



Redakcja: Agnieszka Zajdel (sekretarz redakcji), Szczepan Czarnecki,
Spasimir Domaradzki, Bartłomiej Krzysztan, Łukasz Krzyżanowski,
Sławomir Łukasiewicz, Damian Szacawa, Agata Tatarenko

Nr 1509 (14/2026) | 26.01.2026

ISSN 2657-6996

© IEŚ

Marlena Gołębiowska, Aleksandra Kuczyńska-Zonik

Małe języki, wielkie modele. Strategie państw bałtyckich w walce o cyfrową suwerenność w erze AI

Rozwój generatywnej sztucznej inteligencji zależy od jakości i objętości korpusów, na których trenowane są modele. W przypadku państw z mniejszymi zasobami danych językowych w przestrzeni cyfrowej niska reprezentacja w globalnych zbiorach grozi technologiczną marginalizacją. Aby temu zapobiec, państwa bałtyckie wdrażają różnego rodzaju strategie w walce o cyfrową suwerenność – od udostępniania narodowych korpusów globalnym korporacjom po rozwój własnych wyspecjalizowanych narzędzi.

Zdolności generatywne oraz precyzja wnioskowania dużych modeli językowych – będących fundamentem rozwoju współczesnej sztucznej inteligencji – zależą bezpośrednio od danych, na których zostały wytrenowane. Aby algorytmy poprawnie interpretowały kontekst semantyczny (np. niuanse kulturowe czy rozróżnianie stylu urzędowego od potocznego) oraz specyfikę morfologiczną danego języka (np. rozbudowana odmiana przez przypadki czy aglutynacyjność), wymagają ekspozycji na zbiory liczące biliony tokenów (czyli podstawowych jednostek tekstu, takich jak słowa lub ich fragmenty, na których operuje model).

Stanowi to barierę dla tzw. języków o niskich zasobach (*low-resource languages*). Do grupy tej należą języki państw bałtyckich, którymi posługuje się na świecie kolejno: litewskim – 5,1 mln osób, łotewskim – 2,2 mln i estońskim – 1,3 mln. To naturalnie przekłada się na deficyt dostępnych cyfrowo tekstów w tych językach. Niedostateczna reprezentacja w fazie treningu skutkuje obniżoną jakością treści generowanych przez globalne duże modele językowe: mają one trudności z fleksją, gubią spójność wypowiedzi i wykazują wyższą tendencję do halucynacji, czyli generowania nieprawdziwych informacji. Konsekwencje wykraczają jednak daleko poza niedogodności użytkowników. Słaba wydajność AI w lokalnym języku oznacza, że przedsiębiorstwa nie mogą w pełni wykorzystać potencjału tej technologii, co obniża ich konkurencyjność względem podmiotów z zagranicy. W odpowiedzi na to ryzyko państwa regionu przyjmują różne strategie.

Estonia, jako lider cyfryzacji w regionie, przyjęła model oparty na współpracy z gigantami technologicznymi. Władze w Tallinnie wyszły z założenia, że skoro nie są w stanie zbudować konkurencji dla globalnych modeli, muszą sprawić, aby istniejące modele „nauczyły się” estońskiego kontekstu. Kluczowym zasobem w tej strategii jest Narodowy Korpus Języka Estońskiego, tworzony przez lingwistów z Instytutu Języka Estońskiego (*Eesti Keele Instituut*, EKI). Instytut ten digitalizuje i kataloguje szeroki zakres tekstów estońskich – od literatury klasycznej po współczesne media. Jeszcze w 2013 r. Narodowy Korpus Języka Estońskiego liczył niespełna 0,6 mld słów, w 2017 r. – 1,1 mld, w 2021 r. – 2,4 mld, a w 2023 r. już 3,8 mld, co obrazuje wykładnicze tempo przyrostu danych.

Na początku 2025 r. Ministerstwo Sprawiedliwości i Cyfryzacji podjęło decyzję o udostępnieniu narodowego korpusu w formule danych otwartych. Strategia ta zakłada współpracę z kluczowymi dostawcami technologii (jednym z pierwszych podmiotów, które zadeklarowały wykorzystanie tych zasobów do optymalizacji swoich modeli, była firma Meta). Decyzję tę uzasadniała minister sprawiedliwości i cyfryzacji Lisa Pakosta (Estonia 200), wskazując, że jest to niezbędny krok dla zabezpieczenia przetrwania języka. Jej zdaniem, udostępnienie danych firmom technologicznym to jedyny sposób, by stworzyć warunki do zrozumienia estońskiego kontekstu kulturowego przez globalne modele oraz podnieść jakość usług cyfrowych dla obywateli.

Równolegle procedowana jest nowelizacja prawa, która ma umożliwić wykorzystanie do treningu AI wszystkich publicznych danych tekstowych nieobjętych klauzulą *opt-out*. Działania te są pilne – badania wskazują, że obecnie aż 63% estońskich użytkowników otrzymuje od najpopularniejszych dużych modeli językowych błędne lub



zniekształcone odpowiedzi¹. W tym kontekście naukowcy z grupy TartuNLP przy Uniwersytecie w Tartu opracowali publicznie dostępne narzędzie porównawcze – Estonian LLM Leaderboard. Ten swoisty barometr pozwala użytkownikom na bieżąco zestawiać odpowiedzi różnych modeli na to samo zapytanie i oceniać ich płynność oraz poprawność merytoryczną. Narzędzie jest stale aktualizowane, co pozwala śledzić, który z globalnych modeli robi największe postępy w nauce języka estońskiego, a który wciąż popełnia błędy lub ulega halucynacjom.

