Archive Infrastructure and Spoken Language Corpora for Saami Languages in Finland

Jouste, Marko

CEUR-WS.org
2022

Jouste , M , Mettovaara , J , Morottaja , P & Partanen , N 2022 , Archive Infrastructure and Spoken Language Corpora for Saami Languages in Finland . in K Berglund , M La Mela & I Zwart (eds) , Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) . CEUR Workshop Proceedings , vol. 3232 , CEUR-WS.org , Aachen , pp. 269-278 , Digital Humanities in the Nordic and Baltic Countries Conference , Uppsala , Sweden , 15/03/2022 .

http://hdl.handle.net/10138/350322

Archive Infrastructure and Spoken Language Corpora for Saami Languages in Finland

Marko Jouste[1], Jukka Mettovaara[1], Petter Morottaja[1] and Niko Partanen[2]

[1] *University of Oulu, Finland*
[2] *University of Helsinki, Finland*

Abstract

This study presents the results of an Aanaar Saami pilot project in the Saami Culture Archive, University of Oulu. The project has established a set of conventions to transcribe and annotate Aanaar Saami recordings in the archive's collection and created a mechanism through which grammatically annotated but anonymous versions can be imported to the Korp search interface in the Language Bank of Finland. The practices include wide use of Saami language technology, the use of Finnish computational research infrastructure, and they can be extended later to other Saami languages in the archive.

Keywords

Saami studies, Aanaar Saami, research infrastructure, language technology

1. Introduction

There are three Saami languages spoken in Finland: Aanaar (Inari) Saami, North Saami and Skolt Saami. There are multimedia materials archived there for these languages from over the past one hundred years. Although the materials were previously archived in different institutions around the country, currently various linguistic, folkloristic and ethnomusicological recordings have been organized in a corpus in the Saami Culture Archive of the University of Oulu. This does not mean that all Saami materials in Finland would be accessible in one location, and the neighboring countries have their own archival infrastructures as well. The Giellagas Institute for Saami Studies at the University of Oulu has a nation-wide responsibility to organize, introduce and provide Saami language and cultural studies as well as research at the academic level, and the work within the Saami Culture Archive directly serves these responsibilities. The archive staff also has the needed cultural and linguistic competence to work with the Saami materials.

Additionally, new materials are actively collected. In recent years, this has especially taken place in connection with language revitalization work. The purpose has been to support language teaching, planning and research (for information on language revitalization work in the Saami context, see Olthuis et al. [1], and the studies by Pasanen [2], [3]). The Aanaar Saami and Skolt Saami languages have undergone significant and successful revitalization efforts, and in both languages, there is an increasing demand for language materials that are suitable for second language learners and also fluent community members who want to study the language deeper.

From the point of view of modern language communities and researchers, the archived collections even the ones in the Saami Culture Archive can be considered as what is often called legacy data, referring to the materials that have been collected in a more remote past by researchers who have since passed away or are not themselves working with these materials any longer. The concept of legacy addresses the fact that these materials that are not originally ours are under our curation even today. Holton et al. [4] discuss that in many instances this has resulted in the materials and the information they contain being transferred from the indigenous communities to non-indigenous archives. Our case

The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: marko.jouste@oulu.fi (A. 1); jukka.mettovaara@oulu.fi (A. 2); petter.morottaja@oulu.fi (A. 3); niko.partanen@helsinki.fi (A. 4)
ORCID: 0000-0003-1971-054X (A. 1); 0000-0002-4727-6704 (A. 2); 0000-0001-8584-3880 (A. 4)


study provides a contrary case. These materials are hosted in a Saami archive, within the same institutional structure that also supports Saami education and other fields. It has been argued by Dobrin and Schwartz [5] that the main challenge in the use of legacy data is the way how they are connected to the intentions and ideas of the original creators of the materials, and they have a social history that must be understood when the data is used. We believe that the Saami Culture Archive is as a Saami organization in a strong position to evaluate how the materials can be used in modern times, as the perspective is not that of an outsider but an insider. Discussion about different aspects is certainly still on-going: O'Meara and Good already discussed a decade ago that social and legal concerns in making legacy materials more openly available have not been thoroughly studied [4], and this conversation is certainly on-going even today. We think that case studies like ours illustrate various ways of how wider accessibility can be achieved while ensuring the appropriateness of use. The best practices will develop in time when different approaches can be evaluated in a longer perspective. At the same time we think that the archives should have an active role in this process as well, since especially with the legacy data the archives are in the major role as curators of these materials.

