Primljen: 3.12.2025.
Prihvaćen: 19.12.2025.

Stručni rad
UDK: 004.8:81'322:37.014

Understanding the Language Gap in Artificial Intelligence

Razumijevanje jezičnog jaza u umjetnoj inteligenciji

Marija Krstinić, Lucija Bačić
Tehničko veleučilište u Zagrebu, Hrvatska
marija.krstinic@tvz.hr, lucija.bacic@tvz.hr

Abstract

Artificial intelligence (AI) tools have become part of daily communication, education, and translation. Yet users quickly notice that these systems perform much better in English than in many other languages. This paper explores what might be called the AI language gap; the uneven quality of AI language tools across different languages. The discussion focuses on English and Croatian, drawing attention to how unequal access to data, digital resources, and investment shapes the performance of AI systems. The paper reviews examples from translation and text-generation tools, highlights consequences for education and communication, and reflects on how such differences influence linguistic equality. Finally, it suggests several steps educators, institutions, and policymakers can take to support fairer representation of smaller languages in the digital age

Key words*: artificial intelligence, language gap, digital inequality, education and communication*

Sažetak

Umjetna inteligencija (AI) postala je sastavni dio svakodnevne komunikacije, obrazovanja i prevođenja. Ipak, korisnici brzo uočavaju da su ovi sistemi znatno uspješniji na engleskom nego na mnogim drugim jezicima. Ovaj rad istražuje tzv. *jezični jaz u umjetnoj inteligenciji*, neujednačenu kvalitetu AI jezičnih alata među različitim jezicima. Diskusija se fokusira na engleski i hrvatski jezik, ističući kako nejednak pristup podatcima, digitalnim resursima i investicijama oblikuje performanse AI sistema. Rad razmatra primjere iz oblasti prevođenja i generiranja teksta, naglašava posljedice za obrazovanje i komunikaciju, te ukazuje na to kako ove razlike utiču na jezičnu ravnopravnost. Na kraju se predlažu koraci koje prosvjetni radnici, institucije i kreatori politika mogu poduzeti radi pravednije zastupljenosti manjih jezika u digitalnom dobu.

Ključne riječi: *umjetna inteligencija, jezični jaz, digitalna nejednakost, obrazovanje i*

komunikacija

1. Introduction

In the last few years, artificial intelligence has transformed how people write, translate, and learn languages and communicate. Many of us now use online translation systems, automatic grammar checkers or conversational chatbots. However, users soon discover an uncomfortable truth: these tools are often more reliable in English than in Croatian or other smaller European languages (Bačić, Krstinić i Tolnauer-Ackermann, 2025.). This imbalance, which we can call the AI language gap, is not only a technical issue but also a linguistic and cultural one (Anik et al., 2025.). When digital systems understand some languages better than others, the result is a new form of linguistic inequality. Speakers of smaller languages may find that their texts are translated less accurately, that automatic summaries make more mistakes, or that AI chat systems fail to understand idiomatic expressions.

Linguistic bias in artificial intelligence, that is, the uneven performance of language-processing models across different languages, is increasingly becoming the focus of research, as is determining why modern NLP methods favor the English language and how this affects the availability and quality of AI assistance for low-resource languages. In this paper, we will examine the existing literature on this phenomenon and define the points at which data and methodological shortcomings lead to the so-called language gap, its consequences for education and communication, and recommendations for further research. It draws on recent studies and reports, but its focus is descriptive and reflective, not experimental. The examples come mainly from English and Croatian, but similar patterns can be found across Europe and beyond.

2. The dominance of English in digital language technologies

Artificial intelligence is emerging as a transformative technology in the global educational and professional landscape, but its availability and quality remain drastically unequal across languages.

English has long held a special position in global communication, science, and education. That same dominance now extends into the digital world and the development of AI systems, creating a profound inequality in opportunities for speakers of languages other than English,

including Croatian (Stanford Report, 2025.). Most of the internet's written content is in English, and most AI systems are trained on huge collections of online text. Because there is simply more English data available, AI models learn English patterns more easily and perform better in that language. When developers collect material to train a translation or writing system, they need millions or even billions of sentences. For English, such text is abundant, from websites, newspapers, books, and social media. For Croatian and many other languages, the amount of available material is much smaller, and often less clean or consistent. As a result, the models that "learn" from this text can only form a limited picture of how the language works.

