Towards the speech recognition for Livonian

Valts Ernštreits
University of Latvia Livonian Institute
valts.ernstreits@lu.lv

Abstract

This article outlines the path toward the development of speech synthesis and speech recognition technologies for Livonian, a critically endangered Uralic language with around 20 contemporary fluent speakers. It presents the rationale behind the creation of these technologies and introduces the hypotheses and planned approaches to achieve this goal. The article discusses the four-stage approach of leveraging existing data and multiplying voice data through speech synthesis and voice cloning to generate the necessary data for building and training speech recognition for Livonian.

1 Introduction

In October 2024 University of Latvia Livonian Institute has launched a new project aimed at developing speech synthesis for Livonian. This project has a broader goal of laying the groundwork for future speech recognition technology for the language. This article outlines the rationale behind the creation of speech synthesis and speech recognition technology for Livonian, and presents the hypotheses and approaches intended to achieve this goal.

Recently, as technological advancements progress, significant efforts have been made to develop speech-recognition technologies for underrepresented and critically under-resourced languages, including Uralic languages (e.g. Partanen et al 2020). This work is also being undertaken by two key partners of the UL Livonian Institute in this project – AILAB [1] and the University of Tartu NLP group.

It is worth noting that the current article focuses on the broader context, data production and handling approaches, and hypothetical methods for acquiring data and achieving the desired quality. The article does not cover the actual development of the technologies (speech synthesis and recognition) themselves, which will be conducted later in the project by the UL Livonian Institute's partners.

2 Broader context

The world's linguistic and cultural diversity is in urgent danger. To safeguard endangered languages and intangible cultural heritage, major policy documents and programmes have been introduced, e.g., the UNESCO Convention for the Safeguarding of the Intangible Cultural Heritage (2003), the UN International Decade of Indigenous Languages (2022–2032; further – IDIL) and others. A large part of this work in the digital era has been finding ways to narrow the digital gap between languages and cultures with extensive resources and data and the – mostly endangered – ones lacking resources and especially the data needed for creating digital technologies.

With every new technology, this gap only grows, and closing it is crucial for keeping endangered languages competitive and vital. Thus, in 2024, the Ad-Hoc Group on Digital Equality and Domains was created by the Global Task Force of the IDIL to tackle these issues.

Another crucial aspect of this work is ensuring access to existing data and archives, especially materials only in analogue format. Such archives may be a great resource for academia and communities themselves, e.g., for language acquisition or expanding the language environment, but not all analogue materials can be easily digitised and utilised, as some are scattered

[1] Artificial Intelligence Laboratory at the University of Latvia Institute of Mathematics and Computer Science.

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 76–80

or require considerable skilled human resources to be made digitally accessible.

As endangered languages enter the digital world, this access only becomes more important, as these populations are increasingly exposed to majority populations and subject to free movement and the rapid expansion of the information space – also through recently developed AI-based tools for generating language content (Wang 2024; Haboud, Ortega 2023, 287, 298).

Emerging technologies including AI offer new challenges, but also new possibilities, for tackling these issues. Due to a lack of sufficient data and other aspects specific to critically under-resourced languages and cultures, rule-based and AI technologies cannot and should not be used in the same way as when "big data" is available. Specific methods must be invented to meet the specific needs of languages and cultures with extremely limited resources. Obtaining knowledge and developing such approaches, methods, and tools is a speciality and one of the cornerstones of the research and resources developed by the University of Latvia Livonian Institute for Livonian language and culture.

3 Creating digital resources for Livonian

Like many other Uralic languages, Livonian with around 20 contemporary fluent speakers (Druviete, Kļava 2018) is highly constrained in terms of its domains of use and available resources. While many digital tools developed in the 21st century can potentially offer support, they first require large amounts of digital data. This presents a significant challenge not only for Livonian but also for most Uralic languages and the majority of the world's languages.

Since its founding in 2018, the University of Latvia's Livonian Institute has been developing a range of digital resources and methods for Livonian, focusing on data acquisition, efficient data use, resource-saving workflows, and improved accessibility.

What began as a digital dictionary has now expanded into a cluster of interconnected databases, accessible through Livonian.tech (LT). This includes an up-to-date lexical database with translations of Livonian lemmas and example sentences in Estonian, Latvian, and English; a morphology database; a partially annotated corpus of written Livonian with translations; and a geospatially linked place-name database. New databases focused on documentation sources and individuals are currently in development.

To meet the specific data extraction and management needs of extremely low-resource languages, tailored data processing approaches have been implemented in the construction of this cluster. In collaboration with partners, the institute has also explored opportunities to introduce various digital technologies for Livonian.

However, Livonian continues to face significant obstacles in obtaining new data, particularly due to the extremely small number of contemporary speakers and the limited domains in which the language is actively used. The language is also nearing a "glass ceiling" of available written data[2], much of which has already been digitized.

4 Data extraction or data synthesis?

At the core of every technological advancement in the digital field lies the availability of sufficient data. For languages like Livonian, which have very few contemporary speakers, extremely limited domains, and is entirely overshadowed by dominant language, the natural production of new data is an enormous challenge. Therefore, to obtain the additional data necessary for developing language technologies, one must either rely on extracting data from earlier documentation (if available) or turn to synthetic data produced through the use of technology. However, both approaches come with their own flaws and risks.

