well-suited to the project's ethical and relational orientation. HCD allowed for a flexible yet grounded way to move from empirical insight to design solution, without assuming universal user behaviour or homogenous needs (Norman, 2013). Its iterative nature aligned with the project's sensitivity to complexity, even if time constraints limited the number of actual design iterations. The use of research-based personas helped to capture the distinct epistemic positions of native speakers and learners, highlighting asymmetries in trust, access, and legitimacy that shaped the interface architecture (Miaskiewicz & Kozar, 2011).

Theoretically, the concepts of communities of practice and boundary objects provided valuable frameworks for understanding how engagement around endangered languages is structured, negotiated, and sometimes contested. Wenger's model clarified the fragmentation across mutual engagement, joint enterprise, and shared repertoire, helping frame why so many current tools fail to create sustainable learning ecologies. The boundary object concept informed both the structure and governance of the proposed platform, supporting the idea that digital tools can mediate between communities with different goals, provided they are flexible, co-governed, and interpretable across contexts. In addition to these, the concept of affinity spaces (Gee & Hayes, 2012) provided a valuable lens for interpreting less formalised engagement, particularly among learners. These spaces, which prioritise shared interest over credentialed participation, helped explain the function of WhatsApp groups, Discord servers, and informal peer learning communities that lacked institutional affiliation but nonetheless facilitated important language practice.

The design itself evolved through a process of analytical synthesis, driven not by feature ambition but by ethical, contextual, and temporal constraints. The choice to create a low-fidelity prototype was not simply a matter of feasibility, but a conscious methodological position. This early stage of the prototype enabled more structural reflection, which was particularly important given that many participants expressed ambivalence or scepticism towards digital tools, particularly AI-driven ones. A high-fidelity prototype might have implied a finished solution, rather than a prompt for continued community negotiation and adaptation.

Finally, the decision to design for assisted use, particularly for elder native speakers, marked a critical methodological and ethical stance. Rather than treating limited digital literacy as a barrier to be overcome through training or interface simplification alone, the design acknowledged relational infrastructures already present in many communities, intergenerational help, facilitation, and proxy use (Dantec & Disalvo, 2013). This allowed the platform to accommodate multiple user realities without reducing users to a single ideal type.

In sum, the methodological and theoretical orientation of this project enabled a design process that was not only feasible within the scope of a master's thesis but also reflective of the real tensions and practices present in endangered language communities. The prototype does not

resolve these tensions, but it embodies a design logic grounded in care, asymmetry, and the possibility of co-governed digital infrastructures.

8.4 Implications for Future Work

While this thesis presents a conceptual prototype grounded in empirical and theoretical insights, its broader value lies in offering direction for future work, both in design and research. A key implication is the need to transform the current prototype into an operational, iteratively developed platform. Testing sessions with native speakers, heritage speakers, learners, linguists, and educators across different linguistic communities will allow for the refinement of usability features and trust dynamics. Future work could explore modular platform design in more depth, enabling communities to activate or deactivate components, such as dialectal variation tagging, region-specific orthographies, or content vetting workflows, based on their own governance models. Iterative prototyping cycles with community feedback are essential for moving from conceptual to practical impact. Co-design workshops and longitudinal testing should be integrated early in future development phases to ensure ongoing alignment with community expectations and platform usability.

Another promising direction involves expanding the platform's functionality to include collaborative corpus development, crowd-sourced audio annotation, and "consent-aware" transcription pipelines that allow speakers to control how their contributions are used. These features would extend the platform's utility beyond individual learning to support community archiving and intergenerational teaching practices.

Future development could prioritise a mobile-optimised version of the platform to improve accessibility for communities with limited access to desktop computers. This could also support informal, on-the-go language use, such as WhatsApp-based conversations or voice note exchanges, which several participants reported as key learning practices.

A further area for development is the incorporation of regional identity and linguistic variation, particularly for language families such as the Uralic group, which shaped the initial scope of the thesis. Future iterations could return to this context by piloting the prototype in collaboration with Uralic-speaking communities (e.g., Mansi, Udmurt), testing for dialectal adaptability, cultural fit, and governance alignment (Hämäläinen, 2023). Additionally, future research should also explore governance models that ensure community ownership and decision-making authority over the platform's development and data use, potentially drawing on indigenous data sovereignty principles (Walter & Suina, 2018)

Finally, long-term development could benefit from partnerships with indigenous-led tech organisations, grassroots revitalisation initiatives, or academic centres working on endangered language AI to ensure ethical and sustained impact.