In practice, this means that translation systems, grammar checkers, or chatbots have a richer knowledge of English usage and vocabulary than of smaller languages. English idioms, collocations, and stylistic patterns are represented in far more examples. By contrast, Croatian words that appear rarely online, especially technical or regional terms, may confuse the system or produce unnatural results. This situation reflects what some researchers call digital language inequality (European Language Equality, 2023.). In the same way that English dominates in international education and publishing, it also dominates the datasets that shape how AI understands and produces language.

The nature and scope of the language gap can be summarized into two fundamental problems. The first is the English-centric development of AI technologies. Current large language models and AI systems, including ChatGPT, have been developed primarily on English data, resulting in significant asymmetrical support for English and only a few widely spoken world languages. Research shows that current AI language-processing technology is focused on only 2–3% of the most widely spoken global languages, meaning that the majority of the population does not have access to tools of comparable quality (Nicholas i Bhatia 2023.; Bella et al. 2023.).

The second problem is the quality of language tools, which varies greatly between languages (Lakew et al., 2018.). Croatian, as a minor language with a relatively small amount of digital resources, falls into the category of low-resource languages. An analysis of the current situation shows that Croatian language systems are susceptible to the same problems as other languages with a similar status: translation models often produce lower-quality output, especially for longer sentences and complex structures (Obadić et al., 2023.).

3. Impact on low-resource languages

The historical dominance of English in AI development is crucial, and the language gap in AI directly affects general access to information. Large language models are trained on vast amounts of English text available on the internet, which means that there is an inherent asymmetry in their competence and understanding (Seto et al., 2025.).

The result is that ChatGPT, Claude, and similar tools show significantly better understanding, generation, and translation of English than of Croatian (Reusens, M. et al., 2024.). The differential availability of knowledge and information deepens global inequality. Although AI translation tools have improved, quality remains inconsistent, especially for languages that are not high-resource pairs. This means that access to top international knowledge and research is regularly limited for speakers of minority languages. As a result, millions of speakers of local languages lose access to valuable data and business tools. The European Union reports that 21 out of 31 studied European languages are not supported by machine translation at all, placing about 80% of European languages at risk of "digital extinction" (European Parliament, 2020.). The language gap is specifically manifested in the fact that, for common AI workflows and services, Anglophone performance differs drastically from that for Croatian and similar languages. Translations produced by machine translation systems often lose style and nuance in Croatian, and tools for spell-checking and grammar correction provide significantly weaker feedback than they do for English. For example, the Croatian language has complex morphology (cases, aspects, agreements), which often leads to incorrect word endings or wrong grammatical forms when a language model is not sufficiently trained on such structures.

These differences widen the language gap in everyday use, and the consequences of this imbalance become visible in everyday use. Although some multilingual models have included Croatian, their accuracy for this language is often lower than for English. This is especially evident in specialized domains where Croatian linguistic material is scarce (Levy et al., 2023.). When translating between English and Croatian, online systems often handle short, literal sentences well but struggle with style, nuance, or complex word order (Li et al., 2024.). The availability of data and resources makes a critical difference.

Croatian has very limited digital resources for training language models, which has led to most advanced AI tools being either unavailable in Croatian or of lower quality (Rehm, Grützner-Zahn i Barth, 2025.). Croatian's rich morphology, its system of cases, aspect, and agreement, poses challenges that English-based systems are not always equipped to handle. Word endings may be wrong, gender agreement inconsistent, or idiomatic phrases mistranslated. In writing

assistance tools, the gap is similar. Grammar and spelling checkers for Croatian are far less advanced than those for English. While an English text might receive precise feedback on register, style, and tone, Croatian corrections are often limited to spelling and simple grammar. Automatic summaries or paraphrases can also sound unnatural or incomplete. These differences are not the fault of the language itself but of the resources behind the technology. English simply has a much larger and more diverse digital footprint. Croatian corpora, large electronic collections of text, do exist, such as the Croatian Web Corpus (Ljubešić i Klubička, 2016.), but they are modest compared with the enormous English databases used by global technology companies.

4. Consequences for education

Digital inequality in access to educational resources presents a significant barrier. Speakers of minor languages, including Croatian, face a smaller amount of online educational content in their own languages, which limits their learning opportunities (Chinta et al., 2024.). Research on online knowledge collections shows that contributors working in minor languages have difficulty finding resources to verify their articles, and AI language tools such as translation and spell-checking often produce errors that waste time and hinder their work (Nigatu, Canny i Chasins, 2024.). Biases in AI educational systems can reinforce existing inequalities. Although AI can potentially be used for personalized education, systems developed primarily on English data often display biases when applied to other languages (Chinta et al., 2024). These issues are particularly concerning in the context of assessment and knowledge evaluation, where AI detection can incorrectly classify the work of writers who are not native English speakers as AI-generated (Liang et al., 2023.).

