Addressing Historical and Contemporary Inequalities in Low-Resource Languages through the Lens of Serbian

Smiljana Antonijević Ubois
Ainthropology
smiljana@ainthropology.com

Abstract

The development of large language models prioritizes high-resource languages such as English, while most of the world's 7,000 languages remain underrepresented in AI. Multilingual pre-training and cross-lingual transfer have produced uneven results, favoring languages structurally closer to English and transferring linguistic and cultural biases from the training data into low-resource language (LRL) outputs, creating a double disadvantage in which LRLs are simultaneously underrepresented and misrepresented. This paper argues that developing language technologies for LRLs requires not only technical solutions, but also contextualized ethnographic inquiry into the historical, sociocultural, and institutional conditions that shape their digital future. Using Serbian as a case study, it draws on semi-structured interviews with ten scholars and NLP practitioners working across disciplines. Findings identify barriers common to many LRLs, including limited corpora, insufficient digitization, restrictive copyright regimes, and partial institutional support, which contribute to reliance on low-quality web data and multilingual models that reproduce bias. The paper outlines an alternative approach centered on native language models, interdisciplinary collaboration, inclusive datasets, and transparent data practices, with sustained involvement of humanities and social science scholars across data selection, evaluation, and governance, so that sociocultural and ethical concerns are built into language technologies from the outset rather than treated as post hoc technical interventions or matters of regulatory compliance.

1 Introduction

As we sit down in her office for an interview, mathematics professor Jelena[1] recounts how in 2023 her research team finally secured funding to develop Serbian language models and other language-specific tools from the national government's artificial intelligence (AI) research program. In 2021 and 2022, her team ranked among the top candidates but did not receive funding. "The peer reviewers' explanations were unbelievable to me," Jelena recalls, quoting feedback they received: 'Serbian is a small language, everyone works in English, so why develop language resources for Serbian?'" She notes that the grant competition included dual international and domestic review, and while international reviewers evaluated their application highly, the national committee did not. "To me that was precisely an indication that the need for developing language technologies for Serbian is not recognized. Our national mindset needs to be adjusted so that we understand that it is one of the country's infrastructural projects. Just as we need electricity and internet, we need language technologies," she concludes.

Jelena's experience reflects a broader pattern identified among researchers in this study, who describe the lack of strategic support for advancing Serbia's linguistic digital infrastructure, a

[1] Respondents' names have been replaced with pseudonyms to ensure their anonymity.

situation typical of low-resource languages (LRLs). LRLs "lack the quantity of data necessary for training statistical and machine learning tools and models" due to small speaker populations, limited digital corpora, scarce annotated datasets, or insufficient computational resources (Liu et al., 2022, p. 1). Most of the world's roughly 7,000 languages are LRLs, remaining technologically under-resourced and underrepresented in AI applications in ways that reproduce historical inequalities and limit their digital sustainability, accessibility, and global visibility (Lin et al., 2024; Joshi et al., 2020; Vargas-Parada, 2025). Although large language models (LLMs) are trained on vast datasets, major US-based models such as GPT and Llama have thus far focused predominantly on high-resource languages (HRLs) such as English, which accounts for approximately 90% of their training data, leaving LRLs significantly marginalized (Li et al., 2024).

This disparity has noticeable consequences. Performance of major LLMs sometimes reflects the cultural and linguistic biases encoded in source-language training data to the point that speakers of regional varieties of English feel unrepresented in AI tools (Fleisig et al., 2024). Furthermore, research shows that LLMs default to English in their internal reasoning and semantic processing, even when prompted in another language, and perform most accurately with Western knowledge and context (Schut et al., 2025). Recent initiatives have attempted to address these imbalances through multilingual pre-training and cross-lingual transfer, enabling models trained on HRLs to extend some capacity to LRLs (Hu et al., 2020). However, results remain uneven, favoring languages linguistically closer to the source language, which is often English (Joshi et al., 2020; Lauscher et al., 2020; Ponti et al., 2020).

An alternative approach involves developing native language models, sometimes called “sovereign language models” (Bondarenko et al., 2025; Barbereau & Dom, 2024), trained on data in a specific language to better reflect local linguistic structures, cultural frameworks, and communicative norms. Examples include GPT-NL for Dutch, ERNIE for Chinese, GPT-SW3 for Swedish, Fugaku-LLM for Japanese, BgGPT for Bulgarian, and Jais for Arabic, among others. Research shows that native models trained from scratch outperform multilingual alternatives for a given language, offering deeper semantic understanding and better handling of morphosyntactic complexity, dialectal variation, and domain-specific terminology (Virtanen et al., 2019). They also tend to “demonstrate a superior understanding of cultural nuances compared to general multilingual LLMs” (Mousi et al., 2024, p. 2). However, developing native models demands substantial high-quality training data and computational resources, precisely what LRLs lack.