The materials archived in the Saami Culture Archive are closely connected to other work done at the Giellagas Institute of the University of Oulu, including indigenous cultural work, teaching and research. As these materials originate from an indigenous culture, there are several specific questions that need to be addressed, especially with regard to access and cultural information therein. At the same time, there is value in making the materials as accessible as possible for the communities themselves, ideally online, so that they can be used to their full potential in language teaching, planning and maintenance. This connects to the idea of repurposing the archival data for language reclamation, as discussed by Lukaniec [7, 323], who points to numerous pedagogical tools, teaching aids, dictionaries and other resources which can be created from archival resources – once they are transformed into structured and normalized format that functions as a corpus.

We recognize that for the Saami materials we discuss here, the above-mentioned Saami communities are the foremost group that has the need for and interest in these resources, and for whom access must also be provided and with whom the conditions of use need to be negotiated. At the same time, we also want to find a solution whereby the materials can be used without significant barriers, especially when those barriers arise from the needs and practices of the majority culture. For example, we think that a language learner should be able to easily verify how a word is used from a corpus without the necessity of university-account-based authentication, which would quickly be the situation if the materials are accessible strictly for verified educational use.

In our solution, the materials are described and analysed by specialists in these languages and cultures, many of whom are also native speakers. We use a manual tagging method where personally identifiable and sensitive information is marked and can thereby be restricted or removed from subsequent derivations. These versions will be made available in the Language Bank of Finland[2]. In this way, parts of the archived materials can be used as language learning tools, or as example sentences, e.g., for dictionary infrastructure, while also taking into account the cultural integrity and sensitivity of the data. We use a national computational infrastructure, which is secure and allows for continuous refinement of the materials. What is specific in our approach is the extensive use of modern language technology, combined with a close connection to the needs of language communities.

In our study, we describe the workflow and evaluate the time spent in different work phases, thereby providing concrete numeric guidelines for similar future projects. We also analyse and test in detail the sensitive information tagging method which we have designed, with the goal of being able to estimate how much of the tagged content needs to be pseudonymized and anonymized. Similar tagging methods have been used before [7], but concrete estimations of how well they function are important and still scarce. We also provide accurate statistics about the size of the resulting corpus, and describe it in a manner that will directly benefit the new users, both in digital humanities and other fields of research.

The case study we report was a pilot project for Aanaar Saami that was conducted May–December 2021. This pilot was successful, and, in the future, similar work will be extended to the North Saami and Skolt Saami languages, thereby covering all Saami languages spoken in Finland. The model will certainly be developed further as the work advances, but, at this point, we want to create a solid foundation for later work.

[2] https://www.kielipankki.fi/language-bank

2. Aanaar Saami materials in the Giellagas Corpus of Spoken Saami

As we described in the introduction, during the past decades various recorded Saami materials have been relocated to the Saami Culture Archive. Under these circumstances, the work on them can be planned systematically and on a long-term basis from the perspective of one Saami organization. The oldest recordings in the archive are wax cylinders from 1913, and the majority originate after the 1950s when modern recording equipment started to become available. For a detailed description of the history of Aanaar Saami recordings, see the summary in Jouste's dissertation [8, 50–68].

The Aanaar Saami materials currently hosted in the archive total approximately 92 hours. The earlier archival sources of these materials are the archives of the Finno-Ugrian Society, the archive of the Finnish Literature Society and the Tape Archive of the Finnish Language in the Institute for the Languages of Finland. Besides this, new recordings have been done within the "Complementary Aanaar Saami Language Education" project (CASLE, for more information on the project, see [1]) program for adults, where the recording work was part of the course work. Most of the Aanaar Saami recordings are done by outsider researchers, but especially the recordings connected to language teaching represent a more collaborative and community-oriented approach.

Before the project started, all recordings were digitized but mainly untranscribed. Approximately 5% of the recordings had some level of transcription, half of which was in time aligned XML format, customarily used in linguistic research nowadays. From this starting point, it was deemed important that workflows be built through which the amount of systematically structured and annotated materials could be increased. A pilot project was designed to meet this need.