In terms of using earlier documentation, written records are generally more accessible. Converting analog documentation into digitally usable form requires only basic technologies (such as OCR for printed materials) or staff with basic language skills, knowledge of orthography, and an understanding of transcription methods used in the original documentation. However, the amount of pre-existing written documentation is usually quite limited, as is the case with Livonian.

Transliterating spoken language poses a much greater challenge. Not only is it significantly more time-consuming, but it also requires full

[2] The Livonian corpus representing written Livonian sources currently consists of ca 500 000 tokens; its full extent can be estimated at ca 750 000 tokens.

proficiency in the language to accurately capture what is being spoken. Consequently, manual transliteration can deplete the limited human resources proficient in the language for an indefinite period of time.

When it comes to synthesizing language data with the help of technology, a major risk is that the data produced may be of poor quality, due to the lack of sufficient data for building or training the necessary models. Poor-quality synthetic data may even pollute the digital environment of the language (Ernštreits, Fišel et al., 2022, p. 24; Trosterud, 2009), particularly given the ability of many technologies to generate large amounts of data in a short time—amounts that could far exceed the natural language data available for languages like Livonian.

A relevant example is machine translation, which is often one of the first technologies mentioned in discussions about endangered and low-resource languages. As with human translation, the quality of machine translation depends on its ability to generate accurate output in the target language, which, in turn, relies on the availability of training data and the methods used. Consequently, it is easier—and more beneficial—to develop machine translation from the endangered language to provide access to texts written in it, rather than to create machine translation into the endangered language that can synthesize high-quality data for further use.

Experiments with machine translation have been conducted also for Livonian (e.g., Rikters, Tomingas et al., 2022; Ernštreits, Fišel et al., 2022), but the evaluation of these results has shown that even with the use of all available aligned and monolingual corpora and databases, the data was still insufficient to build a high-quality machine translation model capable of synthesizing quality Livonian data.

However, these two approaches—data extraction and data synthesis—can be combined to generate new language data and significantly enhance the data production capabilities for languages like Livonian.

5 Choosing speech synthesis and recognition

Analyzing technologies that can be most beneficial for Livonian from the perspective of obtaining data and best serving the community, it has been concluded that speech recognition may actually be the most advantageous (Ernštreits, Fišel et al. 2022, 31) at this stage.

As mentioned, Livonian corpus building is nearing the point where no major written sources remain to be added, but there are extensive Livonian speech recordings in several archives and private collections. The contents of these recordings could significantly expand the Livonian language corpus, but to access information beyond phonetic features (to get to vocabulary, morphology, syntax, etc.), these recordings need to be transcribed.

Currently, the only way to achieve this is through the involvement of fully proficient speakers, which is neither viable nor effective. Speech recognition, on the other hand, would significantly accelerate this process. Moreover, it would be a powerful tool for rapidly expanding the Livonian information space (and corpus) by enabling speakers to record and transcribe large volumes of information in Livonian. However, speech recognition is quite challenging to develop and requires a significant amount of data for training.

Conversely, speech synthesis—a technology that is much easier to develop compared to speech recognition—is another technology urgently needed by the Livonian community. The need for speech synthesis arises from the fact that a natural Livonian-speaking environment no longer exists, and most of the data available to the general public and those learning the language is in written form. This means that Livonian speakers and learners primarily read Livonian rather than hear it.

To address this issue, from 2022 to 2024, the Livonian lexicographic database (LT) has been supplemented with audio recordings of lemmas and example sentences spoken by contemporary Livonian speakers, giving users the opportunity to hear Livonian. Thus, the collection of audio data necessary for developing speech technologies has already begun.

Successful voice synthesis would give the general public, and especially the Livonian community, the opportunity to hear all digitized Livonian texts immediately, although having golden-standard level voice recordings should clearly be prefferable.

Speech synthesis would also greatly expand the Livonian audio information space, offering the community opportunities to create content or even develop audio and video media, thereby allowing

Livonian to be encountered in new language domains. Additionally, speech synthesis could provide the conditions and data necessary for creating and training speech recognition technology.

6 How is it planned?

The approach proposed by the researchers of the UL Livonian Institute for developing speech recognition for Livonian involves four stages and is based on using existing data from the Livonian.tech database cluster (LT) and both alligned and non-aligned speech recordings from various informants found in archives.

In the first stage, initial aligned datasets are created using pre-existing audio data from the Livonian.tech databases and the 'gold standard' natural speech data from contemporary Livonian speakers reading texts from written corpora. Subsequently, based on this data, speech synthesis is created.

In the second stage, voice cloning takes place, creating additional synthetic voices by using the natural speech data of previously recorded informants from archives. Initially, the focus is on informants who have been recorded more extensively and whose speech has also been transcribed and published as language samples.

In the third stage, Livonian written data is used to generate a synthetic voice corpus using all available voices (both contemporary voices and those from archive recordings), in order to obtain voice data from all existing written sources. This process effectively multiplies the corresponding data by the number of synthetic voices. In the fourth stage, both natural and synthesized speech data are used to train the speech recognition model.