Linguistic and cultural authenticity in learning remains an issue. Generative AI models often do not produce linguistically pragmatic and contextually adapted language content; they lack the social awareness and cultural authenticity that are essential for high-quality language learning (Godwin-Jones, 2024.). This is a particular problem for Croatian speakers who want to learn English through AI tools, but also for English speakers who want to learn Croatian.

Educational inequality and accessibility are deepening. Although AI can be used to democratize access to education, without thorough changes in the development and distribution of technology, there is a risk that existing inequalities will be reinforced. Students who have access to advanced AI tools in their own language have a significant advantage over those who

do not (Holstein i Doroudi, 2021.).

5. Consequences for communication

Errors in translation and comprehension issues are common when communicating with AI tools. Machine translation models are still not suitable for translating chat and conversation, despite the popularity of translation software (Li et al., 2022.). Složenost dijaloga i neformalnog jezika stvara značajne izazove. In multilingual environments, these limitations can lead to serious misunderstandings (Shaham et al., 2023.). Barriers for speakers of English as a foreign language in professional settings are well documented (Bačić i Krstinić. 2018.; Tolnauer-Ackermann, Bačić i Krstinić, 2025.). Error-checking tools often make mistakes when detecting AI-generated text, incorrectly classifying the work of non-native English speakers as machine-generated, which can have serious implications for education and employment. This is especially problematic because international teams are increasingly expected to use AI tools for communication (Liang et al., 2023.).

Communication barriers between language groups remain significant. Research shows that most speakers of Croatian and other small languages cannot access the same quality of AI communication tools as native English speakers. This can lead to situations in which certain speakers appear less competent or reliable, even though the real problem lies in the quality of the technology (Chen i Liu, 2024.).

6. European perspectives on the language gap in AI

Although this paper focuses on English and Croatian, the same pattern appears across Europe. The digital disconnection of small languages creates multiple inequalities. Croatian and other minor languages face gaps in digital support that make access difficult for speakers, poorly designed digital tools that negatively impact the integrity of the language and scripts, and unique vulnerabilities to surveillance for speaker communities (Zaugg, Hossain, i Molloy, 2022.). Although the European Union funds initiatives such as the European Language Grid, the scope remains fragmented, with unequal availability of resources between languages. Croatian has access to certain tools, but never to the same level as English (Rehm et al., 2020.). Research shows that there is a lack of evidence on the use of technology in minority language education, especially in multilingual contexts. This means that the needs and opportunities are not yet fully understood (Zhao, et al., 2024.).

Studies on language technology readiness (European Language Equality, 2023.) show that most European languages, especially those with fewer speakers, lack sufficient digital resources. Nordic languages such as Icelandic and Finnish have made strong efforts to develop local AI models, but many other languages remain underrepresented. European projects such as *No Language Left Behind* (Costa-jussà et al., 2022.) and *FLORES-200* (Meta AI, 2024.) have begun to address these differences. Their goal is to include as many languages as possible in translation and evaluation tasks. Even so, results consistently show that English and other large languages outperform smaller ones. While the English-Croatian comparison illustrates the uneven performance of AI systems between high- and low-resource languages, similar patterns are evident throughout Europe.

A growing body of research and policy documentation highlights how smaller or morphologically complex languages experience a persistent disadvantage in digital environments. These cases reveal a shared European challenge: linguistic diversity remains poorly represented in the data that underpins large language and translation models.

Slovenian provides a clear example of a language with strong local academic expertise but limited global data representation. Despite comprehensive corpora such as *Gigafida* and initiatives supported by CLARIN.SI, Slovenian lags behind English in advanced applications such as text summarization, dialogue systems, and instructional AI (Krek, 2022; Rehm i Uszkoreit, 2012.).

A similar situation can be observed in **Montenegro**, where emerging research and institutional reports highlight both opportunities and limitations in the use of AI for education and language technologies. The *Council of Europe* (2024.) examined how tools like ChatGPT influence academic integrity and digital literacy in higher education, noting that English-language systems remain far more capable than those handling Montenegrin or other regional languages. Likewise, the *UNDP Artificial Intelligence Landscape Assessment* (2025.) emphasized that national AI initiatives are still at an early stage, with scarce language data and limited local expertise in model training. Although universities such as the *University of Montenegro* and *University of Donja Gorica* have begun integrating AI into curricula, most applications rely on English-based tools. Montenegro thus mirrors the broader European trend: digital inclusion depends not only on technological access but also on linguistic representation within AI systems.