Litwa zamiast masowego udostępniania narodowych zasobów językowych globalnym korporacjom postawiła na budowę wyspecjalizowanych krajowych rozwiązań językowych, które odpowiadają na konkretne potrzeby administracji publicznej i usług społecznych. Prace w tym zakresie prowadzi Instytut Języka Litewskiego (*Lietuvių kalbos institutas*), który współpracuje z uczelniami i sektorem publicznym przy tworzeniu korpusów dostosowanych do potrzeb krajowych systemów AI. Kluczowe jest tu podejście selektywne: dane są gromadzone i udostępniane w sposób kontrolowany, głównie do projektów realizowanych na potrzeby państwa.

Przykładem jest działalność Neurotechnology – jednej z najbardziej rozpoznawalnych litewskich spółek technologicznych na arenie międzynarodowej. Firma rozwija zaawansowane systemy przetwarzania mowy i syntezy głosu, dostosowane do specyfiki języka litewskiego, charakteryzującego się złożoną fleksją i ruchomym akcentem. Rozwiązania te wykorzystywane są m.in. w technologiach asystujących dla osób z niepełnosprawnościami wzroku oraz w systemach e-administracji. Ponadto w maju 2024 r. Ministerstwo Gospodarki i Innowacji Litwy ogłosiło nabór wniosków na kwotę ponad 12 mln euro na rozwój rozwiązań AI dla języka litewskiego. Celem jest stworzenie zasobów umożliwiających trenowanie modeli do zadań specyficznych, takich jak detekcja dezinformacji czy analiza sentymentu w przestrzeni informacyjnej, gdzie modele ogólnego przeznaczenia wykazują niską skuteczność.

Litewska strategia cyfrowej suwerenności nie polega więc na „uczeniu świata litewskiego”, lecz na zapewnieniu funkcjonalności języka litewskiego tam, gdzie jest on krytyczny – w usługach publicznych, edukacji i technologiach inkluzywnych. W przypadku tych pierwszych zgodnie z wnioskami raportu litewskiej Najwyższej Izby Kontroli z lipca 2025 r. pełna automatyzacja procesów administracyjnych przy użyciu AI mogłaby zredukować nakład pracy urzędników nawet o 30%. Obecnie jednak potencjał ten pozostaje niewykorzystany, gdyż z AI w administracji korzysta zaledwie 15% litewskich instytucji.

Łotwa postrzega język narodowy jako zasób wymagający szczególnej ochrony instytucjonalnej, w związku z tym stawia na rozwój własnej infrastruktury narzędzi językowych opartych na sztucznej inteligencji. Kluczowym rozwiązaniem jest Hugo.lv – państwowa platforma technologii językowych rozwijana na potrzeby administracji publicznej i obywateli. Oferuje ona automatyczne tłumaczenia dokumentów i stron internetowych, narzędzia rozpoznawania i syntezy mowy oraz wsparcie językowe dla e-usług publicznych, przy czym dane użytkowników pozostają w infrastrukturze państwowej. Rozwój Hugo.lv koordynowany jest przez Centrum Systemów Informacyjnych Kultury (*Kultūras informācijas sistēmu centrs*, KISC) we współpracy z Łotewską Agencją Językową (*Latviešu valodas aģentūra*).

Wnioski

- Choć Litwa, Łotwa i Estonia mają ten sam cel – zabezpieczyć przyszłość swoich języków w erze rozwoju sztucznej inteligencji – to obrały zupełnie inne strategie, wynikające z odmiennego rozumienia cyfrowej suwerenności.
- Estonia reprezentuje model suwerenności cyfrowej przez dostępność. Władze przyjmują założenie, że brak reprezentacji języka w dużych modelach stanowi większe zagrożenie dla suwerenności niż współpraca z globalnymi korporacjami.

¹ KPMG, *Trust, attitudes and use of artificial intelligence: A global study 2025*, <https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-lithuania-snapshot.pdf.coredownload.inline.pdf>; <https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-estonia-snapshot.pdf.coredownload.inline.pdf> [13.01.2026].



- Litwa i Łotwa realizują natomiast model suwerenności cyfrowej przez kontrolę. Dla obu państw kluczowe jest zachowanie niezależności decyzyjnej nad krytycznymi zasobami językowymi oraz nad infrastrukturą, która obsługuje kluczowe funkcje państwa.
- Uzupełnieniem tych strategii są inicjatywy regionalne. Przykładem jest polski model PLLuM, trenowany m.in. na danych litewskich i łotewskich, który umożliwia korzystanie z narzędzi generatywnej AI językom nieposiadającym własnych modeli. Pokazuje to potencjał suwerenności współdzielonej, opartej na współpracy regionalnej zamiast wyłącznie na rozwiązaniach narodowych.