Instytut Europy Środkowej

IEŚ Commentaries

Editorial Team: Agnieszka Zajdel (Editorial Assistant), Szczepan Czarnecki, Spasimir Domaradzki, Bartłomiej Krzysztan, Łukasz Krzyżanowski, Sławomir Łukasiewicz, Damian Szacawa, Agata Tatarenko

No. 1509 (14/2026) | 26.01.2026

ISSN 2657-6996

© IEŚ

Marlena Gołębiowska, Aleksandra Kuczyńska-Zonik

Small Languages, Big Models. Baltic States' Strategies in the Fight for Digital Sovereignty in the Age of AI

The development of generative artificial intelligence depends on the quality and scale of the corpora on which models are trained. For countries with limited linguistic data resources in the digital space, low representation in global datasets poses a risk of technological marginalisation. To counter this threat, the Baltic states are implementing a range of strategies to safeguard digital sovereignty, from sharing national language corpora with global corporations to developing their own specialised tools.

The generative capabilities and reasoning precision of large language models, which form the backbone of contemporary artificial intelligence, depend directly on the data with which they are trained. For algorithms to correctly interpret semantic context (such as cultural nuances or distinctions between formal and informal registers) and the morphological specificity of a given language (for example, extensive case inflection or agglutination), they must be exposed to datasets comprising trillions of tokens, that is, basic textual units such as words or their fragments, on which the models operate.

This creates a barrier for so-called low-resource languages. This group includes the languages of the Baltic states, spoken worldwide by approximately 5.1 million Lithuanian speakers, 2.2 million Latvian speakers, and 1.3 million Estonian speakers. This naturally translates into a deficit of digitally available texts in these languages. Insufficient representation during the training phase results in lower-quality outputs generated by global large language models: they struggle with inflection, lose coherence, and exhibit a higher tendency toward hallucinations, that is, the generation of false information. However, the consequences extend far beyond user inconvenience. Weak AI performance in local languages means that companies cannot fully leverage the potential of this technology, reducing their competitiveness relative to foreign entities. In response to this risk, countries in the region are adopting evermore diverse strategies.

Estonia, as the regional leader in digitalisation, has adopted a model based on cooperation with global technology giants. Authorities in Tallinn concluded that, since they are unable to build competitive alternatives to global models, they must instead ensure that existing models "learn" the Estonian context. A key asset in this strategy is the Estonian National Corpus, developed by linguists at the Institute of the Estonian Language (*Eesti Keele Instituut*, EKI), which digitises and catalogues a wide range of Estonian texts, from classical literature to contemporary media. In 2013, the Estonian National Corpus contained fewer than 0.6 billion words; by 2017, this had grown to 1.1 billion, to 2.4 billion in 2021, and to 3.8 billion by 2023, illustrating the exponential pace of data accumulation.

At the beginning of 2025, the Ministry of Justice and Digital Affairs decided to make the national corpus available under an open data framework. This strategy involves cooperation with key technology providers, with Meta being among the first companies to declare their intention to use these resources to optimise its models. The decision was justified by the Minister of Justice and Digital Affairs, Lisa Pakosta (Estonia 200), who argued that this step is essential to safeguard the survival of the language. In her view, sharing data with technology companies is the only way to create conditions for global models to understand the Estonian cultural context and to improve the quality of digital services for citizens.

In parallel, legislative amendments are underway to enable the use of all public textual data not covered by opt-out clauses for AI training. These measures are urgent: research indicates that as many as 63% of Estonian users currently receive incorrect or distorted responses from the most popular large language models. In this context, researchers from the TartuNLP group at the University of Tartu have developed a publicly accessible benchmarking tool, the Estonian LLM Leaderboard. This unique barometer enables users to compare the responses of different models to the same prompt in real-time and assess both their fluency and factual accuracy. The tool is continuously updated, making it possible to track which global models are making the greatest progress in learning Estonian and which still produce errors or succumb to hallucinations.

Lithuania, rather than pursuing the large-scale sharing of national language resources with global corporations, has focused on developing specialised domestic language solutions tailored to the specific needs of public administration and social services. Work in this area is led by the Lithuanian Language Institute (*Lietuvių kalbos institutas*), which cooperates with universities and the public sector to build corpora adapted to the needs of national AI systems. A selective approach is central to this strategy: data are collected and shared in a controlled manner, primarily for projects implemented to meet state needs.

An example of this approach is Neurotechnology, one of Lithuania's most internationally recognised technology companies. The firm develops advanced speech processing and voice synthesis systems tailored to the specific features of the Lithuanian language, which is characterised by complex inflection and mobile stress. These solutions are used, among other applications, in assistive technologies for people with visual impairments and in e-government systems. In addition, in May 2024, the Lithuanian Ministry of Economy and Innovation announced a call for proposals worth over 12 million EUR to support the development of AI solutions for the Lithuanian language. The aim is to create resources enabling the training of models for task-specific applications, such as disinformation detection or sentiment analysis in the information space, where general-purpose models tend to perform poorly.

Lithuania's digital sovereignty strategy, therefore, does not focus on "teaching the world Lithuanian" but rather on ensuring the functionality of the Lithuanian language, where it is critical, namely in public services, education, and inclusive technologies. With regard to public services, according to the conclusions of a July 2025 report by Lithuania's National Audit Office, full automation of administrative processes using AI could reduce civil servants' workloads by as much as 30%. At present, however, this potential remains largely untapped, with only 15% of Lithuanian public institutions currently using AI.

Latvia views its national language as a resource requiring special institutional protection and, therefore, prioritises the development of its own AI-based language technology infrastructure. A key solution is Hugo.lv, a state-run language technology platform developed for public administration as well as ordinary citizens. It offers automatic translation of documents and websites, speech recognition and synthesis tools, and language support for public e-services, while user data remains within the state infrastructure. The development of Hugo.lv is coordinated by the Cultural Information Systems Centre (*Kultūras informācijas sistēmu centrs*, KISC) in cooperation with the Latvian Language Agency (*Latviešu valodas aģentūra*).

Conclusions

- Although Lithuania, Latvia, and Estonia share the same goal, safeguarding the future of their languages in the age of artificial intelligence, they have adopted markedly different strategies, reflecting distinct interpretations of digital sovereignty.
- Estonia represents a model of digital sovereignty through accessibility. Its authorities assume that the absence of language representation in large models poses a greater threat to sovereignty than cooperation with global corporations.
- Lithuania and Latvia, by contrast, pursue a model of digital sovereignty through control. For both countries, preserving their decision-making autonomy over critical language resources and over the infrastructure supporting key state functions is paramount.
- These national strategies are complemented by regional initiatives. One example is Poland's PLLuM model, trained in part on Lithuanian and Latvian data, which enables languages without their own models to access generative AI tools. This illustrates the potential of shared sovereignty based on cooperation, rather than exclusively national solutions.