Polish, with over forty million speakers, demonstrates that even relatively large European languages can be under-resourced in AI. Domestic initiatives such as the *HerBERT* model (Mroczkowski et al., 2021.) narrow the gap, yet English-based data and cross-lingual transfer are still required for top-tier results. Evaluation benchmarks are less standardized, and smaller online corpora limit the scope of context-sensitive generation or translation (European Language Equality, 2023.).

In the **Baltic region**, **Latvian** and **Lithuanian**,both official EU languages, still face limited AI coverage. ELE and ELRC reports note that Latvian's text and speech datasets remain modest in size and quality, while Lithuanian lacks sufficient corpora and robust open-source tools for downstream tasks (European Language Equality, 2023.; ELRC, 2023.).

Maltese, another EU language, is among the most digitally under-resourced. Rosner's reports (2022., 2023.) describe a chronic shortage of annotated corpora and models able to handle Maltese morphology and frequent English code-switching. While progress has been made in tagging and translation, the language still lacks resources for competitive AI development.

For **Basque**, a non-Indo-European language spoken in Spain and France, linguistic isolation amplifies disparities. Rigau (2022.) and ELE findings emphasize that Basque's agglutinative grammar and small online footprint challenge mainstream AI pipelines, leaving it outside many commercial language models.

Welsh shows that political will alone does not erase the gap. The Welsh Government's *Language Technology Action Plan* (2018.–2024.) improved ASR and translation for government use, yet English systems still outperform Welsh equivalents in scope and reliability (Welsh Government, 2024.).

Finally, **Icelandic** demonstrates both the problem and a proactive remedy. The *Language Technology Programme for Icelandic 2019–2023* invested heavily in core resources and models. Nikulásdóttir et al. (2020.) report measurable progress in transcription and TTS, while warning that long-term funding is essential.

Together these cases confirm that the AI language gap is structural, not isolated. English dominates because of data volume and commercial focus. European cooperation, through initiatives like the European Language Equality project, remains vital to ensure that linguistic

diversity survives in the digital age.

7. Implications for language teaching and communication

For language educators, the AI language gap has practical and ethical implications. When teaching English as a foreign language, AI tools can be useful. Grammar checkers, vocabulary assistants, and writing generators often produce accurate English examples and explanations. However, these systems are designed primarily for English and may reflect Anglo-American usage and cultural references. When students use the same tools to write in Croatian, the experience changes. Suggestions are less precise, explanations may be missing, and output can sound awkward.

This can frustrate learners or mislead them about what is "correct." Teachers may need to guide students more actively, explaining that AI assistance is uneven across languages and that human judgment remains essential. Although progress is steady, the AI language gap is unlikely to disappear quickly. Technological advances, such as larger multilingual models and improved transfer learning, will continue to reduce differences in quality. However, data imbalance is a structural issue: English will probably remain dominant online for the foreseeable future. The crucial question is not whether the gap can be completely closed, but how small it can become through sustained effort. If current European projects continue and national initiatives actively contribute data, we can expect noticeable improvements within five to ten years.

On the other hand, without consistent support and cooperation, smaller languages may continue to lag behind each new wave of AI innovation. The future of linguistic diversity in artificial intelligence therefore depends not only on technical progress but on the collective will to make all languages visible in the digital world. Reducing the AI language gap requires both technological and social strategies. On the technical side, one promising method is transfer learning. In simple terms, this approach allows a model that has already learned patterns from a large, well-represented language such as English to transfer part of that knowledge to a smaller language such as Croatian. Rather than training a system entirely from the beginning, developers can fine-tune it using a smaller amount of high-quality Croatian data. This process has been shown to improve translation and text-generation accuracy for languages with limited resources, especially when the smaller language is related to a larger one. For example, models trained on several Slavic languages together often perform better on Croatian than those trained

separately. Another way to narrow the gap is through data sharing and collaboration.

The European Union has recognized linguistic inequality as a form of digital inequality. Initiatives such as *European Language Equality* (ELE), CLARIN, and the *European Language Grid* (ELG) aim to provide shared repositories of text, speech, and language models. These platforms allow researchers and institutions across Europe to contribute to and benefit from collective data resources. By pooling efforts, even countries with smaller languages can gain access to the infrastructure and expertise necessary for modern AI development. Education also plays a role. Linguists, teachers, and students should be encouraged to participate in data collection and evaluation projects. Annotating texts, creating bilingual corpora, and checking AI translations are tasks where human expertise is invaluable.

