BalsuTalka.lv – Boosting the Common Voice Corpus for Low-Resource Languages

Roberts Darģis[*], Artūrs Znotiņš[*], Ilze Auziņa[*], Baiba Saulīte[*], Sanita Reinsone[†], Raivis Dejus[‡], Antra Kļavinska[¶], Normunds Grūzītis[§]

[*]Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, Latvia
[†]Institute of Literature, Folklore and Art of the University of Latvia, Mukusalas iela 3, Riga, Latvia
[‡]Latvian Open Technology Association, Raina bulvaris 19-9, Riga, Latvia
[¶]Rezekne Academy of Technologies, Atbrivosanas aleja 115, Rezekne, Latvia
[§]Faculty of Computing, University of Latvia, Raina bulvaris 19, Riga, Latvia
Corresponding authors: {roberts.dargis, arturs.znotins, ilze.auzina, baiba.valkovska}@lumii.lv

Abstract

Open speech corpora of substantial size are seldom available for less-spoken languages, and this was recently the case also for Latvian with its 1.5M native speakers. While there exist several closed Latvian speech corpora of 100+ hours, used to train competitive models for automatic speech recognition (ASR), there were only a few tiny open datasets available at the beginning of 2023, the 18-hour Latvian Common Voice 13.0 dataset being the largest one. In the result of a successful national crowdsourcing initiative, organised jointly by several institutions, the size and speaker diversity of the Latvian Common Voice 17.0 release have increased more than tenfold in less than a year. A successful follow-up initiative was also launched for Latgalian, which has been recognized as an endangered historic variant of Latvian with 150k speakers. The goal of these initiatives is not only to enlarge the datasets but also to make them more diverse in terms of speakers and accents, text genres and styles, intonations, grammar and lexicon. They have already become considerable language resources for both improving ASR and conducting linguistic research. Since we use the Mozilla Common Voice platform to record and validate speech samples, this paper focuses on (i) the selection of text snippets to enrich the language data and to stimulate various intonations, (ii) an indicative evaluation of the acquired corpus and the first ASR models fine-tuned on this data, (iii) our social campaigns to boost and maintain this initiative.

Keywords: speech corpus, crowdsourcing, ASR, low-resource languages, Latvian, Latgalian, open data

1. Introduction

Latvian, an official EU language, can no longer be classified as low-resource for many NLP tasks, including speech recognition. However, most Latvian speech corpora are closed data, available to a limited number of research institutions and language technology companies. Apart from proprietary datasets, speech corpora created by research institutions usually cannot be released as open data due to IPR and GDPR restrictions.

We launched a crowdsourcing campaign to address this issue by creating a relatively large, diverse and open speech corpus for Latvian. To ensure that such corpus is widely used in multilingual research and innovation, we are using the open-source Mozilla Common Voice (CV) platform.[1] Moreover, we are using the instance hosted by Mozilla rather than hosting our own. It has several advantages: Mozilla offers a tried-and-tested infrastructure that can handle large volumes of data without the risk of technical glitches or downtime; being an initiative of Mozilla, the CV platform comes with the assurance of strict privacy and security measures; the multilingual CV data repository is well known to the international NLP community, allowing for seamless inclusion of Latvian into multilingual research and language models thanks to the common data structure. In essence, by exploiting the CV platform, we ensure that the Latvian data is not only preserved and accessible but also readily usable, maximizing its impact.

On the contrary, to promote this crowdsourcing effort and to encourage participation in the recording and validation of speech samples, we used an approach that is both localized and personalized. Instead of promoting the CV platform directly, we established a tailored landing page for this initiative and came up with a catchy domain name that made it viral – BalsuTalka.lv.[2] We believe that a more targeted, relatable, and culturally resonant campaign has resulted in active participation across the whole country and also in the diaspora. BalsuTalka, in its nomenclature and design, is inherently Latvian. The strong local resonance ensured that the contributors felt a stronger

[1] https://commonvoice.mozilla.org

[2] It can be roughly translated as 'voice harvesting', although 'talka' is a quite unique Baltic ethnographic concept meaning communal work to achieve a common goal. https://balsutalka.lv

LREC-COLING 2024, pages 2080–2085
20-25 May, 2024.

personal connection to the project which has direct implications and benefits for the Latvian-speaking community and its future existence. Leveraging a local platform and social campaign allowed us to incorporate cultural nuances, stories, and symbols relevant to our target audience, thereby driving higher participation. With a simple but dedicated landing page, there is a clear and targeted call to action, eliminating potential confusion that might arise when promoting a global platform with multiple objectives and projects.