This creates a fundamental tension about whether to prioritize immediate functionality through multilingual models or invest in the slower, resource-intensive work of building native infrastructure that more authentically represents linguistic and cultural specificity. For many LRLs, this choice is not merely technical but deeply political, involving questions of digital sovereignty, cultural preservation, and linguistic equality. As Khanna and Li (2025) point out, characterizing underrepresented languages as low resource primarily in terms of insufficient training data “obscures more fundamental lines of inquiry: why do some languages with millions of speakers possess virtually no digital footprint? Which historical, political, and structural dynamics give rise to this inequality?” (p. 2). Addressing these questions requires more than technical optimization and calls for contextualized ethnographic inquiry into the historical, sociocultural, economic, and technical factors that shape how each LRL relates to the contemporary AI landscape.

This paper takes such an approach through a case study of Serbian, a language characterized by rich dialectal diversity and the use of dual alphabetic systems, Cyrillic and Latin. Drawing on semi-structured interviews with ten scholars and practitioners working on Serbian natural language processing (NLP), this study examines both historical and contemporary barriers to developing language technologies for Serbian. Interviews were conducted between March–July 2025, predominantly online, and on average lasted one hour. All interviews were audio-recorded and then transcribed, and the resulting material was coded into relevant categories that were systematically developed and refined throughout the analysis. Participants were recruited through purposive snowball sampling within Serbia's relatively small NLP community to ensure diversity across institutional affiliations, educational backgrounds, and career stages. The sample comprised three mathematicians, two computational linguists, one librarian, one data scientist, one philosopher of

language, one software engineer, and one digital humanist. Most interviewees (8/10) were in the age group 35–50, and the majority (6/10) were women.

In what follows, the paper first outlines the historical context of the Serbian language and its current state of digital readiness. It then presents interview findings that highlight key barriers and opportunities in advancing Serbian language technologies, with particular attention to the importance of native models, corpus construction, institutional support, data bias, and interdisciplinary collaboration. Building on these insights, the conclusion identifies strategic areas for intervention in which proposed approaches move beyond post hoc technical fixes and become integral to the systematic development of technologies for low-resource languages.

2 Serbian historical context and digital readiness

Serbian is an Indo-European language that developed a rich medieval literary tradition, including seminal works such as Miroslav's Gospel (c. 1180), Vukan's Gospel (c. 1202), St. Sava's Nomocanon (1219), and Dušan's Code (1349). However, Ottoman occupation (mid 15th–early 19th centuries) devastated this tradition. Monastic libraries, central to manuscript production, faced destruction and looting, with manuscripts burned, stolen, or sold to collectors in Russia, Bavaria, or Venice. The collapse of urban centers, coupled with the dismantling of the Serbian royal court, eliminated patronage for literary and artistic works, while the church's ability to sustain schools and scriptoria became increasingly limited. Secular cultural production also almost entirely ceased, prompting a shift toward oral tradition as the primary means of preserving national identity and culture (Mihailovich & Mikasinovich, 2007). Written production persisted in the Serbian diaspora and at Hilandar Monastery on Mount Athos, laying groundwork for cultural revival (Deretić, 1987).

After liberation from Ottoman occupation (1804–15), Serbia gradually rebuilt its literary and cultural infrastructure, but its textual legacy continued to face setbacks. The Balkan Wars (1912–13) and World War I caused significant destruction, while the most catastrophic loss came in 1941, when German bombs obliterated the National Library of Serbia, destroying over 500,000 volumes, including 1,424 medieval Cyrillic manuscripts (Ristić, 2016).

These losses significantly diminished Serbian linguistic and literary resources, with lasting implications for contemporary language technologies. Literary tradition is not merely a cultural archive, but part of what constitutes language as a historically situated practice, since what survives in written form shapes which registers, styles, and discursive practices get transmitted, institutionalized, and eventually standardized. Destruction of textual heritage thus shapes not only what a culture remembers, but also which forms remain available for later study and, today, for computational modeling. In this sense, historical discontinuities in the written record are also discontinuities in the data on which language technologies are built.

Another historical development with contemporary relevance is Serbian linguistic modernization during the post-Ottoman cultural revival. Serbian philologist and ethnologist Vuk Karadžić led a comprehensive language reform in the early 19th century, transforming not only Serbian but also Croatian linguistic standardization. This influenced the future of Serbo-Croatian and the broader South Slavic linguistic landscape, with the Vienna Literary Agreement of 1850 declaring a shared language with two scripts, Cyrillic and Latin (Greenberg, 2004). Serbo-Croato-Slovene (later Serbo-Croatian) became Yugoslavia's official language after World War I, and subsequent Yugoslav constitutions maintained this status until the language fragmented into nationally framed variants following Yugoslavia's disintegration in the 1990s.

Today, Serbian is classified as a national pluricentric language, spoken across at least two nations where it serves as a national identity marker, enjoys state-sanctioned status with legally prescribed usage in official domains, and has sufficient speakers to establish its own codifying and standardizing linguistic center (Muhr et al., 2020). The ISO 639-3 code "hbs" (Serbo-Croatian) classifies it as a macrolanguage, grouping individual languages that share a common linguistic base but are treated as distinct for cultural, political, or national reasons (Eberhard et al., 2025). Serbian is currently spoken by about 11 million native speakers worldwide (UCLA, 2026), with hbs speakers estimated at 20 million (University of Cambridge Language Centre, n.d.). Serbian is the official

language of the Republic of Serbia and Republic of Srpska, co-official in Bosnia and Herzegovina, Montenegro, and Kosovo and Metohija under UN resolution 1244[2], and as an official minority language in Croatia, North Macedonia, Hungary, Romania, Slovakia, and the Czech Republic.