Collaboration between universities, libraries, and government agencies can help ensure that new language data are representative, ethically collected, and open for research use. Finally, policy support is crucial. If national and European funding bodies explicitly include linguistic diversity as a requirement in digital projects, more attention will naturally shift toward smaller languages. The goal is not only to make AI systems speak Croatian more fluently but also to protect linguistic identity and ensure equal digital participation for all European citizens.

8. Possible solutions and future directions

The AI language gap will not disappear quickly, but several realistic steps can help reduce it. The first step is to increase the amount of good-quality digital text available in Croatian. Universities, public institutions, and publishers can contribute by sharing open data from government documents, educational materials, and professionally edited texts. These collections form the basis for more accurate AI models. Researchers and journals should also encourage per-language evaluation, that is reporting separate results for each language, not only overall averages. When Croatian and similar languages are included in testing, weaknesses become visible and can be addressed directly.

Finally, educators and institutions can develop guidelines for responsible AI use. Awareness is the simplest and most immediate way to prevent misuse. In order to reduce the language gap, future research needs to focus on several key areas. First, on the development and evaluation of language AI models for small languages like Croatian, focusing on the specificities of language structure and culture. It is necessary to develop and evaluate models that better

understand idioms, context, and pragmatics in minority languages. Increasing the quantity and quality of language resources requires systematically building and sharing digital corpora for Croatian and other low-resource languages. Larger and better annotated text sets lay the foundation for more precise models and enable more robust processing of complex grammatical constructions.

Second, it is necessary to standardize evaluation. Scientific publications should report results by language, not just global averages, in order to clearly identify weaknesses for each language separately (McGiff i Nikolov, 2005.).This could eventually form multilingual benchmarks that include Croatian.

Third, the development of specialized language models and techniques for low resources is recommended. Practical examples such as HerBERT (the first Polish BERT model created by optimized pretraining for a fusional language) show that the transfer method from multilingual models to a specific language can be effective (Mroczkowski et al., 2021.).

Similarly, models like BERTić (for Bosnian/Croatian/Montenegro/Serbian) are already proving that a focus on regional corpora brings improvements (Ljubešić i Lauc, 2021). Future research could further develop transfer learning, few-shot, zero-shot techniques adapted to morphologically complex languages. The field of data enrichment (eg synthetic data generation, paraphrasing, back-translation) should also be explored for the enrichment of small corpora (Zhong et al, 2024; Mcgiff i Nikolov, 2025.).

Fourth, strengthening inter-institutional cooperation and support is essential: investments in language technologies, as the EU does through projects (European Linguistic Equality, CLARIN, ELRC), should be continued and expanded. Incentive policies and funding for research projects that explicitly include low-resource languages can accelerate data collection and the development of local models (Misra et al., 2025). Since the inclusive application of AI for language accessibility is a major challenge in itself, it needs to be systematically addressed at the research and policy levels.

Finally, areas that are insufficiently covered by current studies should become the focus of future work. For example, many works are limited to gap detection in tasks such as translation; more attention should be paid to other applications (self-learning, text understanding, voice technologies). Also, research on the impact of language structures (e.g. complex syntax or dialect range) on model performance is lacking. As McGiff and Nikolov (2025.) conclude, it is necessary to “extend the methods to a wider range of languages and work on open challenges

in building equitable generative language systems” In conclusion, future research should combine technical innovations (better algorithms, new learning techniques) and socio-organizational measures (collaboration, education, policy) to slowly close the language gap and enable equal access to AI technologies for all languages.

9. Conclusion

Artificial intelligence has opened exciting possibilities for communication, education, and translation. Yet its benefits are unevenly distributed across languages. English enjoys a clear advantage because it dominates online data, research, and investment. Croatian and many other languages lag behind, resulting in differences in translation accuracy, writing assistance, and access to digital content. The AI language gap is therefore both a technological and a cultural issue. For educators and linguists, understanding this gap is essential for fair and responsible use of digital tools. Steps such as building better language resources, insisting on transparent evaluation, promoting awareness, and supporting European cooperation can help ensure that all languages, not only English, have a place in the digital future.