Before the initiative was launched, the Latvian CV corpus, vers. 13.0 (Mar 2023), contained 18 recorded and 14 (78%) validated hours of data by 321 speakers. In half a year since the initiative began, vers. 15.0 (Sep 2023) already contained 165 recorded and 88 (53%) validated hours by 2,773 speakers, placing Latvian in TOP5 w.r.t. the number of contributors per total number of native speakers. The latest release, vers. 17.0 (Mar 2024), contains 277 recorded and 223 (81%) validated hours by 5,712 speakers (TOP11 language w.r.t. the absolute number of contributors).

2. Related Work

Several speech corpora have been developed for Latvian (~1.5M native speakers) in the last decade for training and evaluating ASR models, for instance, a diverse general-purpose 100-hour corpus (Pinnis et al., 2014), a 10-hour corpus of dictation instructions (Pinnis et al., 2016), a 35-hour corpus of the radiology domain (Dargis et al., 2020), but they all are either proprietary or domain-specific, or very small.

We have been inspired by similar speech data crowdsourcing initiatives for other languages, but the most influential to our decisions have been the campaigns for Finnish (~5.8M native speakers) and Icelandic (~0.3M native speakers).

To collect spontaneous and colloquial speech, the Finnish campaign 'Donate Speech' (Linden et al., 2022) did not use the Mozilla CV platform, since reading text aloud triggers the use of more standardized speech. They collected a large and diverse dataset of speech recordings, but such data without orthographic transcripts could be used only for unsupervised pre-training of ASR models, and it would require for us a more laborious follow-up crowdsourcing campaign to add or post-edit the transcripts. Instead, we try to mitigate the use of standardized speech by careful selection of stylistically and functionally diverse text prompts.

The Icelandic campaign (Mollberg et al., 2020), on the contrary, did use the Mozilla CV platform, but by adapting the application and by hosting a separate instance. Therefore this dataset is not part of the multilingual CV dataset – it has to be discovered and integrated separately.

3. Data Selection

The first step to create and enrich a CV speech corpus is to submit and validate a text corpus – a set of text prompts to be read aloud. The Latvian CV corpus contained around 7k sentences before the campaign, mostly compiled from movie subtitles. To make the text prompts (sentences in the CV terminology, although they can be single utterances and small paragraphs as well) much more diverse and to raise the potential amount of speech recordings that can be submitted (CV imposes a limit of 15 recordings per sentence), we have not only quadrupled the size of the corpus (29k sentences currently) but also made it significantly richer in terms of genres, functional styles, and lexicon.

We followed the guidelines developed by the CV project: no more than 15 words, linguistically correct sentences without numbers, special characters and foreign letters. Another important criterion for the data selection is easily readable sentences of conversational style. To affect the expressiveness of the readings, question and exclamation sentences, fragments of dialogues and short dialogues are included in the data set. Additional considerations were taken into account form the validation of already recorded sentences, making the data augmentation process continuous: diversity of topics (incl. news headlines, food recipes), diversity of the lexicon (incl. named entities), diversity of sentence structures and communicative types of sentences.

The sentences have been selected so that the data can also be used to study various linguistic phenomena. The linguistic coverage of sentences comprises: (1) phonetically rich design, covering all phonemes; (2) phonological processes and alternations (e.g., vocalization of the consonants /v/ and /j/ after vowels in the same syllable). Texts from various sources were added to the Latvian CV corpus, including: (1) well-known and easily recognizable texts, such as Latvian proverbs and sayings, fairy tales; (2) food recipes, user manuals; (3) selected sentences from the Latvian National Corpora Collection (Saulite et al., 2022).

