The Linguistic Marginalization of Low-Resource Languages: Confronting AI's Limits in Digital Cultural Heritage

Annie K. Lamar

Library Trends, Volume 75, Number 1, August 2026, pp. 137-158 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/lib.2026.a999944

The Linguistic Marginalization of Low-Resource Languages: Confronting AI's Limits in Digital Cultural Heritage

Annie K. Lamar

ABSTRACT

Cultural-heritage institutions are increasingly digitizing vast collections to preserve and make accessible the histories of diverse communities, including those forcibly displaced. However, a significant and often overlooked challenge in this endeavor lies in the inherent limitations of current artificial intelligence (AI) models, particularly for non-English and low-resource linguistic communities. This article examines how the scarcity and inadequacy of AI tools for languages such as Palestinian Arabic impede deep scholarly and public engagement with crucial historical narratives. Drawing on a detailed case study of the Palestinian Oral History Archive, this article illustrates how a community-governed, linguistically attentive digital archive exposes the limitations of English-centric, high-resource natural language processing pipelines. This article also addresses broader structural inequities shaped by market incentives, data colonialism, institutional capacity gaps, and the geopolitical constraints surrounding Palestinian linguistic data. These dynamics constitute forms of testimonial and hermeneutic injustice. This article concludes by outlining pathways toward linguistic justice in cultural-heritage AI, emphasizing community-led model development, dialect-sensitive tools, ethical governance frameworks, and collaborative institutional infrastructures that can sustain equitable access for marginalized linguistic communities.

KEYWORDS

artificial intelligence, low-resource languages, algorithmic invisibility, epistemic injustice, oral history, digitization, data colonialism

LIBRARY TRENDS, Vol. 75, No. 1, 2026 ("Cultural Heritage in Crisis," edited by Anna E. Kijas and Andreas Segerberg), pp. 137–158.

Introduction: Cultural Heritage in an Age of Crisis

Across the globe, archives face the dual mandate to preserve at-risk materials and broaden access for the communities whose histories they steward. Large-scale digitization, often heralded as a solution to both, has become a defining strategy of the contemporary archival landscape. Digitization is now "the only way to provide future access" for fragile or endangered materials, yet also "one of the main risks to archival and library collections" (VanSnick and Ntanos 2018, 282). The imperative to digitize is particularly acute for collections documenting communities displaced by war, occupation, and state violence, where the fragility of physical records is inseparable from the precarity of the people represented within them. For many displaced communities, digitization is often the sole mechanism preventing these fragile records from being lost entirely. At the same time, the imperative to digitize places these testimonies into infrastructures that can reinscribe the very power asymmetries that imperiled the records in the first place.

Palestinian archives exemplify this condition of crisis. Decades of dispossession, military assault, and the targeted destruction or seizure of cultural heritage have produced a system of "two violent acts of elimination": the physical elimination of indigenous Palestinians and the epistemic elimination enacted through the destruction or confiscation of archives, schools, and cultural institutions (Moon 2023, 38). Oral history collections such as the Palestinian Oral History Archive (POHA) emerged precisely to counter this silencing. The POHA documents the testimonies of Palestinians displaced in 1948 and their descendants, preserving what POHA calls the "integrity, immediacy, and oral nature of the material at hand" (Sleiman and Chebaro 2018, 65). These testimonies challenge the long-standing "bias in favor of 'objective' archival evidence" and assert that Palestinian history is composed of "clouds of stories" carried in lived experience (Allan 2007, 10).

Digitization has been transformative for these collections. Projects like POHA were built to "preserve and consolidate the wider region's threatened oral history and intangible heritage" and provide "reliable long-term access" through a unified digital platform (Sleiman and Chebaro 2018, 73). Digital Palestinian archives make possible "anti-hierarchical engagements with archival materials" and enable users worldwide to "partake in shaping Palestinian narratives that challenge Israeli ones" (Moon 2023, 46). Even under conditions of occupation, censorship, and fragmentation, such platforms "democratize access to primary sources" and facilitate the recovery of "Palestinian dialects and vernacular expressions" otherwise at risk of erasure (Moon 2023, 46; see Allan 2021).

The accelerating integration of artificial intelligence (AI) into archival workflows introduces another crisis, layered atop the material vulnerabilities digitization was meant to resolve. This crisis stems from a dependence

on AI systems whose design priorities and training choices—shaped by market incentives, institutional decisions, and geopolitical structures—have left non-English, low-resource, and dialect-rich languages systematically underserved. While digitization expands access in theory, the reality is that most contemporary discovery tools—for example, search engines, AI-assisted reference systems—are built for dominant languages, as are the processing pipelines that produce the data they depend on: machine translation, automatic speech recognition (ASR), and automated metadata generation. Their underlying architectures presuppose textual regularity, standardized orthographies, and large, stable corpora. Such assumptions are fundamentally misaligned with the linguistic realities of Palestinian Arabic and with the testimonial forms through which displaced communities narrate experience.

This dynamic reflects what we might call *algorithmic invisibility*: the systematic exclusion of marginalized languages, narratives, and epistemologies from AI-driven infrastructures. Even as oral histories move into digital form, they remain largely illegible to the computational systems that increasingly mediate access to cultural heritage. Consider POHA's own cataloging process. Despite the researchers designing "custom-tailored subject headings and keywords that remained faithful to the language used by the narrators" and building a bilingual thesaurus to capture localized knowledge (Sleiman and Chebaro 2018, 70), the project's materials still depend on researcher-generated metadata because no existing automated tools can reliably process the linguistic and narrative complexity of the recordings in the archive.

Situating this failure within broader structures of power makes clear that the crisis at hand is not merely technical but epistemic. It reflects what Miranda Fricker (2007) terms epistemic injustice, wherein marginalized communities are denied the conditions necessary for their knowledge to be heard, interpreted, or trusted. When AI systems cannot support the languages of displaced peoples, they reproduce the very archival silences that projects like POHA were created to counter (see Trouillot 2015). The erasure of the Nakba[1] has long been maintained through deliberate archival suppression, including the removal, concealment, and controlled release of 1948-era records, which regulate "who has permission to narrate the past" and whose testimony is recognized as authoritative (Anziska 2019, 64). This entrenched pattern of erasure forms the conditions under which contemporary digital systems operate. Digital archives cannot on their own "account for the gaps and silences" produced by settler-colonial destruction (Moon 2023, 38). When computational systems fail to read, transcribe, translate, or index Palestinian Arabic, oral histories become dependent on partial or inaccessible metadata schemas that constrain discovery and interpretation. Failures like these expose the structural limits of current AI model architectures, training pipelines, and evaluation

practices, all of which presume linguistic conditions that low-resource communities often do not share.

