Digitization Work at the Finno-Ugrian Society: Livonian Case Study

Niko Partanen
Finno-Ugrian Society
niko.partanen@helsinki.fi

Jack Rueter
University of Helsinki
Digital Humanities
Language Technology
jack.rueter@helsinki.fi

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

Abstract

This article discusses the recent digitization project of the Finno-Ugrian Society, using the work on Livonian publications, especially those from Seppo Suhonen's *Liivin kielen näytteitä* from 1975 as a case study. We start by contextualization and motivation for these undertakings, both from the point of view of the Finno-Ugrian Society and the University of Latvia Livonian Institute, and then describe the workflows we have developed and foresee for the next steps.

1 Introduction

In last years the Finno-Ugrian Society has systematically advanced their digitization program, with the goal of increasing the digital availability of the materials the Society has published. This paper outlines how the work has progressed, what types of questions have been addressed and which have been identified to still require solutions. We use as an example the Livonian materials recorded and published by Seppo Suhonen, narrated primarily by Pētõr Damberg (Suhonen, 1975). Other Livonian materials the Society has published are Setälä (1953) and Mägiste (2006). The aspect that distinguishes Suhonen's materials from the rest is that recordings were made and have been archived at the Institute for the Languages of Finland. Setälä's and Mägiste's publications are based on transcriptions made on the spot without recordings. The audio recordings open many new possibilities in available workflows that need to be discussed.

2 Context of the University of Latvia Livonian Institute

Compared to many other critically endangered languages, Livonian has been relatively well documented. Nevertheless, much of this documentation has historically been shaped by the academic interests of linguists, resulting in materials that primarily address scholarly audiences. Examples include textual publications and, in particular, lexicographic works dating back to the mid-19th century (e.g., Wiedemann, 1861; Kettunen, 1938), which relied heavily on phonetic transcription and were therefore largely inaccessible to the Livonian-speaking community. The first lexicographic collection written in the Livonian standard orthography did not appear until 1999 (Ernštreits).

Since its establishment in 2018, the University of Latvia Livonian Institute has been developing a suite of dual-purpose databases—serving both research and community needs—which encompass lexicographic and morphological data as well as the Livonian text corpus, all based on the contemporary Livonian orthography (see Ernštreits et al. 2024).

With the rapid expansion of the aforementioned Livonian database cluster over the past five years and the growing interest in Livonian language learning and use, the need for additional documentation has become increasingly evident. While major text collections published or compiled in the standard orthography—such as books, newspapers, and manuscripts—have already been incorporated into the corpus or are planned for inclusion in the near future, the question of how to effectively integrate other sources, such as materials published for academic purposes in phonetic transcription has come to the forefront.

The efficient utilization and resource-conscious normalization of such phonetic sources into the standard Livonian orthography is closely tied to the research presented in this article. In the broader context of developing future technologies serving both the Livonian community and linguistic research, this work is especially timely. A related project currently being implemented at the University of Latvia Livonian Institute focuses on creating an aligned speech corpus, which uses as its speech input texts from the written corpus—particularly those reflecting natural speech situations, such as folklore.

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 112–122

Consequently, the integration of transcriptions from existing audio recordings is highly relevant for the development of future speech technologies. These not only promise to expand opportunities for the use of Livonian but will also facilitate the extraction of additional linguistic data from the substantial number of recorded yet untranscribed Livonian audio materials.

3 Context of the Finno-Ugrian Society

The Finno-Ugrian Society has been publishing scientific materials, both research and language materials, on the Uralic languages since the Society was founded in 1883. The Society has also funded and coordinated large fieldwork material collections throughout the areas where the Uralic languages are spoken. These materials can be primarily found at the Archive of the Finno-Ugrian Society located at the National Archives of Finland. The Society, in a work that has continued to the present day, has been publishing these materials as edited text collections and dictionaries, and new research is continuously being published.

The contemporary demands and expectations toward the digital availability of these resources have led the Finno-Ugrian Society to develop and apply a digitization plan. Although the digitization work has for now primarily focused on published journals, recently work on text collections and dictionaries has also been initiated, and also the first digitization experiments have been conducted with the Society's archives.

Whereas the scholarly output is primarily meant for researchers, the situation is different with materials such as texts and dictionaries. These materials certainly have extensive and important research uses, but at the same time they are very important for contemporary language users and learners. We argue that it is necessary to combine to the digitization process steps which enhance the usability of these materials, and these actions ultimately align very closely with the needs of both the community members and researchers.