The AI language gap poses a serious barrier to opportunity, equality, and security in education, communication, and business for speakers of Croatian and other small languages. While the technology has great potential to improve access to knowledge and communication, current developments highlight that advances in AI technology have primarily benefited speakers of English and a few other high-resource languages. Even in business environments, where the risks of communication errors are highest, AI tools routinely fail to provide a satisfactory level of support for speakers of minor languages. Without significant investment, community development, changes in how development teams approach multilingual AI, and organizational changes that make linguistic diversity valuable, the growing inequality is likely to deepen. This makes it a key issue for the future of global education, communication, and business success.

10. References

1. Adedeji, T., i Olabode, F. (2024). AI pismenost na jezicima s niskim resursima: Stvaranje AI videozapisa na Yoruba, *Računala i obrazovanje Otvoreno, 9*, 100122. https://doi.org/10.1016/j.caeo.2024.100122
2. Anik, M., A., Rahman, A., Wasi, A., T., and Ahsan, M., M. (2025). Preserving Cultural Identity with Context-Aware Translation Through Multi-Agent AI Systems, In Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025), pages 51–60, Albuquerque, New Mexico. Association for Computational Linguistics. https://aclanthology.org/2025.lm4uc-1.7/

3. Bačić, L., Krstinić, M., Tolnauer-Ackermann, T. (2025). A Comparative Study of AI Tool Use in English and Croatian Among the Students of Zagreb University of Applied Sciences, *MIPRO 48th ICT and Electronics Convention*. https://ieeexplore.ieee.org/document/11131825
4. Bačić, L. i Krstinić, M. (2018). Utjecaj straha od učenja engleskog jezika na poslovnu komunikaciju. Obrazovanje za poduzetništvo - E4E, 8 (2), 109-120. https://hrcak.srce.hr/213872
5. Bella, G., Helm, P., Koch, G., and Giunchiglia, F. (2023). „Towards bridging thedigital language divide“. https://arxiv.org/abs/2307.13405
6. Chen, Y. and Liu, Z., (2024). „WordDecipher: Enhancing Digital Workspace Communication with Explainable AI for Non-native English Speak-ers“, *In Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*. Pages 7–10. https://dl.acm.org/doi/10.1145/3690712.3690715
7. Chinta, A. S., Sharma, R., & Patel, P. (2024). „FairAIED: Navigating fairness, bias, and ethics ineducational AI applications“, *Journal of Educational Technology*, 62(4), 124-137. https://doi.org/10.48550/arXiv.2407.18745
8. Costa-jussà, MR i dr. (2022). Nijedan jezik nije ostavljen iza sebe: Skaliranje strojnog prevođenja usmjerenog na čovjeka. Meta AI. https://arxiv.org/abs/2207.04672
9. ELRC (Europska koordinacija jezičnih resursa). (2023.). *Izvješće o nacionalnoj radionici u Latviji.* https://language-data-space.ec.europa.eu/
10. European Parliament. (2020). European Day of Languages: Digital survival of lesser-used languages. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2020)652086
11. Europska jezična ravnopravnost (ELE). (2023). Strateški program i plan prema digitalnoj jezičnoj jednakosti u Europi. https://european-language-equality.eu/
12. Godwin-Jones, R. (2024). Generative AI, pragmatics, and authenticity in second language learning. https://www.researchgate.net/publication/385091269_Generative_AI_Pragmatics_and_Authenticity_in_Second_Language_Learning
13. Holstein, K., & Doroudi, S. (2021). Equity and artificial intelligence in education: Will “AIEd” amplify or alleviate inequities in education? https://doi.org/10.30935/jdet/17297
14. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). Stanje i sudbina jezične raznolikosti i inkluzije u NLP-u. *Zbornik radova s 58. godišnjeg sastanka Udruge za računalnu lingvistiku* (str. 6282–6293). https://doi.org/10.18653/v1/2020.acl-main.560
15. Koponen, M., i Tiedemann, J. (2025). Jesu li višejezični jezični modeli skretnica za jezike s nedovoljnim resursima? *Časopis za višejezične studije umjetne inteligencije, 4*(1), 33–52.
16. Krek, S. (2022). *Izvješće o slovenskom jeziku (ELE Deliverable D1.31).* Europska jezična jednakost. https://european-language-equality.eu/
17. Lakew, S. M.; Erofeeva, A.; and Federico, M. 2018. NeuralMachine Translation into Language Varieties. In Proceed-ings of the Third Conference on Machine Translation: Re-search Papers,

WMT 2018, Belgium, Brussels, October 31 -November 1, 2018, 156–164. Association for ComputationalLinguistics. https://aclanthology.org/W18-6316/