This article argues that the human and institutional decisions embedded in AI systems have produced a new frontier of cultural-heritage inequity in which linguistic marginalization is encoded at the infrastructural level rather than enacted through individual acts. This crisis of inequity is not limited to questions of preservation or access but is rooted in entrenched forms of linguistic and epistemic injustice, that is, in the structural conditions under which certain communities become legible within computational systems while others are rendered invisible. Without culturally informed, community-governed AI tools, the digitization of Palestinian oral histories risks reinforcing, rather than redressing, the marginalization of Palestinian narratives.

Algorithmic Invisibility and Epistemic Injustice

In natural language processing (NLP) and machine translation, so-called low-resource languages are those for which large, standardized, and readily accessible training corpora do not exist (Joshi et al. 2020). This definition, however, obscures the fact that "low-resource" is not a property of languages themselves but, rather, of the political, economic, and infrastructural conditions that determine whose linguistic forms are widely collected, digitized, and valued within global AI ecosystems. As Nekoto et al. note, "'Low-resourced'-ness is a complex problem going beyond data availability and reflects systemic problems in society," and "resource-scarce languages in an NLP context reflect the resource scarcity in the society from which the speakers originate" (2020, 2144–45). Languages with dense state support, extensive publishing histories, and continuous digitization pipelines become "high-resource," while the languages of displaced, colonized, or marginalized communities are persistently underrepresented (see Blasi et al. 2022; Couldry and Mejias 2019; Joshi et al. 2020).

Even multilingual models designed to bridge linguistic inequity reproduce these divides. Multilingual embeddings often bias toward high-resource languages whose statistical patterns dominate shared vector spaces, marginalizing dialects and minority languages by compressing or distorting their semantic relations (Sharma et al. 2025). Machine translation pipelines frequently rely on English as an "interlingua" or pivot language, imposing English-centric structures on languages that do not map neatly onto its grammatical or semantic categories (Tiedemann 2020). Low-resource and dialectal languages suffer from scarcity of high-quality datasets, instability in few-shot performance, and difficulties handling dialectal variation, which leads to degraded performance in classification, translation, and other NLP tasks (Lamin and Aziz 2025).

The dialectal heterogeneity of Arabic is a structural challenge for computational modeling. Many existing NLP and annotation tools built for

Modern Standard Arabic (MSA) or Egyptian Arabic fail or underperform when applied to Palestinian Arabic data (Jarrar et al. 2017, 747). Arabic is a constellation of geographically and socially inflected dialects that differ substantially from one another and from MSA. These dialects have developed through centuries of contact with Berber, Syriac, Turkish, Persian, French, and English and "substantially differ from MSA and each other in terms of phonology, morphology, lexical choice and syntax" (Habash 2010, 2). Arabic dialects "vary widely from region to region and to a lesser extent from city to city," and unlike MSA, often they "are not standardized, they are not taught, and they do not have official status," despite being the primary medium of everyday communication across the Arab world (Bouamor et al. 2014, 1240). Palestinian Arabic belongs to a larger dialect continuum (Levantine Arabic/Shami) but carries region-specific phonological, orthographic, lexical, and sociolinguistic features that complicate computational processing (Jarrar et al. 2014, 18). Additionally, morphosyntactic and lexical variation means that many spoken verbs, negation patterns, and common vocabulary in Palestinian Arabic diverge sharply from MSA norms, thus rendering MSA-trained tokenizers, morphological analyzers, and parsers especially brittle when applied to Palestinian Arabic material.

Orthographic instability in Arabic presents a second structural barrier for computational modeling, one that becomes even more acute in the context of Palestinian Arabic and other dialects. Arabic exhibits highly variable orthographic practices, especially in informal or dialectal contexts. Arabic NLP remains hindered by "complex morphology, complicated orthography and diacritics, [and] ambiguity of significance in context and words," problems that persist even with modern deep-learning methods (Alayba 2025, 2). This ambiguity is amplified in dialectal writing, where no standardized spelling system exists: dialects are often written ad hoc, mixing Arabic script with Romanization, employing locally specific spellings, and omitting diacritics essential for disambiguation (Boumaraf et al. 2022, 1). For Palestinian Arabic in particular, orthographic variation interacts with code-switching, especially with English or other languages such as Hebrew (Makhoul 2022; Mkahal 2016). Corpora and resources for code-switched dialectal Arabic remain extremely limited, reinforcing the structural invisibility of dialects in computational systems in NLP research (Hamed et al. 2025, 4561).

AI systems trained on high-resource, standardized, Anglocentric data are structurally incapable of modeling the linguistic structures through which Palestinians narrate displacement and lived experience. When dialectal variation, orthographic instability, and code-switching push Palestinian Arabic outside the bounds of what NLP systems can parse, the result is a digital archive in which crucial testimonies become mistranscribed, mistranslated, and undiscoverable. This is a mechanism by which

computational infrastructures reproduce the archival silences produced by settler-colonial erasure. In this sense, the technical limitations of AI become inseparable from questions of archival justice.

Scholars across information science, critical data studies, and AI ethics increasingly use the term *algorithmic invisibility* to describe the ways computational systems render certain populations, languages, and forms of knowledge unreadable or absent within digital infrastructures. While the phrase itself circulates in a dispersed scholarly conversation rather than originating with a single author, it resonates closely with Ruha Benjamin's (2019, 9) argument that digital systems often encode forms of "invisible" punishment and harm, producing racialized asymmetries in visibility and legibility within sociotechnical systems. Within archival studies, the concept parallels analyses of archival silence (see Caswell 2014; Trouillot 2015), epistemic erasure (Smith 2022), and "data voids" (Golebiewski and Boyd 2018). Scholars also examine the mechanisms through which (in)visibility is operationalized in computational and digital systems. Platform-ranking and discovery systems actively produce algorithmic visibility for some materials while relegating others to obscurity (Gillespie 2014), and large-scale AI infrastructures generate informational inequalities in which certain groups become systematically invisible to the "infosphere" (Floridi 2014).