In the final stage, voice data manipulations are planned, including techniques such as voice merging, adding disturbances based on recording quality and phonetic peculiarities, and using non-aligned data from both living and deceased speakers for training. This point is particularly important considering the varying time periods and quality of many Livonian recordings.

As the current project is relatively short (a little over a year), only the first stage will be fully completed, which involves creating speech synthesis (text to speech), along with some data multiplication and initial isolated speech recognition (speech to text) experiments. However, this will lay the groundwork for a continuation project, which will focus more specifically on speech recognition.

The tasks to be completed within this stage are:

1. Restructuring the existing aligned (text and voice) corpus and expanding it by using contemporary Livonian speakers;
2. Creating a standard Livonian pronunciation guide to be used for referencing and assessing data quality;
3. Creating speech synthesis using the highest quality “gold standard” data from living speakers participating in the creation of the aligned corpus (text and voice);
4. Expanding the aligned corpus by synthesizing available written text data using the voices of living speakers, thus multiplying the audio data;
5. Assessing the quality and making necessary corrections to improve the synthesized data;
6. Cloning voices of both contemporary speakers and those recorded earlier (for both “gold standard” example data and synthesized data) and building a speech synthesizer using altered “gold standard” data;
7. Conducting a data assessment of archive speaker data and making necessary updates;
8. Performing tests on speech recognition using all created and available datasets and reviewing preliminary findings;
9. Integrating speech synthesis into the Livonian.tech database cluster and making it available for other applications.

The project is expected to conclude with the first preliminary results of the experiments on the creation of speech recognition in February 2026.

7 Final notes

All the necessary preconditions exist for the proposed approach to be successful, at least to some extent. Even a partial acceleration of the transcription process would offer significant benefits for extracting data from audio sources. Gradually expanding the capabilities of Livonian in the sound environment would undoubtedly strengthen the language and make it more competitive in the long run.

Moreover, the methods and approaches developed during this research could serve as a foundation for other research teams seeking opportunities to apply speech technologies to extremely low-resource languages.

If methods for providing high-quality speech synthesis and recognition to languages with extremely limited data are discovered, this would be beneficial to all endangered languages, particularly those lacking domains for everyday language use (speech synthesis) or without a written tradition, where the language is primarily oral (speech recognition). Furthermore, this would provide the varied and abundant data essential for developing other technologies, such as machine translation, chatbots, caption generation, and other solutions powered by advancing AI technologies.

Acknowledgments

This study has been performed as part of the project "Improving access to a critically under-resourced language: AI-based approaches for producing and obtaining Livonian content" financed by the Recovery and Resilience Facility / NextGeneration EU (LU-BA-PA-2024/1-0056).

References

Druviete, Ina & Gunta Kļava. 2018. *The role of Livonian in Latvia from a sociolinguistic perspective. Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 9(2), 129–146. https://doi.org/10.12697/jeful.2018.9.2.06

Ernštreits, Valts, Fišel, Mark, Rikters, Matīss, Tomingas, Marili & Tuuli Tuisk. 2022. Language resources and tools for Livonian. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, *13*(1), 13–36. https://doi.org/10.12697/jeful.2022.13.1.01

Haboud, Marleen & Fernando Ortega. 2023. *Linguistic diversity endangered: the Waotededo language and the effects of intense contact. In: Eda Derhemi and Christopher Moseley (eds.). Endangered Languages in the 21st Century*. Routledge: London and New York.

LT = Ernštreits, Valts (ed. in chief), Vāvere, Signis, Viitso, Tiit-Rein, Damberg, Pētõr, Kurpniece, Milda, Kļava, Gunta, Balodis, Uldis, Tuisk, Tuuli, Kūla, Gita, Tomingas, Marili, Soosaar Sven-Erik, Sedláčková, Anna & Jurgenovskis, Toms. 2024. *Livonian language and culture resource platform "Livonian.tech"*. Riga: University of Latvia Livonian Institute. https://livonian.tech/

Rikters, Matīss, Tomingas, Marili, Tuisk, Tuuli, Ernštreits, Valts & Mark Fishel. 2022. *Machine Translation for Livonian: Catering to 20 Speakers. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,* Volume 2: Short Papers, pages 508–514 May 22-27. Available: https://aclanthology.org/2022.acl-short.55.pdf

Trosterud, Trond. 2009. *Developing Prototypes for Machine Translation between Two Sámi Languages. Proceedings of the 13th Annual Conference of the European Association of Machine Translation,* EAMT09. Allschwil: European Association for Machine Translation.

Wang, Luyi. 2024. *Artificial intelligence's role in the realm of endangered languages: Documentation and teaching. Applied and Computational Engineering, Proceedings of the 4th International Conference on Signal Processing and Machine Learning*, March 2024, 48(1):123-129. DOI: 10.54254/2755-2721/48/20241249

Partanen, Niko, Hämäläinen, Mika & Klooster Tiina. 2020. *Speech Recognition for Endangered and Extinct Samoyedic languages. Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics. Available: https://aclanthology.org/2020.paclic-1.60