When it comes to digital readiness, data from 2024 show that 92.5% of urban and 85.5% of rural households in Serbia had a computer and broadband internet, and that 87.7% of citizens used the internet regularly (Statistical Office of the Republic of Serbia, 2025). Serbia's Information and Communication Technologies sector has significantly grown in its workforce and exports, especially in software development and AI (International Trade Administration, 2025). Since 2017, the Office for IT and e-Government has operated the country's e-Government portal and a National Open Data Portal connected to the European Union's open data platform. After adopting its initial National AI Strategy in 2020, Serbia established the Institute for AI Research and Development and a state Data Center with two supercomputers. In 2024, Serbia chaired the OECD Global Partnership on AI (GPAI) and hosted its inaugural annual summit. The new National AI Strategy (2025–2030), adopted in 2025, prioritizes job creation, improved quality of life, and alignment with EU, UNESCO, and Council of Europe standards.

Overall, Serbia's two state supercomputers, national AI strategy, and dedicated AI research institute place the country in the top 20% globally in AI readiness (Oxford Insights, 2025). Combined with a native speaker population of 11 million, this suggests that Serbian should be relatively well positioned in the domain of language technologies. And yet, Serbian is classified as fragmentarily supported in the European Language Equality framework (Krstev & Stanković, 2023). The ELE framework evaluates digital readiness by assessing technological factors (resources like corpora and tools) and contextual factors (demographic vitality, investment, research, education, institutional status), combining them into a Digital-Language-Equality (DLE) metric that benchmarks European languages against English across five support tiers: excellent, good, moderate, fragmentary, or weak/no support (Gaspari et al., 2022).

Consistent with ELE's classification of Serbian as fragmentarily supported, most interviewees view current language resource development in Serbia as insufficient but improving. Milan, a librarian, notes that Serbia's devastated textual corpora have historically placed it at a disadvantage, and that, compounded by limited awareness among cultural institutions and policymakers of the importance of language technologies, this has resulted in only incremental progress. Today, however, new perspectives are emerging. Ana, a mathematician and an NLP researcher, observes that awareness of digital language technologies' importance has expanded from a small circle of enthusiasts to multiple groups actively advancing the field, an evolution she characterizes as a "promising upward trajectory." As a next step, interviewees advocate developing native LLMs for Serbian, while noting that this goal depends on addressing persistent constraints around corpora, digitization, and institutional support, discussed in the next section.

3 Benefits and challenges of Serbian language technologies

Building on the call for native Serbian LLMs, respondents emphasize that multilingual LLMs, trained on dozens or even hundreds of languages, cannot replace native models because they capture Serbian only superficially through cross-lingual transfer. "You train the model on English, and then it learns how to roughly reproduce that in Serbian, but it lacks nuance, cultural and historical context, it simply doesn't have the breadth or depth," explains Marko, a data scientist. He argues that no cross-lingual transfer can replace native training data as "what isn't in the data won't be in the model." This assessment is supported by research showing that multilingual models may generalize across languages, but low-resource-language models trained from scratch consistently achieve better performance (Virtanen et al., 2019).

This is particularly relevant for Serbian, given its linguistic diversity that includes two scripts, five dialects, two phonological variants, and specific morphosyntactic features. Vera, a philosopher of language, describes Serbian as rooted in dialectal and ethnic pluralism, explaining that in the

[2] A territory that declared independence from Serbia in 2008; not recognized by the United Nations as an independent state.

province of Vojvodina alone the same Serbian dialect connects over twenty ethnic communities, each contributing something of their own. “That level of diversity is not the case for many other European languages that are often far more ethnically monolithic,” she points out.

Building on this and on contemporary views of LLMs as a means of language preservation (Koc, 2025), Jelena observes that “just as languages that never developed a writing system gradually declined and eventually disappeared with their last speaker, so too will languages that fail to build digital infrastructure and resources slowly fade away.” From a similar perspective, Milan asks whether models can meaningfully convey Serbian history, culture, and diversity if they are not trained on Serbian texts, but rely on data drawn from other languages. This aligns with Windsor’s (2022) argument about the role of language technologies and data in understanding international relations (IR): “Because non-English corpora are underrepresented in IR research using text-as-data methods, scholars’ understanding of international political phenomena is filtered through an ethnocentric, Western lens. This can lead to scholars misdiagnosing or failing to understand political phenomena since the experiences of political processes omit the perspectives offered by accounts in non-English languages” (p. 2).