The goal of the Finno-Ugrian Society is not to republish these materials. These materials have already been processed, analyzed and edited by specialists of each language, and we would prefer to frame our current work more as enhancing the usability and accessibility of the already existing works, and not as creation of new publications as such. Of course these boundaries are blurry, and digital versions of the publications are inevitably distinct from the originals. There are situations where they need to be cited separately, and the researchers who were involved in the work with the digital versions also need to be acknowledged. Our stance can still be illustrated by delineations such that when we digitize these works and create digital versions, we refrain from additional tasks such as adding new translations. Tasks such as adding automatically a new normalized transcription layer or ensuring that all lexemes are in the morphological analyzers are more of enhancing background tasks than conducting entirely new research.

This work has not been done in a vacuum, but it connects to the earlier research. Rueter and Partanen (2019); Rueter (2024); Rueter et al. (2024) describe their work on Erzya and Moksha corpora and how they connect to the analyzers of these languages. However, the approach taken here is more extensive, and we aim to keep the connection intact between digitized resources and the later corpora constructed from them.

4 Livonian Case Study

The Livonian recordings carried out by Seppo Suhonen in 1971 in Tallinn and Riga form a large collection of Livonian speech data. A co-interviewer was Karl Kont. Suhonen returned to interview Pētõr Damberg in 1981, but these recordings will only be digitized by the late 2026. Besides Damberg, Suhonen recorded other individuals as well, and Damberg himself was recorded by Eduard Vääri and Unto Miettinen in 1965. These recordings are stored in the Tape Archive of the Finnish Language at the Institute for the Languages of Finland. Jantunen (2025, 9) estimates that Suhonen's recordings are all together approximately 51 hours.

The recordings transcribed and published in *Liivin kielen näytteitä* (Suhonen, 1975) contain in total 2 hours and 50 minutes of speech. The published transcription is displayed in Figure 1, located in the end for convenience. We argue that this is an extremely typical scenario with data on endangered languages: a small part of the material is processed in more detail than the rest. Also in this case Jantunen (2025, 9) describes having transcribed approximately 6 hours of the Suhonen's materials. This means that approximately 20 % of Suhonen's material has been transcribed.

This scenario is at the same time very promising and potentially highly rewarding in contempo-

rary technical landscape. The parts of the dataset that are more finely processed, be it in the form of transcriptions or annotations, can be used as a training data to model the process in question, and thereby the resulting model can be used to analyze the remaining data in comparable style. This way the current transcribed Suhonen's Livonian corpus could ideally be extended to whole 51 hours, which would have significant consequences to the general availability of spoken and transcribed Livonian.

It must also be noted that we are now discussing Livonian materials collected by a few individuals and stored in one language archive: naturally, the scope of all existing Livonian recordings from this time period, and containing speech from the same individuals, among them Damberg, is much larger.

5 Automatic Text Recognition

Automatic text recognition of texts written in the Finno-Ugric transcription system has been a large challenge in the field in the past. However, in last years especially the Transkribus platform (Kahle et al., 2017) has allowed researchers to easily transcribe materials following the transcription conventions they consider best, and then train text recognition models with this data, improving the accuracy rapidly in an iterative manner. At the same time, processing of handwritten documents has also progressed very rapidly (Partanen et al., 2022; Arkhipov et al., 2021; Lamb et al., 2022).

In the context of the Finno-Ugrian Society's Livonian materials, we often find a situation where the same material exists in handwritten, typed and published versions. In these instances our focus is in digitizing the published version, and we take as our starting point that this is the most carefully edited and the most useful version. We can make the information available about the other existing versions, but starting to digitize all of them and creating comparable version would already stray away toward entirely new publications, and is not the point nor the scope of the current work. The goal is not to reconstruct in detail all nuances of the earlier work, but improve the use of language resources that are not currently as accessible as they could be. The language data is in the focus of this work, not the actions of the earlier researchers.

When we create the text recognition models, it seems that Finno-Ugric transcription of Livonian is a category in which the same models are able to generalize up to some degree. However, each publication has small differences and idiosyncrasies that need to be individually addressed. The best Livonian models currently are trained with almost 200 000 transcribed words and reach the character error rate of 0,28 %. A page in Transkribus platform with recognized and manually corrected text is displayed in the Figure 2. When we want to process a new publication, we need to add enough pages to cover the new characters and the new variation, but in our experience this is a very painless and fast process.