8.5 Comparison with Prior Literature

The findings of this study support and expand upon existing literature in several key areas. First, they confirm critiques that mainstream AI tools often neglect the social and epistemic

realities of endangered language communities. Studies such as those by Liu (2022) and Low (2022) have emphasised that without direct community governance, AI tools risk reproducing dominant linguistic ideologies and undermining cultural authenticity.

This thesis responds by proposing a boundary object approach to tool design, one that is flexible, participatory, and sensitive to the diverse roles of users. Such an approach aligns with the frameworks advanced by Mainzinger (2024), who argues for tools that scaffold rather than replace human revitalisation efforts.

The need for modular, role-sensitive design also mirrors recommendations found in prior evaluations of language tools such as LemkoTran.com and Whisper (Orynycz, 2023; Chen, 2023), where technical success was not matched by cultural relevance or governance flexibility. Participants in this study similarly emphasised that tools must accommodate variations in fluency, orthography, and identity, which rigid AI systems often overlook.

Moreover, the fragmented experiences among native speakers and language learners (heritage or second language) identified in this thesis reflect the participatory disparities found in affinity spaces and Communities of Practice literature (Gee & Hayes, 2012; Wenger, 1998). By designing for overlapping yet distinct engagement practices, this study adds practical strategies for mitigating the tensions previously highlighted in the field.

8.6 Methodological and Design Limitations

This research offers valuable insights but also has several limitations that should be addressed in future studies.

From a methodological perspective, the absence of sustained co-design workshops limited the depth of participatory involvement. While interviews and surveys informed the prototype's design logic, live iterative sessions could have offered richer insights into interface preferences and feature priorities.

The survey sample, though diverse in geography and roles, was limited in size and skewed toward digitally engaged respondents. This may underrepresent more marginalised speakers, particularly those with limited access to digital tools or who operate outside dominant literacy models.

The interviews, while thematically rich, lacked longitudinal depth. A more extended engagement over time might have uncovered shifts in participant attitudes, especially concerning emerging AI tools or intra-community tensions.

Design-wise, the prototype remains conceptual. It does not yet account for real-world challenges such as internet connectivity, localisation barriers, or platform governance structures. Furthermore, while the prototype was informed by community needs, it lacks formal validation through user testing or scenario-based evaluation.

Finally, although the original project aimed to focus on Uralic languages, the limited number of participants necessitated a broader scope. As a result, the findings offer generalisable insights but are not deeply situated within any single linguistic or cultural tradition. Revisiting this context in future work could validate whether the prototype design holds relevance across

typologically diverse or highly endangered Uralic languages, among others (Hämäläinen, 2023).

9. Conclusion & Future Work

This thesis set out to explore how AI technologies can be meaningfully and ethically integrated into endangered language revitalisation, with a focus on supporting both native speakers and language learners through a modular, community-governed digital platform. By drawing on literature, expert interviews, and user surveys, the study addressed three core research questions concerning the current landscape of AI-driven tools, sociotechnical needs of language communities, and the principles that support effective design.

The findings revealed that while AI holds considerable potential, particularly in areas like speech recognition, machine translation, and content annotation, many of the existing tools prioritise efficiency over cultural fit. This often reinforces standardised, top-down approaches that marginalise dialectal diversity, oral traditions, and community agency. This was echoed by participants, who voiced concerns about authenticity, consent, legitimacy, and the invisibility of relational and contextual knowledge in current digital environments.

In response, this thesis proposed a prototype platform conceptualised as a boundary object: a flexible infrastructure that supports diverse interpretations and roles while scaffolding coordination and ethical contact. Informed by frameworks such as Communities of Practice (Wenger, 1998), boundary objects (Star & Griesemer, 1989), and affinity spaces (Gee & Hayes, 2012), the design promotes co-governance, dialectal nuance, and participatory engagement. It does not aim to replace human-led revitalisation efforts, but to provide support by recognising that language is not merely data to be processed, but a living relational practice embedded in culture, place, and the people.

By emphasising trust, adaptability, and community control, the prototype advocates for a design approach grounded in the real-world needs of community members. While the scope of this thesis did not allow for full-scale testing and implementation, the work contributes a clear conceptual and ethical foundation for future development.

Ultimately, this project illustrates that the challenge of endangered language revitalisation is not solely technical, but relational. Effective tools must not only be functional but respectful, not only interoperable but interpretable. As such, the work presented here offers not a finished solution, but a starting point: an invitation to build technologies that honour linguistic diversity, foster mutual recognition, and help communities shape their digital futures on their own terms.