Economic and security considerations also figure prominently in arguments for native language models. Respondents note that commercial LLMs are not only costly, but also unfeasible for the public sector. Concerns about data sovereignty, security, and compliance require that any model used in the public sector be domestically trained and hosted, without foreign infrastructural dependence. Beyond such practical arguments, several respondents emphasized the affective and symbolic dimensions of native language model development. “You feel left out when your language isn’t represented,” Ana observes; “culturally and linguistically, it’s an alienating experience.” She points out that this sense of exclusion is especially stark given current European strategies to build LLMs in support of a European language family. “Without an LLM for Serbian we are effectively left out of that family, even though we belong to it,” she concluded. This exclusion, respondents warned, is not just technological, it is cultural.

Yet, developing LLMs depends on large, high-quality corpora, an area where Serbia continues to face critical gaps due to limited data and partial institutional support.

3.1 Limited corpora and institutional support

A corpus (plural: corpora) is a large, structured collection of authentic language data, written or spoken, presented in machine-readable format, purposefully selected and sampled to be representative of a specific language, and often annotated with linguistic information (McEnery et al., 2006, p. 5). Large, high-quality, and representative corpora provide the statistical foundation for computational models to learn linguistic patterns, structures, and nuances necessary to accurately understand and generate language. However, the lack of adequate corpora remains one of Serbia’s most significant challenges. As Nikola, a mathematician, puts it: “The problem with data is that it simply doesn’t exist.” Vera seconds that opinion: “When it comes to language data, there really isn’t enough of it—that’s a fact.” Dragana, a computational linguist, further underscores the issue, noting that “the problem lies in how to create adequate datasets,” and that “in general, there isn’t as much material available [for Serbian] as there is for some other languages.”[3]

Currently, most training data for Serbian comes from scraping web sources. Respondents describe this as low-quality materials biased toward the most prevalent types of content found online—entertainment, casual conversation, and daily news. But even with web scraping as the largest source, insufficient corpora remain a pressing problem. Nikola illustrates this with the case of “Kišobran“ [Umbrella], the largest Serbian-language corpus available on Hugging Face. This project aggregates and de-duplicates data from sources like Common Crawl and Archive.org, yielding about 23 billion words. However, Nikola clarifies that this is not just Serbian, but a combined corpus of South Slavic languages. Serbian alone accounts for only 7 to 9 billion words.

[3] For a detailed overview of existing language models and datasets for Serbian, see Škorić, 2024.

By contrast, English-language models train on between 3 and 14 trillion words. “The gap in data volume is staggering,” Nikola concludes.

This data scarcity stems from centuries of destruction, suppression, and looting of Serbia’s written legacy, as detailed in the previous section, and is compounded by contemporary challenges such as limited digitization, a restrictive copyright regime, and insufficient institutional support.

Respondents acknowledge steady progress in digitization since the 1980s, especially over the past two decades, yet identify several systemic problems. One is the fragmented and uncoordinated nature of present-day digitization, which leads to duplication of work across institutions while leaving significant gaps. Another is the “early-digitization error,” when materials were scanned as images without OCR, rendering them non-searchable and unusable for language-technology projects today. This omission was not unique to Serbia; early digitization efforts elsewhere often omitted OCR due to limited accuracy, cost constraints, and a preservation philosophy that viewed digitization as faithful reproduction rather than textual transformation. Today, “the challenge lies in converting and transforming those images into actual text, especially when we encounter the specific complexities of the [Serbian] language, such as the use of both Cyrillic and Latin scripts,” Ana notes.

Copyright restrictions further intensify challenges of inadequate corpora by limiting access to both born-digital and digitized materials. “High-quality literature, which represents beautiful language, is ideal for language models. However, copyright restrictions make this just a dream,” Ana underscores. When asked why a larger high-quality corpus for Serbian does not exist, Dragana responds simply: “Copyright.” Marta echoes this frustration, stating that she has "no sympathy for copyright" because "any work—scientific, artistic, whatever it may be—is a public good when it concerns language." Scholarship on low-resource languages confirms that copyright restrictions exacerbate data scarcity by limiting access to precisely the kind of materials LRLs need to build comprehensive, high-quality language models (Moody, 2023; Siminyu et al., 2021). As Sandra, a digital humanist, points out, a desired dataset is a “dataset of national importance,” which includes “everything that represents the broader Serbian cultural, scientific, and other contexts.” Dragana concurs, noting that an ideal corpus should cover the full spectrum of language use—spoken and written forms, literature, television broadcasts, daily and weekly newspapers, public speeches, educational and academic discourse, and domain-specific terminology across fields such as law and medicine. Countries like Wales and Iceland have addressed the problem of copyright restrictions by working with writers’ associations and publishers to secure copyright owners’ permissions for using materials free of charge (Helgadóttir et al., 2012; Knight et al., 2020). Serbia's copyright law aligns with EU standards but lacks an AI-training exception comparable to EU Directive 2019/790, which allows text and data mining for AI training unless rights-holders have opted out.

Alongside legal harmonization, Serbia's language technology development also involves EU project funding, which primarily supports multilingual resources while leaving national language development to individual states. Respondents argue, however, that Serbian institutions tasked with advancing language technologies have shown limited engagement. "Even for basic tools like Serbian spell-check or a script converter from Latin to Cyrillic, there isn't a single authoritative tool produced by these institutions," Marko observes. This limited engagement proves particularly consequential given that academia lacks capacity to lead large-scale national projects and the private sector provides no alternative. Respondents note that most Serbian IT firms focus on international markets with no commercial incentive to develop Serbian-specific tools, while domestic-facing companies, primarily serving public administration, remain contract-driven and lack resources for independent research and development. "This is a clear-cut case for state intervention," Marko concludes.