Although Transkribus is not an open-source platform, the spirit and general approach of Transkribus maintainers and the READ Coop that manages the project has aligned well with our goals. Needless to say, one must also consider whether the proofread materials could be deposited in some other environment, so that even open-source text recognition tools could be trained, tested and evaluated with this data.

6 Layout Analysis and Tagging

This part of the process takes place partly before the text recognition, but we discuss it still at this point as adjustments to the layout are done usually after the text recognition, and at the same time part of the tagging is done for the existing text.

Layout analysis refers to the identification of the structures in the document pages. Text regions and text lines are examples of regions, and page number would be an example of a structurally tagged text line. In our approach this tagging is extended very far. We mark headers, descriptions, metadata sections and page numbers separately. This data can be used effectively when the corpus is created at the later steps.

There would be many ways to structure the data, but our goal for now has been to create a minimal structure needed to distinguish Livonian and Finnish elements, first of all. In some of the books discussed here every even and odd page has a different language, in which case we can simply use this information to distinguish the language. At times the texts and translations are on the same page, with possible multiple short texts per page. In these cases it is critical that both original text and translation parts have the same number of elements. Then we can match the Livonian text and translation automatically by the number of elements. Naturally, a more nuanced method could be envisioned, but this convention has worked well for us. There have

been individual cases where a Livonian paragraph is split into two paragraphs in the Finnish translation. In these cases the solution has been to insert a tag for the Finnish translation that tells that we have a non-corresponding paragraph break. Information about the existence of the paragraph break is thereby kept, and the digitized data remains as intact and coherent as possible.

One particular case comes from indentation. In some situations, indentation is distinct enough that we can build a small classifier for the line starting points on a page and identify which are indented. At times, instead of indentation, there is a small vertical space between paragraphs: then these should probably be in different text regions.

Hyphenation is another structural issue. When the hyphen is located at the end of the line, the word can possibly be just hyphenated in this position, or it can be a compound where the hyphen is supposed to occur following the used transcription standards. We have not marked these instances manually, so that hyphenated words where the hyphen is needed are marked distinctly and these hyphens can be retained for later processing.

7 Orthography Normalization

As different publications have used slightly different transcription systems, there is a need to unify these so that comparative searches can be done, and the material can be connected to contemporary language technology. We need to facilitate corpus entries for lexicon and morphological analyzers, and this cannot be done if the transcriptions are wildly different. At the same time we recognize that the transcriptions are often very detailed and may contain dialectal features that are important, but cannot be easily expressed in the literary language and contemporary orthography.

Thereby what we are looking for is sort of a middle way where the representation is brought as close to the orthography as possible, but leaves some wiggle room for original details in the transcription. This is necessarily a partly impressionistic goal. Partanen (2024) discussed this task in their study where Large Language Models were tested in transliteration of endangered Uralic languages, and also in this context the task was not only a transliteration, but toward a normalization as well. As we are also keeping the original transcriptions, no information is lost, and various transcription layers can be envisioned.

If all transcriptions in different sources are essentially phonemic, with additional phonetic features present, one can also envision a solution where the harmonized transcription would have the same phonemic representation in all of them. At the same time, this would not be very useful for the language community and it would remain very unusable from the point of view of language technology. Since the Livonian orthography is actually fairly phonemic, it does not seem reasonable to aim toward anything else.

The workflow we have constructed in the pilot project is that the transcription is automatically transformed toward the orthography with a rule-based Python script. The script is adjusted based on the feedback we receive from the experts at the University of Latvia Livonian Institute. The rules are slightly different for each publication, but the output should match as well as possible. Figure 3 illustrates the transformed text.

The evaluation of texts normalized from phonetic transcription demonstrated that the results were very close to those that could be achieved through manual transcription. While certain orthographic inconsistencies were observed—primarily related to compounding and to the morphological principles applied in Livonian orthography (e.g., *ītõ-kabāl* vs. *īdõkabāl* 'all the time; always'; *jetspēḑõn* vs. *jedspēḑõn* 'away'; and in several cases involving specific verb or noun types such as *tīedist* '[they] did know' vs. *tīedizt* '[they] knew', *taggist* vs. *taggizt* 'ones behind') — the overall output was remarkably close to a gold-standard normalization. This indicates that the process can significantly reduce the effort and resources required for such transcription tasks.