18. Levy, S., John, N., Liu, L., Vyas, Y., Ma, J., Fujinuma, Y., Ballesteros, M., Castelli, V., and Roth, D. (2023. Comparing Biases and the Impact of Multilingual Training across Multiple Languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10260–10280, Singapore. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.634/
19. Li Y, Suzuki J, Morishita M et al. (2022). Chat translation error detection for assisting cross-lingualcommunications. In: Proceedings of the 3rd workshop on evaluation and comparison of NLPsystems, pp 88–95. Online. Association for Computational Linguistics. https://aclanthology.org/2022.eval4nlp-1.9/
20. Li, Y., Suzuki, J., Morishita, M., Abe, K., & Inui, K. (2024). MQM-Chat: Multidimensional Quality Metrics for Chat Translation. https://doi.org/10.48550/arXiv.2408.16390
21. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, Volume 4, Issue 7. https://doi.org/10.1016/j.patter.2023.100779
22. Ljubešić, N. i Lauc, D. (2021). BERTi´c - the transformer language modelfor Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the8th Workshop on Balto-Slavic Natural Language Processing. Kiyv,Ukraine: Association for Computational Linguistics, Apr., pp.37–42. www.aclweb.org/anthology/2021.bsnlp-1.
23. Ljubešić, N., & Klubička, F. (2016). Hrvatski web korpus hrWaC 2.1. *CLARIN.SI.* https://www.clarin.si/repository/xmlui/handle/11356/1064
24. Ljubešić, N., i Lauc, D. (2021). BERTić: Transformatorski jezični model za bosanski, hrvatski, crnogorski i srpski. *arXiv preprint* arXiv:2104.09243. https://arxiv.org/abs/2104.09243
25. Mcgiff, J., Nikolov, N.S. (2025). Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review. https://arxiv.org/html/2505.04531v1
26. Meta AI. (2024). FLORES-200: Višejezično mjerilo evaluacije za strojno prevođenje. *Priroda.* https://doi.org/10.1038/s41586-024-08007-9
27. Misra, A., Zamir, S., W., Hamidouche, W., Inbal Becker-Reshef, I. i Lavista Ferre, J. (2025). AI Diffusion in Low Resource Language Countries. https://www.researchgate.net/publication/397280431_AI_Diffusion_in_Low_Resource_Language_Countries
28. Mroczkowski, R., Rybak, P., Wróblewska, A., & Gawlik, I. (2021). HerBERT: Efficiently pretrained transformer-based language model for Polish. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.1/

29. Nicholas G, Bhatia A. Lost in translation large language models in non-English content analysis. The Center for Democracy & Technology, 2023. https://doi.org/10.48550/arXiv.2306.07377.
30. Nigatu, H. H., Canny, J., & Chasins, S. E. (2024). Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-21). https://doi.org/10.1145/3613904.3642605
31. Nikulásdóttir, AB, Loftsson, H., i Guðnason, J. (2020). Program jezičnih tehnologija za islandski 2019. – 2023. U *N. Calzolari et al.* (ur.), *Zbornik radova s 12. konferencije o jezičnim resursima i evaluaciji (LREC 2020)* (str. 3397–3405). Europsko udruženje za jezične resurse. https://linguist.is/wp-content/uploads/2020/02/nikulasdottir2020language.pdf
32. Obadić, L., Jertec, A., Rajnović, M, and Dropuljić, B. (2023). C-XNLI: Croatian Extension of XNLI Dataset. In Findings of the Association for Computational Linguistics: ACL 2023, pages 2258–2267, Toronto, Canada. Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.142/
33. Rehm, G., & Uszkoreit, H. (ur.). (2012). *Slovenski jezik u digitalnom dobu (META-NET serija bijelih knjiga)*. Springer. https://doi.org/10.1007/978-3-642-30636-5
34. Rehm, G., Grützner-Zahn, A., Barth, F. (2025). Are Multilingual Language Models an Off-ramp for Under-resourced Languages? Will we arrive at Digital Language Equality in Europe in 2030? https://www.researchgate.net/publication/389129940_Are_Multilingual_Language_Models_an_Off-ramp_for_Under-resourced_Languages_Will_we_arrive_at_Digital_Language_Equality_in_Europe_in_2030
35. Rehm, G., Marheinecke, K., Hegele, S., Piperidiset, S. et al. (44 more authors). (2020). The European language technology landscape in 2020: language-centric and human-centric AI for cross-cultural communication in multilingual Europe. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo,H., Moreno, A., Odijk, J. and Piperidis, S., (eds.) Proceedings of the 12th Language Resources and Evaluation Conference. 12th International Conference on Language Resources and Evaluation, 11-16 May 2020, Marseille, France. Association for Computational Linguistics (ACL) , pp. 3322-3332. https://www.researchgate.net/publication/354462223_The_European_language_technology_landscape_in_2020_language-centric_and_human-centric_AI_for_cross-cultural_communication_in_multilingual_Europe
36. Reusens, M. et al. (2024). Native Design Bias: Studying the Impact of English Nativeness on Language Model. Performance. ArXiv. https://doi.org/10.48550/arXiv.2406.17385
37. Rigau, G. (2022). *Nalazi: Izvješće ELE-a o baskijskom jeziku (sažetak radionice STOA)*. Europski parlament. https://www.europarl.europa.eu/
38. Rosner, M. (2022). *Izvješće o malteškom jeziku (ELE Deliverable D1.25)*. Europska jezična jednakost. https://european-language-equality.eu/