But such intervention remains weak, respondents argue, though Serbia's latest (2025) national AI strategy includes language technologies. "It still hasn't reached the level of national understanding that this [language technology development] requires significant investment," Dragana explains. Vera elaborates: "There's pressure and expectations from the [political] top to be highly efficient in

digitally transforming society, but in real life, it's all extremely difficult and barely feasible given the lack of resources. It's as if we expect results to just appear out of thin air."

This critique extends to the state-funded Institute for Artificial Intelligence, which, according to respondents, has not prioritized authentic Serbian-language development. Marta illustrates: "They didn't use texts written in Serbian. They took existing OpenAI content, machine-translated it into Serbian, and then used that as training material. That's why the model generates made-up words like *аффекција*[4]." For respondents, this signals not only a technical failure but also a missed opportunity. Instead of building technologies based on authentic Serbian materials, state-sponsored institutions relied on machine translations that produced flawed training data susceptible to data bias—the very issue explored in the following section.

3.2 Data bias

Data bias in language technologies is not a mere technical flaw, but a reflection of broader societal and epistemological issues. Datasets reflect what is present in a society, reproducing and amplifying socio-cultural, political, racial, economic, and gender-based biases. This stems from multiple sources, such as limited data availability, underrepresentation of certain groups and topics, subjective annotation practices, and unexamined assumptions in corpus design and selection (Bender et al., 2021; Blodgett et al., 2020; Gebru et al., 2021). Although bias may sometimes result from the inherent structure of data, like gendered disease patterns in medicine, interviewees underline that its primary source lies in how datasets are created and curated. "Data bias means lack of representativeness—the model gives answers based on what it has seen most often, frequently favoring particular groups, phenomena, or activities," Dragana explains. "That doesn't mean it's true, or even the majority opinion, it's just what the model has encountered most frequently." She recounts prompting a commercial language model for jokes about Serbs and receiving "a stream of awful, deeply offensive content." Respondents stress that such outputs are not anomalies but structural symptoms of data imbalances, reflecting what the model has and, crucially, has not been exposed to during training.

Data bias is particularly pronounced in low-resource languages, where data scarcity combined with cross-lingual transfer from HRLs produces models that underperform and reinforce existing representational gaps. Multilingual models trained predominantly on HRLs tend to encode and transfer those dominant languages' linguistic, cultural, and social biases, resulting in misclassification of named entities, gendered translation errors, and overall degraded performance. This creates a double disadvantage for LRLs—underrepresentation combined with distorted representations rooted in unrelated linguistic contexts.

In LRLs, data bias is often amplified by the lack of proper data annotation, invisible labor that requires prolonged intellectual and manual work and adequate knowledge. Scarcity of qualified and motivated workforce for such efforts is yet another setback in Serbia: "The linguistic community is small, especially when it comes to computational linguists," Marta explains, while Dragana adds that "there aren't that many people interested in building the foundational resources; AI sounds fashionable but few actually want to do that groundwork."
Petar, a software engineer, adds that proper annotation must be paired with rigorous evaluation, since language models are inherently imperfect and inevitably contain some error. He critiques the historical reliance on quantitative metrics—"poor proxies for model performance," as he describes them— and welcomes the recent shift toward more nuanced evaluation criteria, including human expectations, creativity, linguistic quality, and fairness. "The field is finally engaging with these questions in a meaningful way," Petar concludes, pointing to a growing awareness of the multidimensional nature of model outputs. This shift matters because many of the most consequential errors are not just technical but reflect systematic skews in the data the model learns from. High-resource languages are not immune to data bias either. As Ana notes: "Not even the British national corpus is an adequate, fully representative corpus. It primarily reflects the language

[4] Transliteration-based hallucination of the English word affection.

of the middle class, adult males." This illustrates that bias is not merely a function of data volume but of whose voices are systematically included or excluded.

Developing more equitable language technologies thus requires sustained attention to the sociopolitical contexts of data creation, not just improvements in technical performance. In that regard, interviewees voice concerns that framing bias as an abstract algorithmic problem distracts from its socio-cultural origins. "We're implicitly blaming some abstract entity — an algorithm or dataset — instead of focusing on how datasets are created," Vera explains. Rather, shared human accountability across all stages of AI development is needed, as responsibility for mitigating data bias is necessarily distributed. "We're all responsible [for data bias], from those who collect and label data to those who design the systems," Ana argues. Petar adds that responsibility also lies with the deployer who places the model in front of users and must ensure its behavior aligns with intended purposes. Given this broad responsibility for ensuring representational justice, trustworthiness, and inclusiveness in AI systems, participants emphasize the importance of interdisciplinary collaboration, a theme the following section examines in detail.