3. Pilot project

The pilot project was initiated in spring 2021 and it had four contributors. It is a part of a larger initiative to strengthen the infrastructure in the Saami Culture Archive. The goal was to establish a workflow from transcribed and annotated recordings into the services of the Language Bank of Finland. In a larger perspective, this would need to be connected to an updated archive management system in the Saami Culture Archive, and digital preservation processes that are also organized on a national level. For the time being, however, we have aimed at a solution that reaches concrete outcomes with resources that we already have and that interacts with working practices that already exist. Essentially, in the future when new recordings are transcribed for different purposes, integrating them to this workflow will be easy, and they can, if desired, be made accessible and searchable in ways that allow respectful treatment of personally identifiable information.

The system is very modular in the sense that the current data storage could be changed into something else with little effort, as this would essentially mean copying and transferring the file system. The transcriptions are stored in ELAN files, but this is an XML structure that can readily be transformed into other formats as desired. The main building blocks in this structure are utterance level time stamps and hierarchical annotations at word level. Although we currently provide access through the Korp search interface at Language Bank of Finland, we are not bound to one singular interface or collaborator. Naturally after having invested this much work in our solution, we would not start to implement radical changes without very careful thought, but in the long term we believe this modularity will come to play an important role.

The motive for the pilot project emerged from the specific new value and benefit that the spoken language corpus has. The significance of constructing a spoken language corpus for, e.g., revitalization and educational use can hardly be overemphasized. Speech is the primary form of language, and it differs fundamentally from written communication. For example, much of spoken phonological and lexical variation is oftentimes not present in the codified written standard, and even the ways of constructing sentences and phrases often differ from those of written conventions. A spoken corpus enables us to examine features of natural discourse, such as use of interjections, repetition, self-correction, discourse particles and code-switching, that are characteristically part of spoken communication. The information gained through studying these features helps both educators to teach the language and learners to learn the language in a more natural way. In addition, the contents of the corpus provide a significant source of oral tradition and cultural knowledge.

3.1. Data selection

The transcribed materials were initially selected to represent geographical variants of Aanaar Saami, consisting of recordings from both eastern and western areas of Aanaar Saami, and later on more transcription tasks were initiated by Giellagas Institute students of Aanaar Saami language and coordinated by the Saami Culture Archive. All of the ongoing transcription projects were joined in this pilot project in order to finish the transcription tasks and standardize the transcription conventions used. The transcribed recordings were chosen mainly from the older recordings spanning from 1913 to 1992. The recordings included adequately represent the main traditional speaking areas of Aanaar Saami and the temporal span of the archive recordings as a whole.

3.2. Data storage

The educational institutions in Finland can access the computing environments maintained and developed by CSC – IT Center for Science[3] free of charge. The Saami Culture Archive also uses several of the CSC products in the current project. As the services are built on the national level, institutions like our own can benefit from tools in which the basic security and maintenance are coordinated on a higher level. At the same time, we believe, the practices of one organization should be relatively easily transferable to other organizations within the context of Finland.

The context of the Saami Culture Archive is that of a medium-sized multimedia archive. The collection is large, but not in the millions of items, and although the collection keeps growing, it can still be handled effectively by the archival staff. Several individuals often need concurrent access to the same materials, and while the collections are used, adjacent resources are customarily created. For audio and video recordings this means that new transcriptions are added to the archive, and these improve the further usability of the resources for new archive users.

We have used CSC's Allas object storage system to store the edited materials and to share the files among the archive and project workers. The access can be controlled easily and everyone has access to the same files. The files from Allas can also be accessed in CSC's computing environment Puhti, which has been a large advantage. This setup is not perfect, as we would benefit from more granular version control and logging solutions, but as this is the recommended solution for the research data currently processed in Finnish institutions of higher education, we have wanted to adopt this as our solution, too. In spring 2022 we also participated in Allas service's user interview process, hoping that our experiences will eventually also be beneficial for the implementation of future versions.

3.3. Transcription conventions

The transcriptions were done with ELAN software [10]. The basic, initial transcription (non-standard transcription) is based on the orthographical rules of Aanaar Saami with some additions to mark phonemically relevant quantity distinctions (i.e. the half-long consonants and vowels, short diphthongs, and short consonant clusters). In the non-standard transcription tier, the goal was also to mark the phonemic and morphological variations resulting in non-standard word forms. It was also possible for the transcriber to include utterances characteristic to spoken conversations, like unfinished or interrupted words, laughter, coughing, and pauses, but this was not emphasized in the project. The unclear parts of the speech were marked, with possible guesses of unidentified words by the transcriber.