8 Corpus Creation

In the corpus creation phase we parse the Transkribus Page XML documents with Python through the Transkribus API. The layout structures and tagging described in the earlier section is used to retrieve the correct structure. The resulting corpus contains transcribed Livonian sentences and information about the matching Finnish translation at the paragraph level. At the moment it does not seem to be possible to join Livonian and Finnish automatically at the sentence level. Another option would be to match the lines by position, where the Livonian sentence would have as a translation the roughly corresponding lines or portion in the

Finnish translation. Sentences cannot be directly aligned as there are small editorial differences between the versions, and sentence punctuation in the Finnish does not always correspond perfectly to the Livonian punctuation. Naturally, the text collection has not been created originally with the perfect sentence level matching in mind, and this is just a feature of the material for which we are still deciding the best approach.

As mentioned above, the alignment is based simply on the number of paragraphs. Similarly in the whole book there is a fixed number of texts. The metadata that has been tagged is extracted, and can be accessed directly in the parsing phase. However, we have found it more convenient to store the metadata in a separate table, where it can be extended and clarified. We often have text-specific details, i.e. location as coordinates, which we in any case want to associate with each text, but we do not want to add them to the digitized work. In principle, any format would work for this additional metadata, as we can always merge it with the corpus using the number of the text as a shared field.

9 Forced Alignment

Forced alignment is not a task we have yet applied to the workflow, but aligning the text and audio is a critical phase that should be performed in the cases where the original recordings are available. The current approach is to manually segment a portion of at least some tens of minutes at the utterance level, so that there is some baseline data against which we can evaluate different alignment approaches. This initial work is presented in the Figure 4.

There have been recent experiments in using Montreal Forced Aligner (McAuliffe et al., 2017) within the Uralic-Amazonian collaboration that has been taking place between the Universities of Helsinki and Belem (Rueter and Partanen, 2025). The idea in this work has been to align utterances in Komi-Zyrian and Apurinã at the phoneme level, and the goal has been to test if the alignment model can be trained with this type of data and how good the results are. If the results are positive, we could expect that the same can also be done with other endangered languages with similar resources, a category into which Livonian also fits very well.

With forced alignment it is important to notice that there are at least two fairly different scenarios in which forced alignment can be used. Most typically, it seems, we are discussing a scenario where there is perfect matching with the transcription and the audio segment, and the task is to match every phoneme as accurately as possible. However, this is not what we want to do first with the Livonian materials, but we would be very happy to have it at a later stage.

The situation is very different when there is a long and edited transcription, which corresponds to the audio, but not perfectly. The mechanisms needed here are fairly different, and the model needs specific logic to react to situations where there is no match, or when there is some additional content that cannot be matched. Ideally, in these situations the matching would be done at the utterance level, as the transcription would probably be revised against the audio once it is coarsely aligned. However, it seems that there is less support for this type of fuzzier alignment than there is for phoneme level alignment. We need to investigate what kind of forced alignment tool would work the best in our initial scenario. The tools that focus to phoneme level alignment should be used when the transcription is already aligned at the utterance level, and ideally manually adjusted if needed.

10 Universal Dependencies

There are currently several Universal Dependencies treebanks available for Uralic languages, and among these are also minor Finnic languages. In recent years two treebanks have been published for Karelian (Pirinen, 2019), one for Veps, another for Tundra Nenets, and soon there will be one for Northern Mansi. This is definitely a domain where new progress would be very welcome.

Seppo Suhonen's materials discussed here could suit this type of development work very well, too. The recorded and transcribed texts, that are unproblematic from the point of copyright, would be well fitting for open projects such as the Universal Dependencies. However, there is certainly a need to take into account both spoken and written Livonian, and also older recordings and contemporary speech. The situation is thereby fairly similar to Komi treebanks, for example, where different varieties and genres have been accounted for (Partanen et al., 2018). Similarly, the spoken language treebanks have various questions unique to them, especially in how the speech-specific phenomena are annotated (Dobrovoljc, 2022).

11 Conclusion

As outlined above, it is possible to envision a pipeline where the Livonian transcriptions are aligned with the Finnish translation, aligned with the audio, possibly even on a phoneme or word level, normalized to the Livonian orthography, and analyzed with contemporary morpho-syntactic analyzer for Livonian. This type of resource would be useful in a wide variety of research tasks, but it would also serve the language community in very detailed applications, including searching across the corpus and using both text and audio versions in language learning and education.

