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



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



Label

ASV [ā es vē]

www.rigaslaiks.lv

[vē vē vē rīgas laiks punkts el vē]

<en> schortcuts [šortkats] </en>

{teksts} {---}

min* nu tas labklājības ministrs

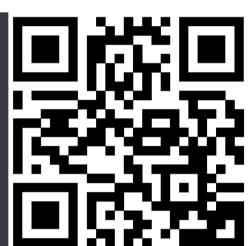
četrdesmit [čēesmit]

(ē) (ā) (m) etc.

<.h>>

<h.>

@





Linguistic Studies

Linguistic studies:

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

Orthographic transcription:

Internet address

Mispronunciation

Vocal hesitation

Acronym

Unclear text

Inhalation

Exhalation

Laugh

Lexical analysis etc.

Type

Foreign language text

Pronunciation variant

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:

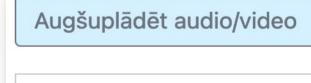
- For a word form or a phrase using regular expressions
- For pauses, inhalations and exhalations, laughing
- For deviations from the norms of pronunciation

Options for playback and downloading selected speech segments.

v10	vienkārši tagad tagad nu tā laicīgi iekārtojāmies tā nu	mamma	un tētis tā tādā labā brīdī tā bet vecmāmiņa un vectēti 🕞
v10	nekā nepaliek jo mammas māsa arī ir prom un <.h>	mamma	ir prom viņi paliek divi vien un tad tā tad tad tur būtu la 🕞
v10	i viņiem iet ? it kā nu ļoti labi viņi ir iejutušies un tā bet	mamma mamm	kā mamma teica ka nu <.h> viņa nevar līdz galam izl
v10	? it kā nu ļoti labi viņi ir iejutušies un tā bet mamma kā	mamma	teica ka nu <.h> viņa nevar līdz galam izbaudīt jo mē ()
v231	nto " paveca grāmata šķiet izdota vēl tad kad mana	mamma	vēl bija jauna šī (eei) šī ir tā grāmata <.h> kuras g
v221	pa* pastī* paskatīšos (ē) nākošreiz man liekas man	mamma	atve* oma atdeva kaut kādas (ē) putnu barības un ši 🕞
f31	a smarža <.h> {big} ozoli daudz ā , oši bij ā , un man	mamma	stāstīja pasakas , man ļoti pati* divas tādas pasakas , 🕞
f31	ι nu un tad , tad protams , es prasīju , kāpēc tā un tad	mamma	<.h> viņa sāka stāstīt , ja , tur to pasaku , ja ka , ka v 🕞

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



Tad es nolemju, ka šajā grāmatā es koncentrēšos tikai uz vienu lietu un praktiski uz... praktiski uz vienu stundu, ja skatāmies hronoloģiski.

Proti, grāmatas galvenā darbība ir viena stunda gara. Un grāmatas stāsts ir par veiksmi. Protams, tā kā es es<mark>mu</mark> ļoti haotisks, tad tur ir arī virkne vēsturisku ekskursu un arī ģeogrāfisku ekskursu dažādu cilvēku dzīvēs. Bet pats galvenais ir stāsts par to, kā cilvēks no biroja dodas pie ārsta.

Un katrā nodaļā tiek aprakstītas dažas minūtes no šī ceļojuma, no šīs kustības. Vienlaicīgi rādot, ko šajā pašā laikā dara citi cilvēki. Proti, brīžiem ir šī grāmatā tik koncentrēta, ka vienas minūtes notikumi tiek



00:00:36 / 00:02:07



Institute of Mathematics and Computer Science University of Latvia













Nokopēt tekstu

Saglabāt kā subtitrus