BALSU TALKA

Crowdsourcing Open Speech Corpora in Latvian and Latgalian

Baiba Saulīte[1], Ilze Auziņa[1], Kristīne Pokratniece[1],
Sanita Reinsone[2], Normunds Grūzītis[1], Roberts Darģis[1],
Artūrs Znotiņš[1], Ilze Ziņģe[1], Antra Kļavinska[3], Raivis Dejus[4]

1 Institute of Mathematics and Computer Science, UL
2 Institute of Literature, Folklore and Art, UL
3 Rēzekne Technology Academy
4 Latvian Open Technologies Association

At the beginning of 2023, there were only a few open Latvian speech corpora available, the 18-hour Common Voice (CV) 13 dataset being the largest one. In the result of a national crowdsourcing initiative, organised jointly by several institutions, the size and speaker diversity of the Latvian CV 17 release have increased more than tenfold in less than a year. A follow-up initiative was also launched for Latgalian.

Latvian speech recognition models:
huggingface.co/AiLab-IMCS-UL

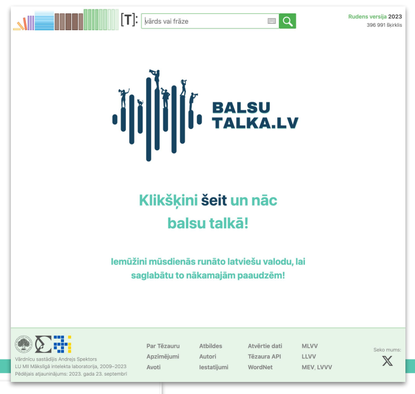
277 recorded hours
223 (81%) validated hours
5,712 speakers

Mar 2024

246 recorded hours
173 (70%) validated hours
5,174 speakers

Jan 2024

24 recorded hours
21 (81%) validated hours
250 speakers



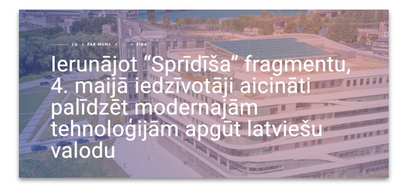
15 recorded hours
9 (60%) validated hours
169 speakers

Sep 2023

XXVII VISPĀRĒJIE LATVIEŠU DZIESMU UN XVII DEJU SVĒTKI

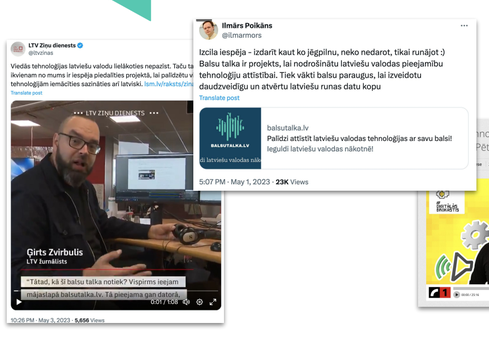
136 recorded hours
88 (65%) validated hours
2,396 speakers

Jun 2023



18 recorded hours
14 (78%) validated hours
321 speakers

Mar 2023



korpuss.lv

INSTITUTE OF LITERATURE, FOLKLORE AND ART, UNIVERSITY OF LATVIA
Institute of Mathematics and Computer Science University of Latvia
RTA RĒZEKNES TEHNOLOĢIJU AKADĒMIJA
volūda Latgalīšu kulturys kusteiba
LATA
unesco Latvijas Nacionālā komisija
VPP Valsts pētījumu programma
Finansē Eiropas Savienība NextGenerationEU
2027 Nacionālais attīstības plāns

EU Recovery and Resilience Facility project "Language Technology Initiative" (2.3.1.1.i.0/1/22/I/CFLA/002)
State Research Programmes' projects LATE (VPP-LETONIKA-2021/1-0006) and DHELI (VPP-IZM-DH-2022/1-0002)