Sentences were initially collected manually.[3] To speed up the process of adding new sentences to the Latvian section of the CV platform, we automatically processed text snippets from the open Corpus of Latvian Pandemic Diaries (Reinsone et al., 2021). This colloquial style corpus of diaries originates from another Latvian crowdsourcing initiative encouraging participants to document their pandemic experiences. Automatically extracted and processed sentences were added also from

[3] Still, all sentences added to the CV platform have to be reviewed and accepted by at least two volunteers before presenting them to the speakers.

autobiographies, parliamentary debates (Auzina et al., 2018), and a Latvian Wikipedia corpus (Dargis, 2022).

4. Social Campaigns

The BalsuTalka initiative was introduced to the broader public on 4 May 2023 – Latvia's Restoration of Independence Day. The publicity campaign unfolded on a remarkable scale: information about the initiative was published across numerous online media platforms, received coverage on national-level TV and radio, and the press releases were picked up by regional newspapers and various online portals. Furthermore, we saw enthusiastic involvement in the dissemination activities from users across social media, including well-known journalists, political figures, technology experts, scholars, and social media personalities with a substantial number of followers.

We have tracked two performance indicators to assess the impact of the campaigns: (1) the number of visitors to the landing page, (2) the number of those who reached the goal – clicked one of the call-to-action buttons (Speak or Validate) on the landing page and were transferred ('converted') via outbound links to the CV platform (the localised CV user interface of the Latvian CV section).

In terms of attracting attention and converting visitors to participants, the initial campaign which actively lasted for one week proved to be highly successful. The landing page received nearly 6,000 visits, resulting in 3,100 conversions. The number of conversions by month is shown in Figure 1.

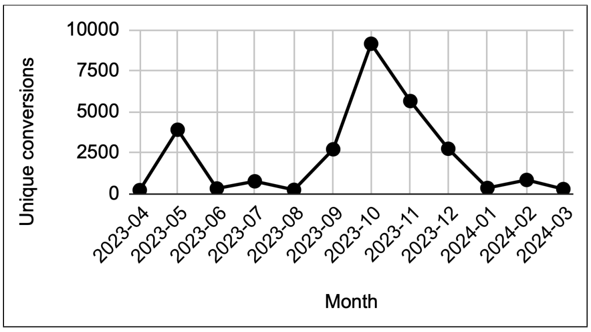

Figure 1: Unique conversions by month

The next two campaigns were less successful in terms of conversions, but they helped to maintain awareness of the initiative. In late June 2023, we introduced a smaller-scale campaign at the World Congress of Latvian Scientists. Meanwhile, substantial efforts were dedicated to encouraging voice recordings from participants at the nationwide Latvian Song and Dance Festival which took place at the beginning of July 2023 with over 40k festival participants. While these two campaigns did contribute to some increase in the number of speech recordings, the expected outcomes were not fully achieved since the festival itself took all the attention and time of its participants.

In contrast, for the fourth BalsuTalka major campaign, we chose to target an audience likely to have a greater interest in language-related matters. At the end of September 2023, we launched a new call for participation via the most popular online dictionary of Latvian, Tezaurs.lv (Spektors et al., 2016), whose monthly user base is ~300k users. This strategy effectively attracted a significant and lasting number of new visitors to the BalsuTalka landing page for the whole 3-month campaign, considerably surpassing the impact of the two previous campaigns. In total, the Tezaurs.lv campaign alone brought in more than 16k unique conversions, although not everyone eventually participated.

Table 1 gives a breakdown of traffic sources to the landing page, and the following conversion rates. We estimate that a large portion of direct traffic and traffic from search engines actually comes from TV and radio coverage. In total, the landing page have had more than 27k unique conversions with 37% conversion rate.

Source	Visitors	Conversions	Rate
Tezaurs.lv	46,483	16,367	35%
Direct traffic	14,710	4,104	28%
Search engines	6,957	4,012	58%
Social media	5,295	2,234	42%
News portals	624	305	49%
Other	449	180	40%

Table 1: Conversions from different sources

The overall progress of the Latvian CV corpus creation is shown in Figure 2. The initial BalsuTalka's target of collecting at least 100 hours was reached in the first month of the campaign. The latest release contains 277 hours (81% are validated).