39. Rosner, M., i Borg, C. (2023). Jezično izvješće malteški. U G. Rehm & A. Way (ur.), *Europska jezična tehnologija 2022./2023* . (str. 223–240). Springer. https://www.um.edu.mt/library/oar/bitstream/123456789/110515/1/Language_Report_Maltese_2023.pdf
40. Seto, S., Hoeve, He Bai, R., Schluter, N., and Grangier, D. (2025). Training Bilingual LMs with Data Constraints in the Targeted Language. In Findings of the Association for Computational Linguistics: ACL 2025, pages 19096–19122, Vienna, Austria. Association for Computational Linguistics. https://aclanthology.org/2025.findings-acl.977/
41. Shaham, U., Elbayad, M., Goswami, V., Levy, O. and Bhosale. S. (2023). Causes and cures forinterference in multilingual translation. In Proceed-ings of the 61st Annual Meeting of the Association forComputational Linguistics (Volume 1: Long Papers),pages 15849–15863, Toronto, Canada. Associationfor Computational Linguistics. https://aclanthology.org/2023.acl-long.883/
42. Stanford Report. (2025). The digital divide and linguistic inequality in AI systems. Stanford HAI Report. https://hai.stanford.edu/news/closing-the-digital-divide-in-ai
43. Tolnauer-Ackermann, T., Bačić, L., Krstinić, M. (2025). Exploring Students' Attitudes Toward Learning Business Communication with AI Assistance. Baška SIF (Spatial Intelligence Forum) Meeting. Baška, Krk Island, Croatia,Baška SIF Meeting
44. UNDP Crna Gora. (2025). Procjena krajolika umjetne inteligencije (AILA). Program Ujedinjenih naroda za razvoj. https://www.undp.org/montenegro/publications
45. University of Donja Gorica. (2024). Sticanje kompetencija u IoT-u i AI – InnovateYourFuture. https://www.udg.edu.me/novosti/innovateyourfuture
46. University of Montenegro. (2023). Vještačka inteligencija u nastavi može da poveća angažovanost i akademski uspjeh studenata. https://www.ucg.ac.me/objava/blog/136886
47. Velška vlada. (2024). *Akcijski plan za velške jezične tehnologije: Završno izvješće (2018–2024).* https://www.gov.wales/
48. Vijeće Europe. (2024). Navigacija granicom umjetne inteligencije: ChatGPT i akademski integritet u visokom obrazovanju u Crnoj Gori. https://www.coe.int/en/web/education
49. Wang, X., Liu, Y., i Lee, J. (2024). Kvantificiranje višejezičnih performansi velikih jezičnih modela. *arXiv preprint* arXiv:2404.11553. https://arxiv.org/abs/2404.11553
50. Zaugg, I., Hossain, A. i Molloy, B. (2022). Digitally-disadvantaged languages. Inter-net Policy Review, 11(2), 1-11. https://doi.org/10.14763/2022.2.1654
51. Zhao, A., Mitchell, J., Gasanabandi, G., Ullah, N., Barnes, K., & Koomar, S. (2024). Minoritised Languages, Education, and Technology: Current practices and future directions in low- and middle-income countries. EdTech Hub. https://doi.org/10.53832/edtechhub.0127
52. Zhao, X., i Singh, R. (2025). Prevladavanje nedostatka podataka u generativnom jezičnom modeliranju za jezike s niskim resursima: Sustavni pregled. *Časopis za istraživanje umjetne inteligencije, 78*, 412–430. https://doi.org/10.1613/jair.1.14456

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