The technical and linguistic asymmetries outlined above constitute a form of epistemic injustice within digital cultural-heritage infrastructures. Epistemic injustice describes conditions under which certain communities are structurally denied the ability to contribute knowledge or to be recognized as legitimate knowers (Fricker 1998). In the context of AI-driven archives, this injustice manifests in two mutually reinforcing ways. First is what Fricker calls *testimonial injustice*, which in her account occurs when a speaker's credibility is deflated due to identity prejudice. In AI-driven archival contexts, this dynamic is reproduced structurally rather than interpersonally: when automated tools cannot accurately transcribe, translate, or index Palestinian Arabic, the resulting errors function as credibility deficits, in that testimony becomes misrepresented, fragmented, or rendered inaccessible, losing evidentiary weight within digital discovery environments.

Second is what Fricker (1998) terms *hermeneutic injustice*, which arises when gaps in collective interpretive resources put someone at an unfair disadvantage. In the context of the POHA, the conceptual frameworks needed to interpret and make meaning of Palestinian historical experience are absent or severely constrained. Large-scale language models and metadata schemas trained and designed, respectively, on dominant linguistic and cultural corpora do not possess the semantic resources to capture the particularities of Palestinian historiography. This absence reflects the broader structural inequality of global knowledge infrastructures,

which concentrate computational resources, data, and technical expertise in institutions aligned with high-resource languages and dominant geopolitical interests. Communities already marginalized in political and archival contexts thus find their histories further occluded within AI-driven environments.

The Palestinian Oral History Archive

Oral history constitutes a particularly demanding test case for contemporary AI systems because oral testimony foregrounds precisely the features of language that computational pipelines are least equipped to handle. Unlike textual records, oral histories emerge from dialogic encounters shaped by memory, affect, trauma, intergenerational transmission, and culturally specific narrative forms. They rely on social, linguistic, geographic, and political contexts to convey meaning. Automated systems built around decontextualized linguistic units, standardized vocabularies, or English-centric semantic frames lack the interpretive resources needed to register these layers of significance. As a result, oral histories are especially vulnerable to forms of algorithmic misreading that flatten or fragment the nuance essential to understanding displacement and dispossession.

The Palestinian Oral History Archive offers a critical lens through which to examine how algorithmic invisibility operates within digital cultural-heritage infrastructures. Developed through a partnership between the American University of Beirut Libraries, community organizations, and regional collaborators, POHA aims to preserve, document, and circulate the testimonies of Palestinians displaced in 1948 and their descendants. POHA was designed to "preserve and consolidate the wider region's threatened oral history and intangible heritage" and to provide "reliable long-term access" to materials that exist nowhere else except in the memories of the displaced and their heirs (Sleiman and Chebaro 2018, 73). Within a cultural landscape marked by archival erasure, confiscation, fragmentation, and political suppression, POHA functions as both a repository of endangered testimony and an intervention against ongoing attempts to obscure or overwrite Palestinian historical experience.

POHA's scope reflects both documentary needs and political realities. The archive includes hundreds of recorded testimonies from survivors of the Nakba and their descendants. Its founding philosophy is explicitly community-centered: interviews were conducted in partnership with local researchers and community experts who understood the linguistic and cultural nuances embedded in the narratives. Rather than pursuing a homogenizing, positivist approach to oral testimony, POHA emphasizes what it calls the "integrity, immediacy, and oral nature of the material at hand," acknowledging that oral histories carry forms of knowledge that exceed textual representation (Sleiman and Chebaro 2018, 65).

One of the central challenges posed by oral testimony is that such testimony is often nonlinear rather than expository. That is, narrators often move back and forth across time, return to earlier events, digress, and layer recollection with reflection about their experiences. Consider the following example. In response to the question "What were the subjects they [i.e., schoolteachers] used to teach?," Ibrahim Mahmoud Blaybil, a Palestinian born in 1920 in Taytabah, recounts not a list of school subjects but a sequence of memories from his childhood:

> And during springtime as well, we would help our parents plant the tomatoes, the zucchini, the wild cucumbers, the okra… People used to own sheep, goats, and they had to take care of them, cows as well. I used to hate it. When I got to fourth or fifth grade, we went to Safad. There were a large number of students from my town, as well as yours. From the Nayif and al-Zaghmut families, as well as the Hamad brothers, Qasim Hamad, Ibrahim Hamad and ʿAli Hamad. No, they were cousins. Two of them were brothers and one of them was a cousin. They were from Syria. We had rented a place in Safad and that is where we studied for the fifth, sixth and seventh grade. (2004)

Blaybil situates agricultural work, family obligations, and migration to Safad not as digressions from education but as constitutive conditions of it. The narrative moves fluidly between seasonal labor, emotional response ("I used to hate it"), family relations ("brothers … [n]o, they were cousins"), and geographic relocation. From the perspective of oral history, this digression is a form of contextualization that conveys how education was embedded in everyday village life. From the perspective of AI-driven systems, however, such nonlinear narration poses a significant challenge. Automated tools trained to extract topics or classify segments into predefined categories, for instance, are likely to treat these shifts as noise rather than as meaning-bearing structures. As a result, the very features of oral testimony that make it historically rich are at risk of being flattened or fragmented when subjected to computational processing.

Meaning in oral testimony is transmitted not only through spoken content but through affect, hesitation, repetition, emphasis, and silence, as well as through gestures and interactions (verbal and gestural) among multiple speakers. These dimensions of meaning resist reduction to stable textual units. Oral testimony is also fundamentally co-constructed; it emerges through an interaction between narrator and interviewer, shaped by prompts, clarifications, shared assumptions, and moments of mutual recognition that guide what is said, how it is said, and when it becomes sayable.

Consider another example. The excerpt below is from an interview conducted in 2003 in the Ayn al-Hilweh refugee camp in Sayda by Mahmoud Zeidan with Jabr Muhammad Yunis and Khalidiya Muhammad Yunis. As Zeidan presses for clarification about what occurred when the Syrian

army[2] entered Jabr and Khalidiya's village, the exchange unfolds through repetition and gradual specification rather than through a single declarative statement. Jabr initially echoes the interviewer's question, seeking to establish what is being asked, before naming the event ("They invaded it") and elaborating on its consequences. The account reaches its strongest evidentiary force when Jabr points to his body to show his scars. At the same moment, Khalidiya supplements the narrative with additional detail. At one point, Jabr and Khalidiya speak at the same time. Gestures and corroboration across speakers function as integral components of testimony:

ZEIDAN: When they entered the village, were the inhabitants still inside or—
JABR: Yes. Yes. Yes—in the village.
ZEIDAN: What did they do to the people who were in the village?
JABR: What did they do to the people?
ZEIDAN: Yes.
JABR: What do you mean, what did they do to the people?
…
JABR: They invaded it. When they invaded it, they entered people's homes and picked out the men and lined them up and fired shots at them. That's what happened. I was one of them. I'm not lying, I was one of them.
ZEIDAN: I see.
JABR: Here—two bullets (he points to his leg). They went in through here and came out through there.
ZEIDAN: Let me see, can you raise your hand a little?
JABR: (He raises his elbow to reveal a deep scar).
KHALIDIYA: He fell hard on his head.
ZEIDAN: The shot in your leg, is it visible? Can you raise your trousers?
JABR: I'm not sure they go high enough.
KHALIDIYA: Here it is. Look, this is where the bullets entered. Let him see (she points to her husband's knee). Here it is, look. There … It's huge.
JABR: Here it is.
KHALIDIYA: They hit him here and went up his leg, there.
JABR: They got stuck and my sister took them out with a needle.
ZEIDAN: How many bullets?
JABR and KHALIDIYA: Two.
KHALIDIYA: And another one brushed his side.
(Maḥmūd Yūnus and Muḥammad Yūnus 2003)

In the context of the Nakba, oral history is both a documentary method and a political practice. Palestinian oral histories contain dense clusters of place-names, kinship structures, agricultural terms, and village-specific idioms and other forms of knowledge that resist straightforward transcription or translation or whose meaning is inseparable from the landscapes and communities violently disrupted in 1948 and after. These

testimonies also show the marks of generational rupture: trauma recollections, code-switching shaped by exile, and narrative strategies developed in response to silencing, censorship, or the loss of physical places (Lamar et al. 2025). Automated systems trained on high-resource languages and standardized corpora cannot reliably interpret or categorize these features.

In contexts of settler-colonialism and forced displacement, mistranslation or decontextualization is a form of harm that risks reinscribing the very epistemic asymmetries oral history was created to counter. Palestinian oral histories, which are often the sole surviving records of villages that no longer exist on contemporary maps, depend on accurate rendering of linguistic and cultural detail for their evidentiary force. When computational pipelines mistranscribe a village name, mistranslate an agricultural term, or strip affective markers from speech, they distort a historical record that has already survived multiple attempts at erasure. Within digital infrastructures, these errors then quickly propagate across search systems and other downstream applications, directly impacting what researchers, educators, journalists, and community members can find and how they can make sense of it.

The digital turn thus introduces a paradox. Digitization has made Palestinian oral histories more widely accessible than ever before, yet the computational systems that now mediate their discovery (and increasingly, their interpretation) risk adding new layers of erasure. Ensuring that oral histories of forced migration remain intelligible, credible, and contextually grounded requires tools and frameworks that can honor the specificities of the narrative forms through which displaced communities articulate their histories. Without such context-sensitive approaches, the digital preservation of oral testimony may unintentionally deliver these records into infrastructures that cannot recognize what they contain, rendering them vulnerable to renewed forms of digital displacement and invisibility.

On the level of interface design, POHA offers multiple pathways for discovery: free-text search, faceted browsing, interview summaries, controlled vocabularies, geographic filters, and thematic descriptors. Users may navigate by topic (e.g., agriculture, displacement, resistance), by location (e.g., villages and towns depopulated during or after 1948), or by biographical metadata about interviewees. The platform's bilingual structure (Arabic and English) widens accessibility and allows researchers who do not read Arabic to engage with the collection at a basic descriptive level. POHA's descriptive framework draws on standard metadata structures while layering onto them custom controlled vocabularies tailored to oral history materials (including a bilingual thesaurus and subject headings developed specifically to capture Palestinian historical categories), helping prevent the imposition of interpretive frameworks drawn from

generic library science rather than from the communities represented in the archive (Sleiman and Chebaro 2018).

Yet these strengths coexist with structural constraints imposed by the wider computational environment in which POHA must operate. As with most digital collections, POHA's search engine and metadata are built on top of conventional schema standards and indexing systems that assume textual regularity, semantic stability, and standardized vocabularies. The video testimonies themselves resist these assumptions: they exhibit dialectal variability, code-switching, recurrent references to erased geographies, and culturally specific idioms in ways that current automated tools simply cannot accommodate. Transcribed village names appear in multiple historical spellings; kinship terms have layered meanings that exceed one-to-one English equivalents; agricultural descriptions reference practices that fall outside the ontologies embedded in mainstream NLP pipelines. The bilingual metadata that POHA carefully constructs to enable access for non-Arabic-speaking users must therefore navigate a constant tension between intelligibility in English and fidelity to Palestinian categories of meaning. In this sense, POHA shows both the possibilities and the limits of digital access: the platform actively democratizes entry points, even as it remains constrained by broader infrastructural assumptions about language.

POHA's metadata creation process offers some of the clearest evidence of how algorithmic invisibility materializes in practice precisely because the project refuses to compromise on linguistic and contextual fidelity. The archive's creators emphasize that existing cataloging, indexing, and descriptive tools were fundamentally inadequate for the linguistic and narrative demands of the collection. As they note, the thesauri, indices, and subject headings available in the field "were developed primarily for written material and official political histories rather than for oral material narrated by first-person eyewitnesses," and transcription itself "risks transforming an audiovisual medium into a two-dimensional textual one that flattens the interviewees' contexts and layered meanings" (Sleiman and Chebaro 2018, 70, 69). POHA thus adopted a deliberately labor-intensive, human-centered methodology: developing a bilingual thesaurus, creating "custom-tailored subject headings and keywords that remained faithful to the language used by the narrators," and manually expanding tag lists to capture local place-names, kinship terms, and culturally specific concepts absent from any existing controlled Arabic vocabulary (Sleiman and Chebaro 2018, 70).

POHA demonstrates how linguistic inequity in AI manifests within cultural-heritage infrastructures. Far from being an example of technical failure, POHA is an exemplary digital archive. It is precisely this excellence that exposes where contemporary AI systems fall short. POHA's

practices function as an index of AI's linguistic hierarchies: They show how computational tools built around English and a handful of high-resource languages cannot accommodate the dialectal variation, code-switching, historical specificity, and embodied knowledge characteristic of Palestinian testimony; this is why the archive's descriptive infrastructure required the kind of sustained, expert human labor that most institutions cannot sustain at scale. Even so, the computational systems through which users actually encounter the collection—for example, search engines, indexing layers, and discovery interfaces—remain calibrated to the assumptions of high-resource languages. Human-generated metadata that uses dialectal terms, historically specific place-names, or culturally particular categories may still fail to surface correctly within these systems or may be legible only to users who already know enough to search for it. POHA's practices thus function as an index of AI's linguistic hierarchies at two levels: They show what automated tools cannot produce, and they show what automated tools cannot yet adequately recognize. In this way, POHA illuminates the systemic conditions under which low-resource languages become computationally illegible. The case study thus exemplifies this article's broader claim that AI technologies, as currently constituted, reproduce global patterns of linguistic and epistemic inequality, rendering displaced communities newly vulnerable within digital archival systems.