5. Results and Evaluation

5.1. Linguistic Studies

We have made the latest release of the Latvian CV corpus available for linguistic analysis via a NoSketchEngine[4] instance as part of the Latvian National Corpora Collection[5] (Saulite et al., 2022). The dataset is morpho-syntactically annotated using an open-source tagger for Latvian (Paikens et al., 2013) to facilitate the linguistic analysis.

The data can be used to study various phonetic and phonological phenomena of Latvian, for example, the pitch accent or syllable tone that is in-

[4] https://nlp.fi.muni.cz/trac/noske

[5] https://korpuss.lv/en/id/BalsuTalka

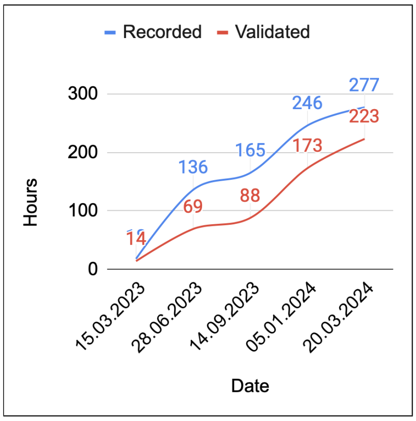

Figure 2: Progress of Latvian CV corpus creation

dependent of stress and is characteristic to each long syllable. There are great differences between the tonal systems of speakers from different areas. The same sentences read by different speakers are valuable data to study the information structure and the variation in the word order and sentence intonation to mark the focus.

5.2. Automatic Speech Recognition

We conducted fine-tuning and evaluation of pre-trained Whisper (Radford et al., 2023) and wav2vec 2.0 (Baevski et al., 2020) models, using the Latvian CV datasets, vers. 15.0–17.0, to gauge their efficacy, impact and potential enhancements. Additionally, we evaluated the models on a broadcast test set LATE-media (Dargis et al., 2024) to measure how well these models, trained on CV data, generalize out of domain.

We used data splits generated by the Common Voice Diversity Check,[6] with the v1 split (voice per split; includes all validated sentences) performing better than the default s1 split (sentence per split with no duplicates). Results are shown in Table 2.

Fine-tuning exclusively with CV 17.0 data (CV17) did improve performance on CV17 test data when compared with Whisper-large-v2 and MMS-1b-all (Pratap et al., 2023) baselines. The fine-tuned Whisper model, however, exhibited excessive over-fitting on CV17 data, leading to diminished performance on LATE-media data. In contrast, wav2vec 2.0 fine-tuning using wav2vec2-xls-r-300m as the base model showcased much more stable performance, with results improving on LATE-media as training progressed.

Eventually (with the CV17 release), the fine-tuned wav2vec 2.0 models ('w2v' in Table 2) outperformed both the Whisper and the MMS baselines on LATE-media test data. Notably, a pronounced disparity exists between the results on the CV17 and LATE-media test sets, pointing to the distinct nature of the CV dataset (Likhomanenko et al., 2020). Moreover, merging the CV17 and LATE-media training sets[7] did not enhance performance on LATE-media test data compared to a pure LATE model, but the combined model most likely generalizes better. We further experimented with different CV training sets, which were formed by sampling the corresponding full training split.

In the Latvian CV15 dataset, each sentence was recorded 9 times on average by different speakers. Utilizing a maximum of 3 recordings per sentence for training produced similar results on the LATE-media test set as the full CV15 training set, suggesting the necessity for greater linguistic diversity. In the improved Latvian CV17 dataset, each sentence is recorded 7 times on average, and the number should keep dropping.

The Latvian CV15 dataset predominantly consisted of recordings featuring rather short and relatively simple sentences, with an average of 5 words per sentence and a duration of ~3.9 seconds. To assess the effect of sentence length on performance, we sorted the dataset by audio duration and divided it into two subsets with equal effective audio length (excluding initial and terminal silences). Employing longer sentences showed significantly better results, which has been taken into account while moving towards the CV17 release (~6 words and ~4.4 seconds per sentence).