4 Interdisciplinary collaboration

Among interviewed experts, interdisciplinary collaboration is common and valued in language technology development, with teams typically composed of computational and general linguists, mathematicians, computer scientists, librarians, and domain experts. Yet, respondents stress the persistence of two intellectual traditions: the engineering perspective, which often treats language as just another type of data and emphasizes efficiency, and the humanistic perspective, which views language as socially and culturally embedded.

On this subject, interviewees recounted scenarios where programmers, regularly operating under pressure to deliver results, underestimated the complexity of linguistic or cultural representation. For instance, Vera recalls advocating for greater inclusion of women authors in the training corpus, a move grounded in principles of diversity and representational justice, only to find some teammates prioritizing feasibility within the available computational and financial resources. Interviewees thus accentuate that humanities and social science (HSS) research contributes not only through ethical oversight, but also methodologically, by questioning the assumptions guiding dataset and model design. While engineers may prioritize performance and functionality, HSS scholars can unpack the epistemological and ethical implications of data selection and technical choices, identifying which decisions carry consequences beyond legal compliance.

However, HSS scholars' ability to effectively contribute to language technology development and assessment presupposes essential understanding of how these systems work, particularly their inherent limitations and typical error patterns. As Sandra remarks, foundational technical literacy, such as recognizing the kinds of mistakes statistical models tend to produce, is indispensable for meaningful engagement. Petar notes that HSS scholars provide invaluable frameworks for interrogating what gets encoded in training data and how certain populations may be included or excluded in computational representations, but such critique is less effective without the technical grounding needed for meaningful contributions to AI ethics. While interviewees do not expect HSS scholars to become data scientists or software engineers, they maintain that working knowledge of data structures and machine learning workflows can significantly enhance their impact. For instance, Marta emphasizes that annotating corpora and evaluating tagged data requires an understanding of both linguistic and computational constraints, especially when working with large datasets. Marko agrees, arguing that concepts such as "tokenization" and "hallucination" should be part of the analytical vocabulary of digital humanists and computational social scientists.

Ultimately, more robust collaboration between HSS and technical domains requires educational reform, interdisciplinary training, and institutional recognition of digital humanities and critical AI studies as essential fields for the development and evaluation of language technologies. Without this, the divide between social critique and technological design risks producing language technologies that fail to reflect the values, diversity, and realities of the societies they are meant to

serve. As Milan warns, if HSS scholars cannot explain their relevance to the broader public, or to engineers, they risk becoming sidelined in one of the most consequential technological developments of our time.

5 Conclusion

Developing language technologies for low-resource languages like Serbian presents both significant opportunities and persistent challenges. Using a contextualized ethnographic approach, this study situates these challenges within a broader struggle for knowledge equity, digital justice, and linguistic inclusion. Insights from in-depth interviews with Serbian scholars and practitioners enable close examination of the structural factors shaping language technology development in the AI age. Drawing on these accounts, this section synthesizes key findings and outlines strategic interventions that address both historical and contemporary barriers to building NLP capacity for LRLs such as Serbian.

One of the most significant insights concerns the inherent limitations of cross-lingual transfer in multilingual LLMs and the importance of native language models. Approaches relying on machine-translated data or localized versions of multilingual LLMs, optimized primarily for dominant languages, reveal deep methodological and linguistic weaknesses. These methods provide only shallow representations of LRLs, risking culturally inauthentic, nonsensical, or even offensive outputs that reproduce bias and reinforce linguistic marginalization. Studies confirm that native language models for LRLs, trained on diverse authentic data, surpass multilingual alternatives, demonstrating the value of language-specific training corpora over cross-lingual transfer.

Beyond technical performance, native models carry broader cultural significance by preserving languages and linguistic heritage in the AI age, enhancing cultural visibility and addressing the alienation citizens feel when their mother tongue is excluded from the global AI arena. Native models also address economic and security concerns, as public-sector applications require domestically trained and hosted models to ensure data sovereignty, regulatory compliance, and independence from foreign and/or commercial providers. This study thus strongly supports respondents' call to treat language technologies as core infrastructure, with sustained investment in curated, representative, open-access corpora and native models developed in the public interest, as illustrated by the ALIA infrastructure (ALIA, n.d.).

Another key insight concerns data scarcity, which Serbia exemplifies acutely, since much of the available training data comes from the web and varies in quality, influencing model behavior and sometimes producing inaccurate or biased outputs. Shaped by historical disruptions, this scarcity is deepened by limited digitization and restrictive copyright requirements. To address these challenges, successful digitization initiatives offer useful models for Serbian and other low-resource languages, demonstrating how coordinated high-quality scanning, strong metadata, and open-access policies can support the creation of representative repositories and link libraries, archives, and cultural institutions within a shared digital ecosystem (NARA, 2014, 2023; United Nations, 2020). Similarly, to navigate copyright constraints, Serbia and other LRL contexts could adopt models similar to those used in Iceland and Wales, encouraging authors to contribute their materials voluntarily, thus accelerating the creation of high-quality datasets.

