Speech Corpora for Facilitating Linguistic Research and Tool Development

Ilze Auziņa, Guna Rābante-Buša, Roberts Darģis, Artūrs Znotiņš, Didzis Goško

Institute of Mathematics and Computer Science, University of Latvia

LATE conver-sations

Corpus size: ~ 30 hours (320k tokens)

Data:

- Recordings of private conversations, interviews and public appearances
- Orthographic transcription
- Morpho-syntactically annotated
- Metadata added

LATE media

Corpus size: ~ 50 hours (424k tokens)

Data:

- Recordings of media broadcasts
- Orthography of Standard Latvian, observing also the principles of punctuation
- Morpho-syntactically annotated

Balsu talka

Corpus size: 277 recorded and 223 (81%) validated hours by 5,712 speakers

Data:

- Derived from the Latvian *CommonVoice17* dataset
- Pre-selected sentences read by people of different ages and nationalities
- Morpho-syntactically annotated

Bolsu tolka

Corpus size: 10k Latgalian sentences, 24 recorded (by 250 speakers) and 21 validated hours

Data:

- Derived from the Latgalian *CommonVoice17* dataset
- First manually POS-tagged and lemmatized Latgalian corpus

Nacionālā korpusu kolekcija

NoSketchEngine

Linguistic Studies

Linguistic studies:

- Pronunciation analysis of words and word combinations
- Colloquial/spontaneous speech syntax
- Sentence intonation analysis
- Lexical analysis etc.

Orthographic transcription:

Type	Label
Acronym	*ASV [ā es vē]*
Internet address	*www.rigaslaiks.lv* [vē vē vē rīgas laiks punkts el vē]
Foreign language text	<en> *schortcuts [šortkats]* </en>
Unclear text	{*teksts*} {---}
Mispronunciation	*min* nu tas labklājības ministrs*
Pronunciation variant	*četrdesmit [čēesmit]*
Vocal hesitation	(ē) (ā) (m) etc.
Inhalation	<.h>
Exhalation	<h.>
Laugh	@

Corpus Search

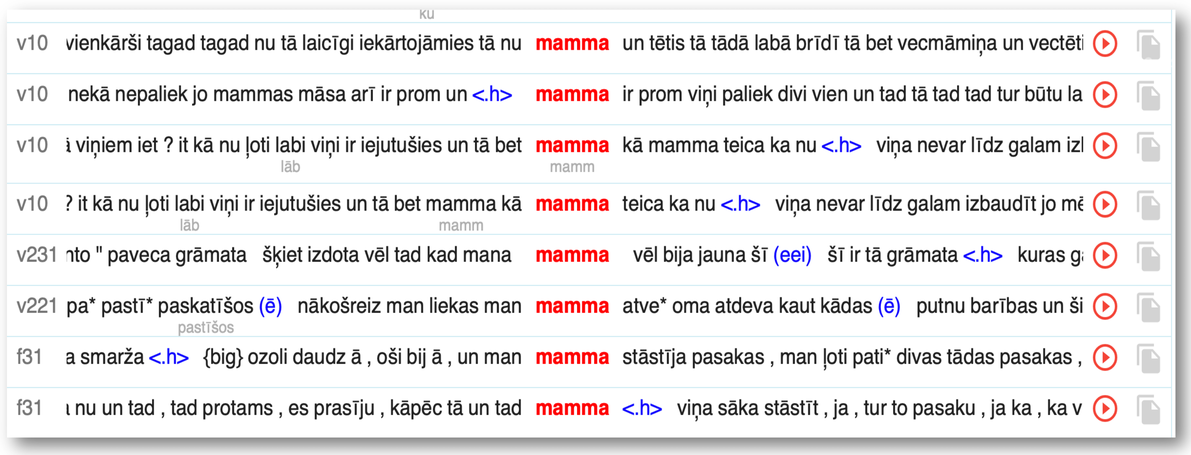
Corpora are included in the **federated search** as part of the **Latvian National Corpora Collection**.

Corpora are available for linguistic analysis via a ***NoSketchEngine*** instance.

Searching the corpus:

- For a word form or a phrase using regular expressions
- For pauses, inhalations and exhalations, laughing
- For deviations from the norms of pronunciation

Options for playback and downloading selected speech segments.



Tool Development

- Development of speech technology systems – an open source model for automatic Latvian language speech recognition (ASR) has been developed, refining the open source Whisper model.
- **https://late.ailab.lv/**
- Development of a text processing tool.
- Latvian speech recognition models: **huggingface.co/AiLab-IMCS-UL**

Institute of Mathematics and Computer Science, University of Latvia

INSTITUTE OF LITERATURE, FOLKLORE AND ART, UNIVERSITY OF LATVIA

RTA RĒZEKNES TEHNOLOĢIJU AKADĒMIJA

LATA

Funded by the European Union NextGenerationEU

2027 National Development Plan

SRP State Research Programme

This work has received financial support from the State Research Programmes under grant agreements No. VPP-LETONIKA-2021/1-0006 and No. VPP-IZM-DH-2022/1-0002; EU Recovery and Resilience Facility under grant agreement No. 2.3.1.1.i.0/1/22/I/CFLA/002