At the same time this data would be useful in tasks such as training automatic speech recognition tools. Especially in the context where there is a large number of recordings from a few individuals, it seems realistic to reach a very high recognition accuracy with the current methods, as reported in similar scenarios half a decade ago by Partanen et al. (2020) and outlined for Livonian recently by Ernštreits (2024).

One particularly promising outcome of the successful normalization of Livonian texts documented in phonetic transcription would be the integration into Livonian databases of materials from the 1938 Livonian–German dictionary (Kettunen, 1938), which contains a substantial amount of linguistic data, especially lexemes, not yet represented in the current Livonian database cluster and providing valuable expansion of vocabulary acccessible for Livonian speakers and learners.

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).

The digitization project of the Finno-Ugrian Society described in the study has been done with the funding by the Finnish Association for Scholarly Publishing.

References

Alexandre Arkhipov, Anna Barinskaya, and Roman Shtefura. 2021. Using handwritten text recognition on bilingual Evenki-Russian manuscripts of Konstantin Rychkov. *Scripta & E-Scripta*, 21.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Valts Ernštreits. 2024. Towards the speech recognition for Livonian. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 76–80, Helsinki, Finland. Association for Computational Linguistics.

V. Ernštreits. 1999. *Lībiešu-latviešu-lībiešu vārdnīca*. Līvõ Kultūr Sidām, Rīga.

Valts Ernštreits, Signis Vāvere, Tiit-Rein Viitso, Pētõr Damberg, Milda Kurpniece, Gunta Kļava, Uldis Balodis, Tuuli Tuisk, Gita Kūla, Marili Tomingas, Sven-Erik Soosaar, Anna Sedláčková, and Toms Jurgenovskis. 2024. *Livonian Language and Culture Resource Platform "Livonian.tech"*. University of Latvia Livonian Institute, Riga.

Santra Jantunen. 2025. *Livonian Verbal Derivation: Inherited Characteristics and Contact-Induced Change*. Doctoral dissertation (article-based), University of Helsinki, Helsinki. Doctoral Programme in Language Studies.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.

L. Kettunen. 1938. *Livisches Wörterbuch*, volume 5 of *Lexica Societatis Fenno-Ugricae*. Finno-Ugrian Society, Helsinki.

William Lamb, Beatrice Alex, and Mark Sinclair. 2022. Handwriting recognition for Scottish Gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 60–70.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, volume 2017, pages 498–502.

Julius Mägiste. 2006. *Muistoja Liivinrannasta: Liivin kieltä Ruotsista*. Number 250 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki.

Niko Partanen. 2024. Using large language models to transliterate endangered Uralic languages. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 81–88.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW*

2018), November 2018, Brussels, Belgium, pages 126–132.

Niko Partanen, Rogier Blokland, Michael Rießler, and Jack Rueter. 2022. Transforming archived resources with language technology: From manuscripts to language documentation. In *The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference, Uppsala, Sweden, March 15-18, 2022*, pages 370–380. University of Oslo Library.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 523–533, Hanoi, Vietnam. Association for Computational Linguistics.

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Jack Rueter. 2024. On searchable Mordvin corpora at the Language Bank of Finland, EMERALD. *Journal of Data Mining & Digital Humanities*, (V. The contribution of corpora).

Jack Rueter, Olga Erina, and Nadezhda Kabaeva. 2024. On Erzya and Moksha corpora and analyzer development, ERME-PSLA 1950s. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 67–75.

Jack Rueter and Niko Partanen. 2019. On new text corpora for minority languages on the Helsinki korp.csc.fi server. In *Èlektronnaâ pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy. Ufa, 11–12 dekabrâ 2019 goda*, pages 32–36.

Jack Rueter and Niko Partanen. 2025. Language technology for the Uralic languages in an Amazonian context. *Dutkansearvvi dieđalaš áigečála*, 9(1):137–147.

E. N. Setälä. 1953. *Näytteitä liivin kielestä*. Number 106 in Mémoires de la Société Finno-Ougrienne. Finno-Ugrian Society, Helsinki.

Seppo Suhonen. 1975. *Liivin kielen näytteitä*. Number 5 in Castrenianumin toimitteita. Helsinki.