The Political Economy of AI's Linguistic Inequities, or, The Structural Conditions of Linguistic Marginalization

The linguistic inequities that render Palestinian oral histories computationally illegible do not arise from isolated gaps in tooling. Even as cultural-heritage institutions increasingly rely on AI for discovery, transcription, translation, and metadata generation, the systems available to them are the product of an ecosystem in which low-resource languages, and especially the languages of displaced or colonized communities, have never been a priority. Understanding why AI technologies consistently fail to accommodate Palestinian Arabic therefore requires shifting focus from individual models or datasets to the political economy of AI itself: the market incentives that drive corporate multilingual systems, the geopolitical hierarchies that shape training data pipelines, the institutional constraints faced by libraries and archives, and the specific political conditions that render Palestinian linguistic data particularly difficult to obtain.

At the most basic level, the AI systems that cultural-heritage institutions can access—such as commercial ASR engines, machine translation services, large language models, and indexing tools—are developed by corporations whose priorities are shaped by market demand. This means that computational support is concentrated overwhelmingly in high-resource languages with large user bases and strong purchasing power. English remains the global default; languages such as French, Mandarin, and

Spanish benefit from state investment and broad commercial markets. Dialect-rich, nonstandardized, or politically marginalized languages, by contrast, rarely generate sufficient commercial incentive to be included in development pipelines.

Multilingual models nominally gesture toward inclusivity, but linguistic diversity is often treated as an engineering afterthought; additional languages are bolted onto systems whose architecture and training pipelines were optimized for English from the outset. Even when Arabic is included, it tends to be represented through Modern Standard Arabic or a narrow set of dialects for which corpora are readily available, for example, Egyptian, Gulf, and occasionally Lebanese or Moroccan. Commercial ASR and machine translation tools therefore default to the linguistic norms of high-resource markets, rendering low-resource dialects computationally invisible. Palestinian Arabic, shaped by displacement, multilingualism, and regionally specific histories of colonial violence, is almost never included. This dynamic is not unique to Palestinian Arabic; it is characteristic of commercial AI systems globally. But for archives working to preserve the oral histories of communities marginalized by war and occupation, these market forces have especially acute consequences: the languages most in need of infrastructural support are the ones least likely to receive it.

Beyond market demand, the structural imbalance in AI capabilities is also driven by what scholars have termed *data colonialism*: the concentration of computational power, data resources, and research infrastructure in the Global North (Habib Lantyer 2025; Mejias and Couldry 2024). Large-scale AI models depend on vast amounts of training data, most of which is harvested from English-dominant online environments or curated by institutions aligned with Western geopolitical interests. The pipelines that gather and preprocess data for machine learning prioritize the forms of language most easily scraped, digitized, and standardized. Spoken dialects, community-specific registers, code-switching practices, and culturally embedded terminologies—features central to Palestinian oral history—fall outside these pipelines by design. As a result, even multilingual models lack exposure to the linguistic realities of communities living under displacement or occupation. When these models encounter Palestinian Arabic, they attempt to force it into the representational space carved by high-resource languages, producing hallucinations, mistranslations, and semantic distortions.

Moreover, community-generated data, like the recordings, testimonies, and vernacular expressions preserved in archives like POHA, is seldom incorporated into training corpora. Ethical concerns around privacy, consent, and cultural sovereignty rightly limit the indiscriminate use of such materials. It is a natural and unfortunate consequence that the data ecosystems driving AI development thus remain structurally skewed toward dominant linguistic communities.

While the AI industry operates according to market incentives, libraries, archives, and cultural-heritage institutions often lack the computational or funding resources required to develop alternatives. Most memory institutions do not have the infrastructure to build bespoke ASR or machine translation models for low-resource languages; they depend instead on commercial vendors whose tools are optimized for general-purpose markets. Even when archivists possess linguistic expertise or deep knowledge of community-specific contexts, translating that expertise into computational models requires collaboration with NLP specialists, access to GPUs or research computing clusters, and sustained funding for model training, evaluation, and maintenance.

Furthermore, many cultural-heritage institutions rely on vendor-provided platforms like digital asset management systems or integrated library systems. These platforms often offer minimal transparency about their underlying models. When these systems fail to recognize dialectal Arabic, misclassify metadata, or tokenize Palestinian place-names incorrectly, archivists may have little recourse. They cannot adjust the training data or retrain the model. Archives bear responsibility for preserving marginalized languages and histories, yet the computational systems they depend on remain largely outside their control.

This capacity gap is even more pronounced for community-led or underfunded projects, common conditions under which low-resource languages are researched. This includes many Palestinian archives operating in diaspora or under conditions of political precarity. Even where there is expertise in oral history, ethnography, or linguistic documentation, there may be insufficient institutional infrastructure to support computational innovation. As a result, human labor remains the backbone of archival work.

Finally, the specific political conditions surrounding Palestinian linguistic and cultural data create additional barriers not fully captured by market or infrastructural explanations. Palestinian Arabic is shaped by decades of displacement, multilingual contact, and military occupation. These are conditions that disrupt both the production and the circulation of linguistic data. Many villages referenced in oral histories no longer exist on contemporary maps; place-names appear in colonial, Ottoman, Mandate-era, and vernacular forms; kinship structures have been altered by forced migration; and code-switching practices have emerged from encounters with Hebrew, English, and regional Arabic dialects.

Moreover, the collection of Palestinian linguistic data is complicated by censorship, surveillance, and geopolitical restrictions. Researchers working in the West Bank, Gaza, or refugee camps in Lebanon face material and political barriers to fieldwork. Palestinian institutions that hold linguistic or oral history materials may be subject to threats or resource constraints. Digitizing, storing, and sharing such data often involves navigating

politically sensitive terrain, where questions of intellectual property, privacy, and community safety intersect with concerns about surveillance and misinformation. Under such conditions, building large, open-access corpora suitable for training AI models becomes exceptionally difficult.