10. References

Avetisyan, H., & Broneske, D. (2023). Large Language Models and Low-Resource Languages: An Examination of Armenian NLP. *Proceedings of the International*

Conference on Natural Language Processing (pp. 199–210). Association for Computational Linguistics.

Bird, S. (2020). Decolonising Speech and Language Technology. *Proceedings of the 28th International Conference on Computational Linguistics*. (pp. 3504–3519). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Blasi, D. E., Anastasopoulos, A., Neubig, G., & Roark, B. (2022). Systematic inequalities in language technology performance across the world's languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. (pp. 5486–5505). Association for Computational Linguistics.

Braun, V., & Clarke, V. (2014). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), (pp. 77–101). https://doi.org/10.1191/1478088706qp063oa

Chekole, A. K., Asfaw, T. T., Mengestie, T. N., Kebie, B. T., Negia, M. K., & Worku, Y. A. (2024). Effect of Parallel Data Processing Model on Bi-Directional English-Khimtagne Machine Translation Using Deep Learning. 2024 *International Conference on Information and Communication Technology for Development for Africa* (ICT4DA), (pp. 189–193). IEEE. https://doi.org/10.1109/ict4da62874.2024.10777148

Chen, C., Chang, C., & Hsu, Y. (2023). Accelerating Hakka Speech Recognition Research and Development Using the Whisper Model. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023),* (pp. 367–370). The Association for Computational Linguistics and Chinese Language Processing.

Dantec, C. A. L., & Disalvo, C. (2013). Infrastructuring and the formation of publics in participatory design. *Social Studies of Science, 43*(2), (pp. 241–264). https://doi.org/10.1177/0306312712471581

Dwivedi, P., Mathews, S., & Majumder, S. (2020). Predicting Language Endangerment: A Machine Learning Approach. *11th International Conference on Computing, Communication and Networking Technologies* (ICCCNT). (pp. 1–7). https://doi.org/10.1109/ICCCNT49239.2020.9225576

Elsner, M., & Needle, J. (2023). Translating a low-resource language using GPT-3 and a human-readable dictionary. *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. (pp. 1–13). Association for Computational Linguistics.

Gee, J. P., & Hayes, E. (2012). Nurturing Affinity Spaces and Game-Based Learning. In C. Steinkuehler, K. Squire & S. Barab (Eds.), *Games, Learning, and Society: Learning and Meaning in the Digital Age* (pp. 95–112). Cambridge University Press.

Haidir, H., Sinar, T. S., Mulyadi, Setia, E., & Saragih, E. (2023). Contextualizing Revitalization of Panai Malay Archaic Vocabularies Through Formal Learning. *Theory and Practice in Language Studies, 13*(6), (pp. 1506). https://doi.org/10.17507/tpls.1306.19

Hämäläinen, M., Rueter, J., Alnajjar, K., & Partanen, N. (2023). Working Towards Digital Documentation of Uralic Languages With Open-Source Tools and Modern NLP Methods. *Proceedings of the Big Picture Workshop*. (pp. 18–27). Association for Computational Linguistics.

Hinton, P. R. (2014). Chapter 2: Descriptive statistics. *Statistics Explained* (pp. 5–25). Routledge.

Houde, S., & Hill, C. (1997). What do prototypes prototype? *Handbook of Human-Computer Interaction* (2nd ed., pp. 367–381). Elsevier.

IDEO (Firm). (2015). *The field guide to human-centered design : design kit*(1st ed.). IDEO.

Kirk, A. (2012). *Data Visualization: A Successful Design Process*. (pp. 79-117, 119-158). Packt Publishing.

Liu, Z., Richardson, C., Hatcher, R., Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers*). (pp. 3933–3944). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.computel-1.4

Low, D. S., McNeill, I., & Day, M. J. (2022). Endangered Languages: A Sociocognitive Approach to Language Death, Identity Loss, and Preservation in the Age of Artificial Intelligence. *Sustainable Multilingualism, 21*(1), (pp. 1–25). https://doi.org/10.2478/sm-2022-0011

Mainzinger, J. (2024). Technology and Language Revitalization: A Roadmap for the Mvskoke Language. *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*. (pp. 7–12). Association for Computational Linguistics

Miaskiewicz, T., & Kozar, K. A. (2011). Personas and user-centered design: How can personas benefit product design processes? *Design Studies, 32*(5), (pp. 417–430). https://doi.org/10.1016/j.destud.2011.03.003

Miyagawa, S., Kato, K., Zlazli, M., Carlino, S., & Machida, S.Building Okinawan Lexicon Resource for Language Reclamation/Revitalization and Natural Language Processing Tasks such as Universal Dependencies Treebanking. *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains.* (pp. 86–91). Association for Computational Linguistic

Norman, D. A. (2013). *The design of everyday things (Revised and expanded ed.)*. MIT Press.