F. J. Wiedemann. 1861. Joh. Andreas Sjögren's Livische Grammatik nebst Sprachproben. In *Joh. Andreas Sjögren's Gesammelte Schriften. Band II. Teil I*. Kaiserlischen Akademie der Wissenschaften, St. Petersburg.

Nuotanveto

vadà kìedə̂D ə̄ttə̂ mūnda kę̣rD / mūndan ə̄ttə̂ pitkìmə̂D mūndan
at lī̆ttə̂mə̂D / no neì kuolīmsadà vīššadà mēttə̂rt pitkàD / vẹ̀i-
bə̂D võlda kìedə̂D ka // siZ kìedə̂D jẹvà sidàbə̂D vadà jūr rānda
pə̄l / ùondźə̂l ku īrgə̂B lə̄'də̂ mie'rrə̂ // vadàn ə̄ttə̂ ka / vadàn
um kuo'ť / vadà sū um / ja sə̄lə̂Z kus̄ um vadàn oùk tuoìZ tuťkà-
mə̂Z sie um vadà pierà // siZ vadàn ə̄ttə̂ tībə̂D.

nu vadà tībə̂D jūr / vadà tībə̂D jūš ja vadà sū immə̂r ə̄t
il'peìn kor̂kkə̂D agà lẹ̄də̂D / ja alìlə̂pedə̂n ə̄ttə̂ / vanàst vọ'l'ť́ə̂
kivìD / pie'rrə̂ pa'n'ť́ə̂ svinàD // se um las̄ pī'lə̂G vadà sū
vəldiž // vadà tībə̂D jūr sidìZ kìedə̂D / vọ'l'ť́ə̂ e'dáist kìedə̂D
ja / ta'ggist kìedə̂D / ī'də̂n vọ'l' kìer ī̆ttiZ tuoìzə̂n kìer vọ'l'
tuoìstiZ // e'dáiZ kìer / e'dáist kìedə̂də̂n vọ'l' ī̆ttiZ kìer /
ta'ggist kìedə̂də̂n tegìž tuoìstiZ vọ'l' kìer / neìku kìerə̂D
vo'l'ť́ə̂ si'zzə̂l-pēdə̂n / nu / kìerə̂D lekštə̂ si'zzə̂l-pēdə̂n vadà
sū pùol ja vadà tībə̂D jūšsə̂ vel vọ'l' / vọ'l'ť́ə̂ ìerə̂D / kìe-
də̂D tuťkamə̂è ìerə̂D vọ'l'ť́ə̂ / sìepieràst ku alg̃ə̂ kìedə̂D kìe-
rə̂gə̂D nẹ'r'ŕə̂ / siZ ìerə̂D laškìstə̂ kìerə̂m kìed- kìedə̂də̂n immə̂r
ī̆ttə̂-kabàl / mẹìťtiZ aš ìeridi äb_vọlks siZ / kìedə̂D lə̄'kstə̂
nẹ'r'ŕə̂ / ja nēdí / ne lə̄'ks tikkiž / mə̄'də̂ks järà.

no siZ ku lekštə̂ ni mie'rrə̂ / va'ddə̂ vied̀istə̂ kakš mìes-
tə̂ // kakš pušnikkə̂ / e'dáí mìeZ ja ta'ggi mìeZ / ku / sẹ̀i-
dist agà pūŕťtist sel'líiZ kùozə̂ kus̄ ni vọ'l' / mẹ̄list kus̄ ni
īrgə̂B ve'iǰjə̂ / siZ amà e'ź̆mə̂ks àigist vieťtà / mits̄ sīlda
um vieťtà // no siZ mẹ̄list neì / no ni'm kūš_šīlda vieťtà /
ēťtam sīń vadà // lieštàD ju ist ùoťtə̂ vä'ggi tevàs vie'tsə̂ //
si'kšpùol siZ ne lekštə̂ tẹ'vvə̂ vie'ddə̂ jemìń // sẹ'uvvə̂ mūnda
kę̣rD vọ'l'ť́ə̂ lieštàD neì aǐgàZ ku / iz_vẹì lọ̄jaks mitt̀ə̂i'l'

Figure 1: Example of the scanned page showing the Livonian transcription of Suhonen (1975, 6)