The best comparable results with a restricted training dataset were achieved through three similarly effective strategies: (1) limiting each sentence to a maximum of 3 recordings, (2) retaining the split containing longer sentences, (3) selecting the subset of longer sentences and limiting each sentence to a maximum of 5 recordings. These strategies each reduced the dataset to roughly one-third of the full CV15–CV17 v1 training datasets. This finding suggested further improvements in the sentence selection and confirmed the importance of constant enlargement of the text corpus to ensure that the same sentence is read by less speakers and the average sentence length increases.

The best-performing ASR model trained on Latvian CV17 data is available from a public Hugging Face repository.[8]

6. Latgalian CV Dataset

In parallel, we kick-started the creation of a Latgalian CV corpus. Latgalian has ~150k speakers and is recognised as an endangered language. It is a low-resource dialect of Latvian, having only two text corpora, and one speech corpus aimed

[6] https://github.com/HarikalarKutusu/common-voice-diversity-check

[7] LATE-media training data: 42 hours (see Table 2).

[8] https://huggingface.co/AiLab-IMCS-UL/wav2vec-xls-r-300m-lv-cv17

Model	**CV17-test (WER)**	**LATE-media (WER)**			**CV-train (hours)**		
		CV15	CV16	CV17	CV15	CV16	CV17
Baseline models							
whisper-large-v2	27.3	37.9			–		
mms-1b-all	16.4	29.3			–		
Fine-tuned models							
ft-whisper-cv-v1	**5.9**	52.9	47.9	46.0	66	123	167
ft-w2v-cv-v1	8.8	28.9	27.1	25.6	**66**	**123**	**167**
ft-w2v-cv-v1-d1	9.3	33.7	28.8	27.0	9	20	32
ft-w2v-cv-v1-d3	8.3	31.5	27.0	**25.0**	26	43	69
ft-w2v-cv-v1-long	9.0	31.7	27.8	26.0	27	50	68
ft-w2v-cv-v1-short	11.6	33.7	31.1	29.2	39	73	99
ft-w2v-cv-v1-long-d5	8.9	32.1	27.6	26.0	18	29	48
ft-w2v-cv-s1	11.7	37.8	30.5	28.0	4	11	18
ft-w2v-late	13.8	23.8			42		
ft-w2v-cv-v1-late	8.6	24.1	24.5	24.0	108	165	209

Table 2: The various CV15–CV17 training data splits: their sizes in hours, and word error rates (WER) of the different models trained on these datasets, evaluated on the CV17 and LATE-media test sets. Where: 'ft' denotes fine-tuned models; 'dN' denotes models trained on data where each sentence in the CV dataset is repeated up to N times by different speakers; 'long'/'short' refers to models trained with longer or shorter sentences; 's1' denotes the CV data split created by the s1 algorithm.

at the acquisition, documentation and research of Latgalian. The Corpus of Contemporary Latgalian Texts (Briska et al., 2022) contains 2M words and includes texts published in the Latgalian standard written language. The Latgalian Speech Corpus (Martena et al., 2023) currently contains 23 hours of audio recordings with orthographic transcriptions totalling 187k words. It documents natural, spontaneous speech, including field research recordings, interviews, TV and radio broadcasts.

To create a Latgalian CV corpus, we have localised the Mozilla CV user interface and the BalsuTalka.lv landing page[9] for the Latgalian campaign, and the first 5k Latgalian sentences were selected and submitted to the CV platform. The most important criteria in the selection of texts were their compliance with the orthography norms of the standard Latgalian, as well as phonetical, intonational (narrative, question, exclamatory sentences) and content diversity. Texts from dictionaries, short dialogues as well as phraseology from fiction and non-fiction texts were manually added (with prior agreement with the authors).

The latest Latgalian CV 17.0 dataset includes almost 10k Latgalian sentences, 24 recorded (by 250 speakers) and 21 validated hours. Latgalian speech corpus BolsuTolka[10], derived from the Latgalian CV17 dataset, is also the first manually POS-tagged and lemmatized Latgalian corpus included in the Latvian National Corpora Collection.

7. Conclusions and Future Work

The decision to launch and promote the initiative via a dedicated landing page and campaign while using the Mozilla CV platform has paid significant dividends. Not only we did notice upticks in participation rates, but the feedback we received from participants also indicated that they felt a stronger sense of ownership and pride in contributing to a 'local' initiative. This confirmed our guess that a localized approach, steeped in cultural relevance and relatability can outperform broader global campaigns, especially when the objective is to rally community support for the cause.