These political constraints help explain why Palestinian Arabic is underrepresented not only in commercial AI systems but also in open-source NLP initiatives. The data scarcity is not merely technical but the result of historical and ongoing forms of violence that shape what linguistic materials can be collected and digitized in the first place. Addressing the inequities in cultural-heritage AI therefore requires not only technical solutions but a reorientation of the political and economic structures that determine which languages are computationally legible. Low-resource languages demand new models of infrastructural investment and linguistic justice.

Consequences for Cultural-Heritage Work

For communities whose histories have long been suppressed, misrepresented, or destroyed, preservation via digitization is both a technical act and a political one. When AI systems cannot reliably process the languages and narrative forms through which displaced people articulate their histories, the result is a renewed form of archival vulnerability.

The most immediate consequence of AI's linguistic hierarchies is the reproduction of digital silence. Oral histories that cannot be accurately transcribed, translated, or indexed become difficult to find within digital repositories. When search mechanisms depend on English keywords or MSA-trained tokenizers, testimonies in Palestinian Arabic may surface only partially or not at all. This mirrors and extends the long-standing archival silences produced by settler-colonial erasure: even when communities succeed in preserving their histories, the computational infrastructures meant to increase access may inadvertently render those histories invisible. In this sense, algorithmic invisibility is a new modality of archival silence.

AI's limitations also narrow the interpretive horizons of scholars, educators, journalists, and community members who rely on digital discovery systems. Researchers without Arabic language expertise may only access interviews that have been manually translated or tagged, creating a skewed or partial picture of the archive. Educators hoping to integrate oral histories into classrooms may find that search tools fail to surface relevant testimonies because dialectal terms are unrecognized or indexing relies on limited English metadata.

Digital humanities and digital memory projects, many of which rely on automated metadata extraction, topic modeling, or other computational techniques, also struggle to incorporate low-resource languages. Because these systems cannot reliably process dialectal Arabic or culturally specific

terminology, Palestinian oral histories are often left out of public-facing digital exhibits or appear only in partial, decontextualized form. The same limitations constrain projects that seek to visualize or analyze geographic data tied to villages destroyed or depopulated in 1948, since place-names rendered inconsistently or misrecognized by automated tools cannot be geotagged with accuracy. These constraints in turn hinder the development of participatory or community-curated digital memory projects, where descendants and community members might otherwise contribute contextual knowledge. As a result, the computational gaps in multilingual AI narrow the possibilities for public scholarship and collective memory work.

When AI systems cannot support the linguistic complexity of Palestinian oral histories, interpretive authority shifts toward intermediaries such as translators, platform administrators, external vendors, or English-speaking scholars. This outsourcing creates two risks. First, there is a risk of a loss of interpretive sovereignty for the communities whose histories are being preserved. Second, outsourcing creates dependence on actors who may lack cultural, linguistic, or political expertise. Even well-intentioned intermediaries may inadvertently reproduce biases embedded in platform defaults, English-language metadata, or general-purpose AI tools. Addressing AI's linguistic inequities within cultural-heritage work requires rethinking both technical development and institutional practice.

Low-resource and community-specific AI cannot be built without the communities themselves. Involving Palestinian linguists, historians, archivists, and community elders in the design, training, and evaluation of models ensures that computational tools reflect lived linguistic knowledge. Community participation is particularly crucial in defining meaningful categories of metadata, identifying culturally salient terminology, setting accuracy thresholds appropriate to oral history contexts, and overseeing ethical use of community-generated data.

A growing body of research in natural language processing suggests several promising pathways for developing AI systems capable of handling dialect-rich, low-resource languages in more equitable and contextually sensitive ways. Dialect-specific ASR models, trained on regionally grounded corpora rather than generalized Modern Standard Arabic, have shown measurable improvements in accuracy for Levantine and North African dialects when sufficient community-collected data is available (Ardila et al. 2020; Hanani et al. 2016). Approaches such as federated learning further enable model training on sensitive or community-held datasets without requiring centralized aggregation, an important consideration for archives whose materials contain politically or personally sensitive testimony (McMahan et al. 2017). Finally, hybrid human-machine workflows, in which automated tools are used to support, rather than replace, human expertise, are increasingly recognized as essential for oral history

contexts, where affect or gesture, which may not be machine-readable, can be meaningfully scaffolded through semiautomated transcription, alignment, or metadata suggestion (Cherukuri et al. 2025). Instead of optimizing for high-volume, English-first generalization, these strategies prioritize adaptability and contextual fidelity.

Because Palestinian oral histories contain sensitive political and genealogical information, AI development must be guided by robust ethical frameworks. Central among these is data sovereignty, which asserts the authority of displaced and stateless communities to govern how their cultural materials are collected, stored, processed, and circulated, particularly in contexts where external institutions have historically controlled Palestinian records. Equally important are trauma-informed metadata practices that acknowledge how testimony often conveys profound grief, loss, dispossession, and memory under conditions of surveillance or censorship; such practices require descriptive systems that do not flatten emotional nuance or expose narrators to new forms of risk. Finally, consent frameworks aligned with cultural norms and community expectations are essential for oral histories. Together, these ethical commitments ensure that computational tools function as extensions of community-led archival care rather than as extractive mechanisms, enhancing rather than undermining the integrity of the histories these archives seek to preserve.

No single institution—whether a library, archive, or university lab—can build the kind of dialect-sensitive AI required to support languages like Palestinian Arabic. Meaningful progress will depend on collaborative infrastructures that redistribute expertise, resources, and decision-making power. This includes library and archive consortia that can pool funding and share corpora; cross-university research clusters that bring together linguists, computer scientists, oral historians, and community experts; and sustained partnerships with open-source NLP communities, especially those committed to linguistic justice for low-resource languages. Initiatives such as Masakhane, which has transformed NLP for African languages through community-driven, continent-spanning collaboration, demonstrate the potential of transnational, grassroots research networks to build models grounded in local linguistic knowledge (Adelani et al. 2022). Adopting similar structures for Palestinian Arabic would allow cultural-heritage institutions to move beyond isolated, underresourced experimentation and toward a coordinated, community-led ecosystem capable of supporting genuine linguistic equity.

Toward Linguistic Justice in Cultural-Heritage AI

The analysis above demonstrates that the challenges facing Palestinian oral history collections are not anomalous, nor are they resolvable through incremental technical improvements. They reflect structural inequities in how linguistic diversity is (not) supported within contemporary AI

infrastructures. Cultural heritage in crisis must therefore be understood as encompassing both material vulnerability and algorithmic exclusion: the risk that marginalized communities' historical records become illegible, misinterpreted, or absent within computational systems that increasingly mediate access to knowledge. The neglect of low-resource and dialectal languages functions as a form of ongoing epistemic violence.