Nuotanveto

vadà kìedâᴅ āťtâ mūnda kẹrᴅ / mūndan āťtə piťkìməᴅ mūndan
aẗ līťtəməᴅ / no neì kᵘol̄msadà vīššadà mēťtârt piťkàᴅ / vẹì-
bâᴅ vǭlda kìedâᴅ ka // siz kìedâᴅ jẹvà sidàbâᴅ vadà jūr rānda
pā̂l / ùondž́âl ku īrgâʙ lā̂ʼdâ mᴵeʼrrâ // vadàn āttâ ka / vadàn
um kᵘoť / vadà sū um / ja sā̂lâz kuš um vadàn oùk tᵘoìz tuťkà-
mâz sᴵe um vadà pᴵerà // siz vadàn āťtâ tībâᴅ.
nu vadà tībâᴅ jūr / vadà tībâᴅ jūš ja vadà sū im̀mər aẗ
iʼĺpeìn koŕkkâᴅ agà lẹ̄dâᴅ / ja aĺlâpedân āťtâ / vanàst vọʼĺťťâ
kivìᴅ / pᴵeʼrrâ paʼńťťâ svinàᴅ // se um las‿pīʼləɢ vadà sū
vāldiž // vadà tībâᴅ jūr sidìz kìedâᴅ / vọʼĺťťâ eʼd́dist kìedâᴅ
ja / taʼggist kìedâᴅ / īʼdân vọʼĺ kìer īťtiz tᵘoìzân kìer vọʼĺ
tᵘoìstiz // eʼd́diz kìer / eʼd́dist kìedâdân voʼĺ īťtiz kìer /
taʼggist kìedâdân tegìž tᵘoìstiz vọʼĺ kìer / neìku kìerâᴅ
voʼĺťťâ siʼzzâl-pēdân / nu / kìerâᴅ lek̄štâ siʼzzâl-pēdân vadà
sū pùol ja vadà tībâᴅ jūššâ vel vọʼĺ / vọʼĺťťâ ìerâᴅ / kìe-
dâᴅ tuťkamâš ìerâᴅ vọʼĺťťâ / sìepᴵeràst ku al̄gâ kìedâᴅ kìe-
râgâᴅ nẹʼŕŕâ / siz ìerâᴅ laškìstâ kìerâm kìed- kìedâdân im̀mər
īťtâ-kabàl / mẹìťtiz aš̄ ìeridi äb‿vọl̄ks siz / kìedâᴅ lā̂ʼkstâ
nẹʼŕŕâ / ja nēd́i / ne lā̂ʼks tik̀kiž / mā̂ʼdâks järà.
no siz ku lek̄štâ ni mᴵeʼrrâ / vaʼddâ vᴵedìstâ kak̄š mìes-
tâ // kak̄š puš̄nik̀kâ / eʼd́di mìez ja taʼggi mìez / ku / sẹ̀i-
dist agà pūŕťtist seĺĺiz kùož́â kuš̄ ni vọʼĺ / mẹtlist kuš̄ ni
īrgâʙ veʼiǰǰâ / siz amà eʼž́mâks àigist vᴵeťtà // mit̄s sīlda
um vᴵeťtà // no siz mẹt̄list neì / no niʼm kūš‿šīlda vᴵeťtà /
ēťtam sīń vadà // lᴵeštàᴅ ju išt ùoťtâ väʼggi tevàš vᴵeʼtsâ //
siʼkšpùo] siz ne lek̄štâ tẹʼvvâ vᴵeʼddâ jemìń // sẹʼuvvâ mūnda
kẹrᴅ vọʼĺťťâ lᴵeštàᴅ neì aïgàz ku / iz‿vẹì lǭjaks mit̀tâ iʼĺ

Figure 2: Example of the text recognized Unicode text showing the Livonian transcription corresponding to the text in Suhonen (1975, 6)

vadā kīedõd ātõ mūnda kõrd, mūndan ātõ pitkīmõd mūndan a
t lītõmõd, no nei kuolmsadā vīžsadā mētõrt pitkād, võibõ
d vȱlda kīedõd ka. siz kīedõd jõvā sidābõd vadā jūr rānd
a pǟl, ūondžõl ku īrgõb lǟʾdõ mieʾrrõ. vadān āttõ ka, va
dān um kuoţ, vadā sū um, ja sǟlõz kus um vadān ouk tuoiz
tutkāmõz sie um vadā pierā. siz vadān ātõ tībõd.