The success of the initiative underscores the importance of understanding and tapping into local contexts and sentiments, especially when mobilizing community-driven initiatives. While global platforms provide invaluable infrastructure and reach, it is often the localized touchpoints that inspire meaningful engagement and active participation.

Our evaluation highlights two contrasting objectives: while it is valuable for linguistic studies to have each sentence recorded by multiple speakers, linguistic diversity represented by fewer duplicates and longer sentences yields better performance for training ASR models.

We have released the best-performing wav2vec 2.0 CV17-LV version as an open source model. Its accuracy can be further increased by using a general or domain/task-specific language model.

Regarding the corpus itself, we still have to continue validation efforts to close the gap between recorded and validated data.

[9] https://balsutalka.lv/ltg/

[10] https://korpuss.lv/en/id/BolsuTolka

8. Acknowledgements

This work was supported by the EU Recovery and Resilience Facility project Language Technology Initiative (2.3.1.1.i.0/1/22/I/CFLA/002) in synergy with the State Research Programme projects LATE (VPP-LETONIKA-2021/1-0006) and DHELI (VPP-IZM-DH-2022/1-0002).

9. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Roberts Dargis, Normunds Gruzitis, Ilze Auzina, and Kaspars Stepanovs. 2020. Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian. In *Human Language Technologies - The Baltic Perspective*, volume 328. IOS Press.

Tatiana Likhomanenko, Qiantong Xu, and Vineel Pratap et al. 2020. Rethinking Evaluation in ASR: Are Our Models Robust Enough? *ArXiv*, abs/2010.11745.

Krister Linden, Tommi Jauhiainen, and Mietta Lennes et al. 2022. Donate Speech. In *CLARIN: The Infrastructure for Language Resources*. De Gruyter.

David Erik Mollberg, Olafur Helgi Jonsson, and Sunneva Þorsteinsdottir et al. 2020. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Peteris Paikens, Laura Rituma, and Lauma Pretkalnina. 2013. Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*.

Marcis Pinnis, Ilze Auzina, and Karlis Goba. 2014. Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*.

Marcis Pinnis, Askars Salimbajevs, and Ilze Auzina. 2016. Designing a speech corpus for the development and evaluation of dictation systems in Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Vineel Pratap, Andros Tjandra, and Bowen Shi et al. 2023. Scaling Speech Technology to 1,000+ Languages. *arXiv*, abs/2305.13516.

Alec Radford, Jong Wook Kim, and Tao Xu et al. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Baiba Saulite, Roberts Dargis, and Normunds Gruzitis et al. 2022. Latvian National Corpora Collection – Korpuss.lv. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Andrejs Spektors, Ilze Auzina, and Roberts Dargis et al. 2016. Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

10. Language Resource References

Auzina, Ilze and Dargis, Roberts and Bojars, Uldis and Paikens, Peteris and Znotins, Arturs and Rabante-Busa, Guna. 2018. *Corpus of the Saeima (the Parliament of Latvia)*. IMCS at University of Latvia. [link].

Briska, Anna and Zinge, Ilze and Dargis, Roberts and Pokratniece, Kristine. 2022. *Corpus of Contemporary Latgalian Texts 2022*. IMCS at University of Latvia; Rezekne Academy of Technologies. [link].

Dargis, Roberts. 2022. *Latvian Wikipedia*. IMCS at University of Latvia. [link].

Dargis, Roberts and Znotins, Arturs and Auzina, Ilze and Rabante-Busa, Guna. 2024. *LATE Dev&Test Set for Latvian ASR*. IMCS at University of Latvia. [link].

Martena, Sanita and Nau, Nicole and Klavinska, Antra and Jusko-Stekele, Angelika and Kocins-Kucens, Armands and Sprukte, Ausma and Briska, Anna and Gusāns, Ingars and Mazure, Laura. 2023. *Latgalian Speech Corpus*. Rezekne Academy of Technologies. [link].

Sanita Reinsone and Ilze Laksa-Timinska and Justine Jaudzema. 2021. *Corpus of Latvian Pandemic Diaries 2020–2021*. ILFA at University of Latvia. [link].