This study also demonstrates that LRLs face distinctive forms of data bias, including representational bias (skew toward low-quality web content), amplification bias (the reinforcement of existing social imbalances, such as gender disparities in available texts), and annotation bias (arising from limited expert workforce in small linguistic communities). A constructive approach to these challenges is to treat bias mitigation not as a post hoc technical intervention, but as an integral component of corpus design and language model development. Grounded in interdisciplinary collaboration and community dialog, this approach supports inclusive datasets, equitable representation across linguistic varieties and social groups, and transparent documentation of data selection criteria. It recognizes that absences in datasets, such as missing dialects, underrepresented demographics, or excluded registers, perpetuate linguistic marginalization. In doing so, it reframes

language technology development as stewardship rather than extraction, offering concrete, replicable strategies for creating more representative and culturally authentic language models for LRLs.

For this approach to be effective, it is vital to strengthen interdisciplinary collaboration and raise public awareness of the importance of language technologies. Echoing Snow's (1959) "Two Cultures," participants' examples suggest that engineers often treat language as generic data, prioritizing speed and ease of deployment, while humanists emphasize cultural grounding and interpretive nuance. Political and economic pressures reinforce the engineering-driven mindset, summed up by one participant as "if it works, it's good enough," and favor adapting English-trained LLMs over developing language-specific infrastructure. Overcoming this pattern requires HSS scholars to participate as equal decision-makers across data, evaluation, and governance processes. Respondents thus advocate adding programming to linguistics curricula and AI literacy to humanities programs to build shared understanding of algorithmic bias. Institutionally, joint funding, interdisciplinary labs, and ethics-centered workshops could enhance collaboration.

Finally, raising public awareness of language technologies' importance for Serbian and other LRLs is critical for building societal and political attention and will to act. Effective awareness campaigns should frame language technologies as core infrastructure, not a luxury. This should be done with an understanding that smaller countries and LRLs face inherent tensions between the imperative to "keep up" with AI developments in larger, wealthier countries and the sustained investment required for language technology infrastructure. Such pressure manifests both externally, through dependence on foreign technologies that may misrepresent or marginalize the language, and internally, through policy agendas that favor flashy "digital transformation" over slower, less visible work of corpus building, annotation, and evaluation. Sustainable language technology development demands balancing competing pressures while maintaining commitment to patient, foundational work and fostering public engagement that demonstrates why such investment is essential for cultural continuity and digital sovereignty.

By situating Serbian within the broader global challenges facing low-resource languages in the AI era, this study argues that sustainable progress depends on three interlinked commitments: building authentic, high-quality corpora; practicing interdisciplinarity rather than merely invoking it; and mobilizing public awareness to catalyze political and financial support. If these commitments are met, Serbian can move from fragmentary support toward greater digital sustainability and visibility, offering a model for other low-resource languages under similar pressures. Significantly, after this study had been completed, such an opportunity did emerge, when the Serbian Office for Information Technologies and eGovernment, in collaboration with some of the interviewees in this study, announced an initiative to develop a large language model for Serbian, closely aligned with the approach advocated in this paper (Office for Information Technologies and eGovernment of the Republic of Serbia, 2026). Through the Serbian case, the paper thus argues that developing language technologies for low-resource languages extends beyond technical design and regulatory compliance, becoming a question of cultural stewardship and public responsibility.

References

ALIA. (n.d.). *ALIA | English: The public AI infrastructure in Spanish and co-official languages*. https://alia.gob.es/eng

Barbereau, T., & Dom, L. (2024). GPT-NL: Towards a public interest large language model. In J. Müller, S. Rümmele, & J. Ziegler (Eds.), *Proceedings of the 2nd Workshop on Public Interest AI co-located with the 47th German Conference on Artificial Intelligence (KI 2024)* (CEUR Workshop Proceedings, Vol. 3958, pp. 13–18). CEUR-WS. http://ceur-ws.org/Vol-3958/piai24-short2.pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bondarenko, M., Lushnei, S., Paniv, Y., Molchanovsky, O., Romanyshyn, M., Filipchuk, Y., & Kiulian, A. (2025). *Sovereign large language models: Advantages, strategy and regulations* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.04745

Deretić, J. (1987). *Kratka istorija srpske književnosti*. BIGZ.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2025). Serbo-Croatian. In *Ethnologue: Languages of the world* (28th ed.). SIL International. https://www.ethnologue.com/language/hbs/

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D. (2024). *Linguistic bias in ChatGPT: Language models reinforce dialect discrimination* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.08818

Gaspari, F., Grützner-Zahn, A., Rehm, G., Gallagher, O., Giagkou, M., Piperidis, S., & Way, A. (2022). *D1.3: Digital language equality (full specification)*. European Language Equality. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_3.pdf

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Greenberg, R. D. (2004). *Language and identity in the Balkans: Serbo-Croatian and its disintegration*. Oxford University Press.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., & Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages—SaLTMiL 8, AfLaT 2012* (pp. 67–72). CLARIN-IS. https://clarin.is/en/resources/mim/

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (PMLR, Vol. 119, pp. 4411–4421). PMLR. https://proceedings.mlr.press/v119/hu20b.html