nu vadā tībõd jūr, vadā tībõd jūs ja vadā sū immõr āt iʾ
ļpein koŗkõd agā lȭdõd, ja allõpeḑõn ātõ, vanāst vȯʾļţõ
kivīd, pieʾrrõ paʾņţõ svinād. se um laz pīʾlõg vadā sū v
āldiž. vadā tībõd jūr sidīz kīedõd, vȯʾļţõ eʾḑḑist kīedõ
d ja, taʾggist kīedõd, īʾdõn vȯʾļ kīer ītiz tuoizõn kīer
vȯʾļ tuoistiz. eʾḑḑiz kīer, eʾḑḑist kīedõdõn voʾļ ītiz
kīer, taʾggist kīedõdõn tegīž tuoistiz vȯʾļ kīer, neiku
kīerõd voʾļţõ siʾzzõl-pēdõn, nu, kīerõd lekštõ siʾzzõl-p
ēdõn vadā sū pūol ja vadā tībõd jūssõ vel vȯʾļ, vȯʾļţõ ī
erõd, kīedõd tutkamõs īerõd vȯʾļţõ, sīepierāst ku algõ k
īedõd kīerõgõd nõʾŗŗõ, siz īerõd laškīstõ kīerõm kīed- k
īedõdõn immõr ītõ-kabāl, mõitiz aš īeridi äb vȯlks siz,
kīedõd lǟʾkstõ nę̄ʾŗŗõ, ja nēḑi, ne lǟʾks tikkiž, mǟʾdõks
järā.

no siz ku lekštõ ni mieʾrrõ, vaʾddõ viedīstõ kakš mīestõ
. kakš pušnikkõ, eʾḑḑi mīez ja taʾggi mīez, ku, sõidist
agā pūŗtist seļļiz kūožõ kus ni vȯʾļ, mõtlist kus ni īrg
õb veʾijjõ, siz amā eʾžmõks āigist vietā. mits sīlda um
vietā. no siz mõtlist nei, no niʾm kūž šīlda vietā, ētam
sīņ vadā. liestād ju ist ūotõ vä'ggi tevās vieʾtsõ. siʾ
kšpūo] siz ne lekštõ tõʾvvõ vieʾddõ jemīņ. sõʾuvvõ mūnda

Figure 3: Example of the automatic orthography normalization, corresponding to the text in Suhonen (1975, 6)

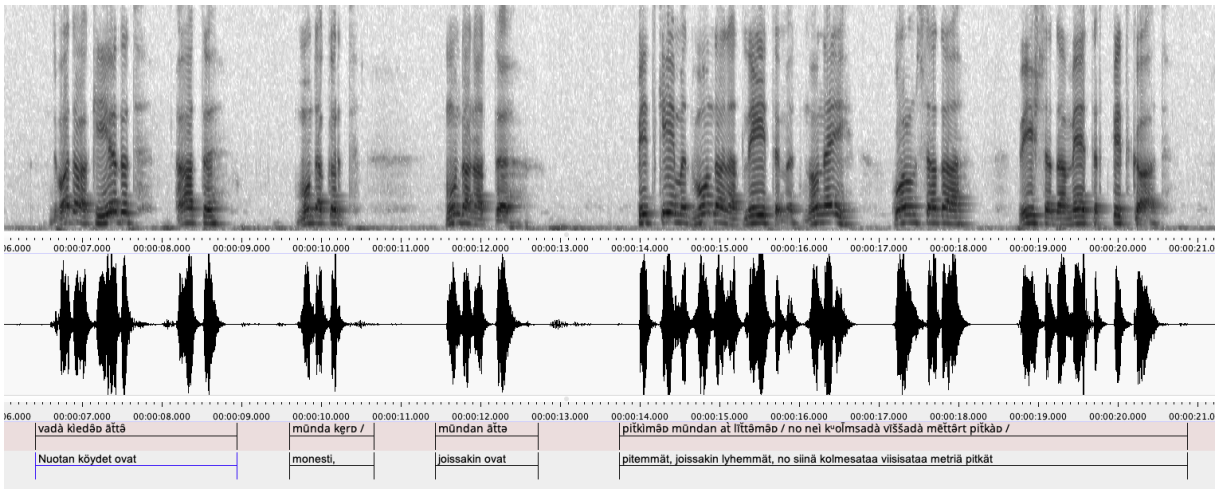

Figure 4: Example of the text recognized Unicode text showing the Livonian transcription corresponding to the text in Suhonen (1975, 6). This figure displays an experiment and the materials will likely be structured differently and managed in other more suitable environments.