Ondiba, H. (2025). Proactive AI-Driven Cybersecurity for Endangered Language Preservation: Safeguarding the Suba Linguistic Corpus. *4th IEEE International Conference on Artificial Intelligence in Cybersecurity, ICAIC 2025,* (pp. 1–5) Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/icaic63015.2025.10848675

Orynycz, P. (2022). Say It Right: AI Neural Machine Translation Empowers New Speakers to Revitalize Lemko. In Degen, H., & Ntoa, S. (Ed.), *Artificial Intelligence in HCI. HCII 2022. Lecture Notes in Computer Science* (pp. 567–580). Springer. https://doi.org/10.1007/978-3-031-05643-7_37

Orynycz, P. (2023). BLEU Skies for Endangered Language Revitalization: Lemko Rusyn and Ukrainian Neural AI Translation Accuracy Soars. (pp. 135). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35894-4_10

R Core Team. (2023). R: A language and environment for statistical computing [computer software]. R Foundation for Statistical Computing:

Romero, M., Gómez-Canaval, S., & Torre, I. G. (2024). Automatic Speech Recognition Advancements for Indigenous Languages of the Americas. *Applied Sciences, 14*(15), 6497. https://doi.org/10.3390/app14156497

Sanders, N. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign, 4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Soylu, D., & Şahin, A. (2024). The Role of AI in Supporting Indigenous Languages. *AI and Tech in Behavioral and Social Sciences, 2*(4), (pp. 11–18.) https://doi.org/10.61838/kman.aitech.2.4.2

Star, S., & Griesemer, J. (1989). Institutional Ecology, "Translations," and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology. *Social Studies of Science, 19*(3), (pp. 1907–1939)

Taylor, J., & Kochem, T. (2020). Access and empowerment in digital language learning, maintenance, and revival: a critical literature review. *Diaspora, Indigenous, and Minority Education, 16*(4), (pp. 234–245). https://doi.org/10.1080/15595692.2020.1765769

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003, March). *Language vitality and endangerment*. Paper presented at the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, Paris, France. https://unesdoc.unesco.org/ark:/48223/pf0000183699

Vo, H. N. K., Le, D. D., Phan, T. M. D., Nguyen, T. S., Pham, Q. N., Tran, N. O., Nguyen, Q. D., Vo, T. M. H., & Quan, T. (2024). Revitalizing Bahnaric Language through Neural Machine Translation: Challenges, Strategies, and Promising Outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*. (pp. 23360–23368). https://doi.org/10.1609/aaai.v38i21.30385

Walter, M., & Suina, M. (2018). Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of Social Research Methodology, 22*(3), (pp. 233–243). https://doi.org/10.1080/13645579.2018.1531228

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business School Press. https://doi.org/10.5465/amle.2009.41788855

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

Wickham, H. (2022). stringr: Simple, consistent wrappers for common string operations [R package] [computer software]

Wickham, H. (2023). tidyverse: Easily install and load the tidyverse [R package] [computer software]

Wickham, H., François, R., Henry, L., & Müller, K. (2023). dplyr: A grammar of data manipulation [R package] [computer software]

Wickham, H., & Girlich, M. (2023). readr: Read rectangular text data [R package] [computer software]

Zariquiey, R., Oncevay, A., & Vera, J. (2022). CLD²: Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages. *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 20–30). Association for Computational Linguistics

Zhong, T., Yang, Z., Liu, Z., Zhang, R., Liu, Y., Sun, H., Pan, Y., Li, Y., Zhou, Y., Jiang, H., Chen, J., & Liu, T. (2024). *Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research.* https://arxiv.org/abs/2412.04497

11. Appendices

Appendix A - <Literature approval>

Rikke Magnussen <rikkem... Monday, 26 May 2025 at 12.24

To: Patrik Bugar

Hi Patrik
Thank you. The reference list looks good. Hereby approved.
Best
Rikke

Appendix B - <Interview outline>
Appendix C - <Survey results>
Appendix D - <Transcript Mansi>
Appendix E - <Transcript Chamorro>
Appendix F - <Transcript Irish heritage>
Appendix G - <Transcript Irish fluent>
Appendix H - <Transcript Nahuatl>
Appendix I - <Transcript Linguist>
Appendix J - <Literature bookkeeping>
Appendix K - <R Code for survey analysis>

Table of Figures

Tables