Addressing these inequities requires a shift from ad hoc technical adaptation to a comprehensive rethinking of how AI systems are designed and evaluated. A future oriented toward linguistic justice would involve multilingual, culturally informed, and community-governed approaches to model development. Such approaches are necessarily grounded in interdisciplinary collaboration among linguists, computer scientists, oral historians, archivists, and the communities represented in the archives.

Institutions should prioritize sustained investment in tools and workflows designed for low-resource languages rather than relying on generic, English-centric platforms. Funding structures should support long-term capacity-building, including the development of regionally grounded corpora, maintenance of language models, and staff positions dedicated to computational support. Short-cycle digitization grants are insufficient without corresponding infrastructural commitments.

Institutional governance frameworks should explicitly address linguistic equity in the design, deployment, and evaluation of AI tools. This includes requirements for transparency around training data and model assumptions, careful assessment of potential harms associated with mistranslation or misclassification, and guidelines that prioritize open, auditable systems over opaque proprietary solutions. Policies should also incorporate principles of data sovereignty; that is, policies should prioritize ensuring that displaced communities retain agency over how their linguistic and cultural materials are used in model development.

AI systems require ongoing care analogous to the stewardship of physical collections. Libraries and archives should develop infrastructure plans that incorporate versioning, retraining schedules, documentation practices, and mechanisms for auditing model outputs. Avoiding "digitize and forget" outcomes requires recognizing that language models and datasets are themselves cultural-heritage assets whose maintenance bears directly on future accessibility and interpretability.

Ensuring meaningful access to collections such as the Palestinian Oral History Archive is a question of archival justice, particularly in the context of Palestine, where decades of dispossession and archival suppression have made oral testimony a primary vehicle for historical continuity and cultural survival. As AI systems become increasingly embedded in the workflows of libraries and memory institutions like museums, they will play a central role in determining which voices become legible within digital environments.

Libraries, archives, and cultural-heritage organizations therefore have a critical responsibility to intervene in the design and governance of AI systems whose default assumptions render Palestinian histories structurally illegible. By prioritizing linguistically inclusive technologies, institutionalizing governance frameworks grounded in community accountability, and cultivating interdisciplinary collaborations that center Palestinian knowledge holders, these institutions can help ensure that digital infrastructures do not replicate the erasures produced through decades of dispossession and archival suppression. Instead, they can contribute to the development of computational systems capable of sustaining the narrative agency of displaced communities and the historical specificity of Palestinian experience.

This analysis has focused on a single, exemplary case. POHA is a well-resourced, institutionally supported archive; the constraints it faces are likely to be even more acute for community-led projects operating with fewer staff and less infrastructure. The argument developed here is therefore necessarily preliminary: it identifies structural patterns rather than offering a complete account of all the ways in which AI's linguistic hierarchies manifest across different archival contexts. More empirical work is needed, particularly comparative studies that trace AI's effects across archives serving different low-resource language communities. Archivists and institutions looking for ways to act on these commitments may begin by composing a record of this failure itself. Documenting the contexts in which automated tools fail generates the evidentiary base needed both for future model development and for institutional arguments that linguistically equitable AI requires sustained investment. That work also creates pressure points for more immediate interventions, such as demanding transparency about dialect support before adopting vendor tools and building those requirements into contracts rather than assuming them after the fact.

Notes

1. "Meaning 'catastrophe' in Arabic, the term 'al-Nakba' (النكبة) is often used—as a proper noun, with a definite article—to refer to the ruinous establishment of Israel in Palestine, a chronicle of partition, conquest, and ethnic cleansing that forcibly displaced more than 750,000 Palestinians from their ancestral homes and depopulated hundreds of Palestinian villages between late 1947 and early 1949" (Eghbariah 2024). See Sa'di and Abu-Lughod 2007; United Nations, n.d.
2. Jabr specifies in an earlier portion of the interview that the army who came to the village was "the Syrian army. They sent Jaysh al-Inqadh [the Arab liberation Army]" (brackets in the original).

References

Adelani, David Ifeoluwa, Graham Neubig, Sebastian Ruder, et al. 2022. "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition." arXiv:2210.12391. arXiv, November 15. https://doi.org/10.48550/arXiv.2210.12391.

Alayba, Abdulaziz M. 2025. "Arabic Natural Language Processing (NLP): A Comprehensive

Review of Challenges, Techniques, and Emerging Trends." *Computers* 14 (11): 497. https://doi.org/10.3390/computers14110497.

Allan, Diana, ed. 2021. *Voices of the Nakba: A Living History of Palestine*. Pluto Press.

Allan, Diana K. 2007. "The Role of Oral History in Archiving the Nakba." *Al-Majdal* 32. https://badil.org/publications/al-majdal/issues/items/1173.html.

Anziska, Seth. 2019. "Special Document File: The Erasure of The Nakba in Israel's Archives." *Journal of Palestine Studies* 49 (1): 64–76. https://doi.org/10.1525/jps.2019.49.1.64.

Ardila, Rosana, Megan Branson, Kelly Davis, et al. 2020. "Common Voice: A Massively-Multilingual Speech Corpus." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, et al. European Language Resources Association. https://aclanthology.org/2020.lrec-1.520/.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Blasi, Damian, Antonios Anastasopoulos, and Graham Neubig. 2022. "Systematic Inequalities in Language Technology Performance Across the World's Languages." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376.

Blaybil, Ibrāhīm Maḥmūd. 2004. "Interview with Ibrāhīm Maḥmūd Blaybil." Interview by Maḥmūd Zaydān, February 7. Video, 1:23:21. Al-Nakba Collection, Palestinian Oral History Archive. https://n2t.net/ark:/86073/b39p6c.

Bouamor, Houda, Nizar Habash, and Kemal Oflazer. 2014. "A Multidialectal Parallel Corpus of Arabic." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, et al. European Language Resources Association. https://aclanthology.org/L14-1435/.

Boumaraf, Assia, Sonia Bekal, and Joël Macoir. 2022. "The Orthographic Ambiguity of the Arabic Graphic System: Evidence from a Case of Central Agraphia Affecting the Two Routes of Spelling." *Behavioural Neurology* 2022:8078607. https://doi.org/10.1155/2022/8078607.