International Trade Administration. (2025, December 2). Serbia—Information and communications technology market. *Country Commercial Guides*. https://www.trade.gov/country-commercial-guides/serbia-information-and-communications-technology-market

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (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* (pp. 6282–6293). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.560/

Khanna, S., & Li, X. (2025). *Invisible languages of the LLM universe* [Preprint]. arXiv. https://arxiv.org/abs/2510.11557

Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2020). Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh. *Language Resources and Evaluation, 55*(3), 789–816. https://doi.org/10.1007/s10579-020-09501-9

Koc, V. (2025). *Generative AI and large language models in language preservation: Opportunities and challenges* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.11496

Krstev, C., & Stanković, R. (2023). Language report Serbian. In G. Rehm & A. Way (Eds.), *European language equality: A strategic agenda for digital language equality* (pp. 203–206). Springer Nature. https://doi.org/10.1007/978-3-031-28819-7_32

Lauscher, A., Vulić, I., Ponti, E. M., Reichart, R., Korhonen, A., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4483–4499). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.363

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N., & Du, M. (2024). *Quantifying multilingual performance of large language models across languages* [Preprint]. arXiv. https://arxiv.org/abs/2404.11553

Lin, P., Ji, S., Tiedemann, J., Martins, A. F. T., & Schütze, H. (2024). *MaLA-500: Massive language adaptation of large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.13303

Liu, Z., Richardson, C., Hatcher, R., Jr., & Prud'hommeaux, E. (2022). *Not always about you: Prioritizing community needs when developing endangered language technology* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2204.05541

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Mihailovich, V. D., & Mikasinovich, B. (Eds.). (2007). *An anthology of Serbian literature*. Slavica Publishers.

Moody, G. (2023, May 16). A "blatant no" from a copyright holder stops vital linguistic research work in Africa. *Walled Culture*. https://walledculture.org/a-blatant-no-from-a-copyright-holder-stops-vital-linguistic-research-work-in-africa/

Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., & Alam, F. (2024). *ARADICE: Benchmarks for dialectal and cultural capabilities in LLMs* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.11404

Muhr, R., Mas Castells, J. Á., & Rueter, J. (Eds.). (2020). *European pluricentric languages in contact and conflict* (Österreichisches Deutsch—Sprache der Gegenwart, Vol. 21). Peter Lang. https://doi.org/10.3726/b16182

National Archives and Records Administration. (2014). *Strategy for digitizing archival materials for public access, 2015–2024*. U.S. National Archives. https://www.archives.gov/digitization/strategy.html

National Archives and Records Administration. (2023). *NARA bulletin 2023-04: Managing records created on collaboration platforms*. U.S. National Archives. https://www.archives.gov/records-mgmt/bulletins/2023/2023-04.html

Office for Information Technologies and eGovernment of the Republic of Serbia. (2026, January 14). *Srbija razvija veliki AI jezički model za srpski jezik*. https://www.ite.gov.rs/vest/sr/11704/srbija-razvija-veliki-ai-jezicki-model-za-srpski-jezik.php

Oxford Insights. (2025). *Government AI Readiness Index 2025*. https://oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025/

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2362–2376). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.185

Statistical Office of the Republic of Serbia. (2025). *Upotreba informaciono-komunikacionih tehnologija*. https://data.stat.gov.rs/?caller=2702&languageCode=sr-Latn

Ristić, D. (2016). *Kuća nesagorivih reči: Narodna biblioteka Srbije, 1838–1941*. Narodna biblioteka Srbije.

Schut, L., Gal, Y., & Farquhar, S. (2025). *Do multilingual LLMs think in English?* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2502.15603

Siminyu, K., Kalipe, G., Orlić, D., Abbott, J., Marivate, V., Freshia, S., Sibal, P., Neupane, B., Adelani, D. I., Taylor, A., Touré Ali, J., Degila, K., Balogoun, M., Diop, T. I., David, D., Fourati, C., Haddad, H., & Naski, M. (2021). *AI4D—African Language Program* [Preprint]. arXiv. https://arxiv.org/abs/2104.02516

Škorić, M. (2024). *Novi jezički modeli za srpski jezik* [Preprint]. arXiv. https://arxiv.org/abs/2402.14379

Snow, C. P. (1959). *The two cultures and the scientific revolution*. Cambridge University Press.

UCLA Department of Slavic, East European and Eurasian Languages and Cultures. (2026). *Background information (Serbian)*. https://slavic.ucla.edu/languages/bcs/serbian-background-info/

United Nations. (2020). *Roadmap for digital cooperation*. https://www.un.org/en/content/digital-cooperation-roadmap/

University of Cambridge Language Centre. (n.d.). *Bosnian, Croatian, Montenegrin & Serbian*. https://www.langcen.cam.ac.uk/resources/langb/bosnian.html

Vargas-Parada, L. (2025, November 27). Large language models are biased—local initiatives are fighting for change. *Nature*. https://doi.org/10.1038/d41586-025-03891-y

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1912.07076

Windsor, L. C. (2022). Bias in text analysis for international relations research. *Global Studies Quarterly, 2*(3), Article ksac021. https://doi.org/10.1093/isagsq/ksac021