Advancing Language Resources

Infrastructure, Data & Benchmarking

Eleri Aedmaa
Kristel Uiboaed

Estonian Language Institute
eki.ee

November 19, 2025

EESTI KEELE INSTITUUT

DATA COLLECTION

At Scale

Why Data Collection?

EESTI KEELE INSTITUUT

Linguistic Research

Support advanced language studies and analysis

Language Technology

Build better tools and LLMs for Estonian

Digital Vitality

Ensure Estonian thrives in the digital age

How We Do It?

EESTI KEELE INSTITUUT

COLLECT

Diverse texts across domains

→

PROCESS

Clean and standardize

→

SHARE

Open access for all

Goal: Create a rich, diverse corpus representing Estonian across genres, registers, and time periods

Estonian Language Corpus Progress

EESTI KEELE INSTITUUT

8.03bn

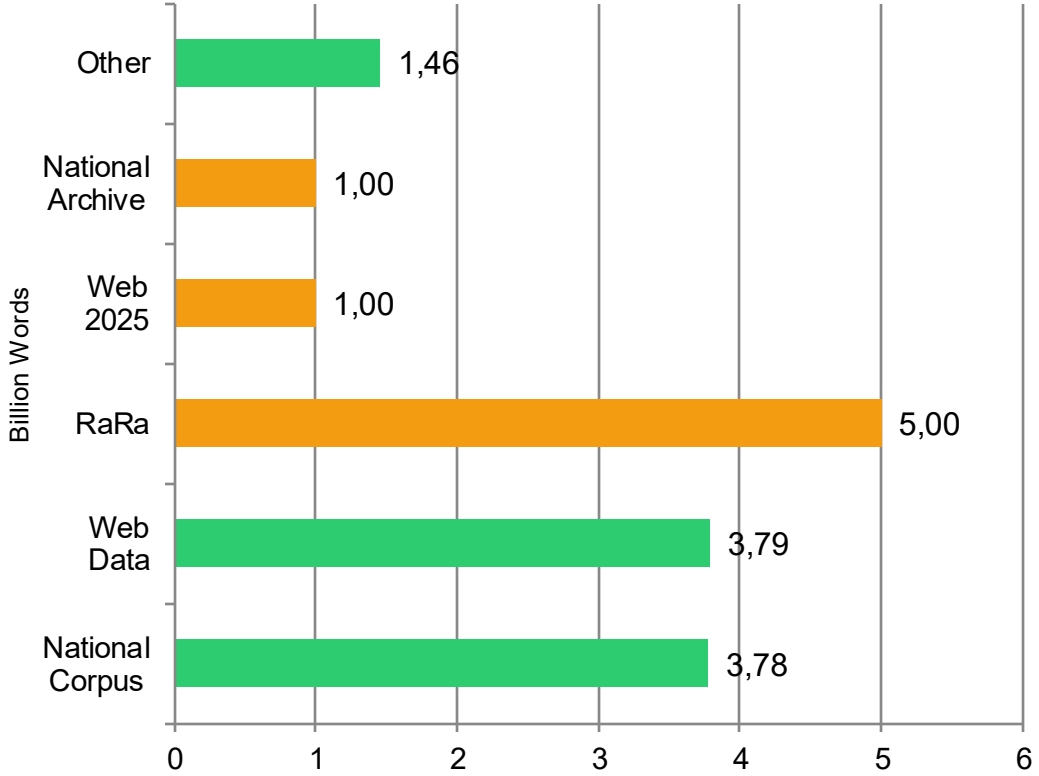
Words collected

54%

of 15bn target

Status

Completed	54%
In progress	43%
On hold	3%



Collaboration Partners

EESTI KEELE INSTITUUT

EE Estonian Partners

EstLLM (University of Tartu & TalTech)

EU European Initiatives

OpenEuroLLM (via HPLT)
TildeOpen LLM

🌐 Global AI Companies

Meta
Mistral AI

EESTI KEELE INSTITUUT

INFRASTRUCTURE

Building the Foundation

Keeleandmete Teadustaristu (KeTa)

Language Data Research Infrastructure

EESTI KEELE INSTITUUT

Complete infrastructure for collecting, preserving, and reusing Estonian language data with comprehensive processing tools.

Partners

EKI • TÜ HVEE
ATI • KirMus
TalTech

Fields

50%
AI Research

50%
Linguistics

KeTa Infrastructure Components

EESTI KEELE INSTITUUT

Data Repository

Storage, management, and preservation with metadata and version control

Data Collection

Text, speech, video, and sign language across multiple modalities

Language Software

Tools and services hosted on ETAIS infrastructure

Collaboration

CLARIN and ALT-EDIC networks for European integration

EESTI KEELE INSTITUUT

BENCHMARKING

Measuring What Matters

What is Benchmarking?

EESTI KEELE INSTITUUT

Definition

Evaluating language models using standardized tests to measure performance in reasoning, accuracy, safety, efficiency, and robustness

Fair Comparison

Objective metrics across models

Track Progress

Monitor improvements over time

Find Weaknesses

Identify areas for improvement

Choose Wisely

Select right model for use case

Why Estonian-Specific Benchmarks?

EESTI KEELE INSTITUUT

Measure What Global Benchmarks Miss

Estonian grammar, morphology, dialects, cultural context, and local knowledge

Transparency in Evaluation

Open, reproducible, locally governed tests that anyone can inspect and verify

Support Local AI Development

Shared benchmarks help build better Estonian-language tools and products

Why Estonian-Specific? (cont.)

EESTI KEELE INSTITUUT

Protect Small Languages

Ensure Estonian is represented and properly evaluated in AI systems

Improve Trust and Safety

Test hallucinations, bias, and misinformation in Estonian context

Improve Model Selection

Objective criteria for choosing LLMs for specific Estonian-language use cases

Benchmarking Strategy

Creating a transparent, reliable evaluation ecosystem

EESTI KEELE INSTITUUT

Define Standards

Map resources, identify gaps, set metrics

Design Benchmarks

Create task suites, datasets, scoring rules

Automate

Develop pipelines for recurring tests

Collaborate

Involve academia, industry, institutions, AI leap

Research

Conduct studies, publish findings

Improve

Refine through feedback and evidence

Current Benchmarking Work

EESTI KEELE INSTITUUT

📦 Evaluation Datasets

Comprehensive set by TartuNLP and TalTech available on Hugging Face

`huggingface.co/collections/tartuNLP/estonian-llm-evaluation`

📊 Tehisaru Baromeeter

Public leaderboard comparing LLM performance on Estonian tasks

`baromeeter.tartunlp.ai`

Building on strong foundations to expand, systematize, and maintain long-term

Key Takeaways

EESTI KEELE INSTITUUT

Data Collection

8+ billion words collected toward 15bn goal

Infrastructure

KeTa provides complete ecosystem for language data

Benchmarking

Transparent evaluation strategy for Estonian LLMs

"Let's make Estonian thrive in the digital age!"

Questions?

Eleri Aedmaa
eleri.aedmaa@eki.ee

Kristel Uiboaed
kristel.uiboaed@eki.ee