Caswell, Michelle. 2014. "Toward a Survivor-Centered Approach to Records Documenting Human Rights Abuse: Lessons from Community Archives." *Archival Science* 14 (3): 307–22. https://doi.org/10.1007/s10502-014-9220-6.

Cherukuri, Komala Subramanyam, Pranav Abishai Moses, Aisa Sakata, Jiangping Chen, and Haihua Chen. 2025. "Large Language Models for Oral History Understanding with Text Classification and Sentiment Analysis." arXiv:2508.06729. arXiv, August 8. https://doi.org/10.48550/arXiv.2508.06729.

Couldry, Nick, and Ulises Ali Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Culture and Economic Life. Stanford University Press.

Eghbariah, Rabea. 2024. "Toward Nakba as a Legal Concept." *Columbia Law Review* 124 (4): 887–992.

Floridi, Luciano. 2014. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press.

Fricker, Miranda. 1998. "IX—Rational Authority and Social Power: Towards a Truly Social Epistemology." *Proceedings of the Aristotelian Society* 98 (1): 159–78. https://doi.org/10.1111/1467-9264.00030.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009.

Golebiewski, Michael, and Danah Boyd. 2018. "Data Voids: Where Missing Data Can Easily Be Exploited." *Data & Society*, May.

Habash, Nizar Y. 2010. "What Is 'Arabic'?" In *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer.

Habib Lantyer, Victor. 2025. "Data Colonialism: The Geopolitics of Information." SSRN Scholarly Paper no. 5236304. Social Science Research Network, April 29. https://doi.org/10.2139/ssrn.5236304.

Hamed, Injy, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. "A Survey of Code-Switched Arabic NLP: Progress, Challenges, and Future Directions." In *Proceedings of the 31st International Conference on Computational Linguistics*,

edited by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.307/.

Hanani, Abualsoud, Aziz Qaroush, and Stephen Taylor. 2016. "Classifying ASR Transcriptions According to Arabic Dialect." In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, edited by Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi. The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4817/.

Jarrar, Mustafa, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. "Building a Corpus for Palestinian Arabic: A Preliminary Study." In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, edited by Nizar Habash and Stephan Vogel. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3603.

Jarrar, Mustafa, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. "Curras: An Annotated Corpus for the Palestinian Arabic Dialect." *Language Resources and Evaluation* 51 (3): 745–75. https://doi.org/10.1007/s10579-016-9370-7.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560.

Lamar, Annie K., Rick Castle, Carissa Chappell, et al. 2025. "Cognitive Geographies of Catastrophe Narratives: Georeferenced Interview Transcriptions as Language Resource for Models of Forced Displacement." In *Proceedings of the First International Workshop on Nakba Narratives as Language Resources*, edited by Mustafa Jarrar, Nizar Habash, Mo El-Haj, et al. Association for Computational Linguistics. https://aclanthology.org/2025.nakbanlp-1.3/.

Lamin, Nor Zakiah, and Azwa Abdul Aziz. 2025. "Cross-Lingual Sentiment Analysis in Low-Resource Languages: A Recent Review on Tasks, Methods and Challenges." *International Journal of Advanced Computer Science and Applications* 16 (11). http://dx.doi.org/10.14569/IJACSA.2025.0161144.

Maḥmūd Yūnus, Khālidīyah, and Jabr Muḥammad Yūnus. 2003. "Interview with Khālidīyah Maḥmūd Yūnus and Jabr Muḥammad Yūnus." Interview by Maḥmūd Zaydān. Video, 1:06:46. Al-Nakba Collection, Palestinian Oral History Archive. https://n2t.net/ark:/86073/b33324.

Makhoul, Ward. 2022. "Code-Switching Practices in Palestinian Arabic." Master's thesis, Theoretical and Applied Linguistics, Universitat Pompeu Fabra. https://www.recercat.cat/handle/10230/54106?locale-attribute=es.

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. "Communication-Efficient Learning of Deep Networks from Decentralized Data." *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017* 54. https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf.

Mejias, Ulises Ali, and Nick Couldry. 2024. *Data Grab: The New Colonialism of Big Tech and How to Fight Back*. University of Chicago Press.

Mkahal, Iyad Ahmad Hamdan. 2016. "Code Switching as a Linguistic Phenomenon Among Palestinian English Arabic Bilinguals with Reference to Translation." Master's thesis, Applied Linguistics and Translation, An-Najah National University.

Moon, Roxy. 2023. "Outside the Locus of Control: Palestinian Digital Archives Resist Israeli Settler-Colonial Erasure." *Journal of Palestine Studies* 52 (4): 37–52. https://doi.org/10.1080/0377919X.2024.2304322.

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, et al. 2020. "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages." In *Findings of the Association for Computational Linguistics: EMNLP 2020*, edited by Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195.

Sa'di, Aḥmad, and Lila Abu-Lughod, eds. 2007. *Nakba: Palestine, 1948, and the Claims of Memory*. Cultures of History. Columbia University Press.

Sharma, Nikhil, Kenton Murray, and Ziang Xiao. 2025. "Faux Polyglot: A Study on Information Disparity in Multilingual Large Language Models." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by Luis Chiruzzo, Alan Ritter, and

Lu Wang. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.411.
Sleiman, Hana, and Kaoukab Chebaro. 2018. "Narrating Palestine: The Palestinian Oral History Archive Project." *Journal of Palestine Studies* 47 (2): 63–76. https://doi.org/10.1525/jps.2018.47.2.63.
Smith, Linda Tuhiwai. 2022. *Decolonizing Methodologies: Research and Indigenous Peoples*. 3rd ed. Bloomsbury Academic.
Tiedemann, Jörg. 2020. "The Tatoeba Translation Challenge—Realistic Data Sets for Low Resource and Multilingual MT." In *Proceedings of the Fifth Conference on Machine Translation*, edited by Loïc Barrault, Ondřej Bojar, Fethi Bougares, et al. Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.139/.
Trouillot, Michel-Rolph. 2015. *Silencing the Past: Power and the Production of History*. Beacon Press.
United Nations. n.d. "About the Nakba." The Question of Palestine. Accessed December 10, 2025. https://www.un.org/unispal/about-the-nakba/.
VanSnick, Sarah, and Kostos Ntanos. 2018. "On Digitisation as a Preservation Measure." *Studies in Conservation* 63 (Suppl. 1): 282–87. https://doi.org/10.1080/00393630.2018.1504451.

Annie K. Lamar is an assistant professor of classics and, by courtesy, of linguistics at the University of California, Santa Barbara. She is also the director of the Low Resource Language (LOREL) Lab.