In the next step of the transcription, the non-standard transcription was standardized (tier Orthography). In this state, the phonemic diacritic markers, discourse analytical and unclear words were removed or simplified, resulting in a transcription that follows the orthographical rules of Aanaar Saami. The main idea for the orthographical tier is to make it easier to conduct word queries and use the orthography-based Aanaar Saami morphosyntactic analyser designed by Giellatekno group of UiT – The Arctic university of Norway[4]. The phonologically more accurate transcription, however, is saved and can be added to queries based on orthographical word forms and morphosyntactic descriptions.

[3] https://www.csc.fi/en/solutions-for-research
[4] https://giellatekno.uit.no/

In the initial state of the transcription, the recording was screened for possibly sensitive or private materials in three categories: 1) place names 2) personal names 3) sensitive information. These parts were flagged to be filtered or censored in later stages. Identifying place names and personal names is generally a mundane task for the transcriber. The third category, however, gives more room for interpretation. We took a relatively cautious stance with instructions for transcribers to flag anything that might feel like sensitive information, and the final judgment of flagging of possible sensitive information was tasked to core staff of the project.

3.4. Annotation conventions

The transcriptions are annotated with the methods presented by Gerstenberger et. al. [11]. They have previously been used in documentation of Pite Saami [12] and Komi [13]. The idea is that a rule-based morphosyntactic analyser is run over ELAN's transcription tier, and the resulting annotations are written directly to the ELAN file. The resulting tiers are then manually corrected. In the initial stage, the output of the analyser was inspected and analysed before manual correction was even considered. This allowed correcting typing errors and especially transcription specific conventions such as hesitation and unclear words, which are marked as described above. After the manual correction was done, an extensive list of remaining issues was created. This makes it possible to address these questions in following project phases.

Especially in lemmatization and morphological analysis, the quality was so high that even the most common unrecognized words were relatively rare. We used this as a justification to continue the manual correction in order to achieve a small gold corpus. It would be possible and desirable to envision workflows where the analyser is developed more in parallel with the transcription work, but, as our project team was small, resources were limited, and the goal of the work being a concrete annotated corpus, we followed the process described here.

Although the analyser returns information about Aanaar Saami morphology and syntax, the latter was not included in the manual correction phase after initial tests were done. The syntactic analysis was simply still at too rudimentary a stage with missing and wrong syntactic tags. This information was kept in the files, but it was too slow to correct it entirely by hand. More information about the time estimations also for this work phase is included in table 1.

The resulting annotations include first the lemma, which means the word form that is found as the headword in dictionaries. In the subordinate tiers, we store part-of-speech information and the morphological analysis. Part of speech contains only one tag, as there is only a small number of available categories. However, for the morphological analysis a string of feature tags is provided. These tags convey information such as tense, person, number, case, derivational elements and other grammatical-category information. The system is similar for all languages in the GiellaLT infrastructure and extensively documented[5]. This approach is very beneficial for the work on Aanaar Saami, as the same searches should be relatively easy to compare between different Saami varieties when annotated with a similar system. From this point of view, inclusion of more languages in the workflow described in this study is very important for the Saami Cultural Archive as well.

3.5. Anonymization

There are many situations where personal and identifiable information in the archived data should not be made public. Information about living persons falls under the European GDPR regulation, and as the materials discussed here have been recorded much earlier than the current legislation, the conventions used when the original recordings were made are often very different from modern practices. We are currently applying relatively strict measures of anonymization for the archival data, although situations may arise where the old age of the materials or possibility of making new agreements with individuals participating in the recordings could allow more open sharing. At the same time, we have to be careful when applying practices developed for the majority cultures in an indigenous context. When the participants in the recordings can still be reached, discussing and agreeing with them

[5] https://giellalt.github.io/lang-smn/

on new online distribution could be one way to improve the understanding on how the materials can be shared. Naturally, when the original work was done decades ago, it would have been impossible to even envision that someone would like to listen to the recording online, for example.

The flagging conventions to which the anonymization is based on was described above for the initial transcription. For this, we use an independent tier that is not connected to any other tier, so adjustments can be done at all times and on varying granularities, and they are entirely independent from changed annotation boundaries on other tiers as they are connected to time codes instead of annotations. To illustrate this, an annotation that tags some segment as containing identifiable information can span a very short segment or a longer span, depending on the situation and certainty of the annotator. When the actual filtering of this information is done, the segments that overlap the annotation even partially are hidden. This way the identifiable information is removed relatively roughly, so that we can confidently make the remaining texts openly available. At the same time, if there is a later need to refine any annotation, we can simply adjust, remove and add the tags in new versions. We do err on the side of caution but recognize that this kind of tagging scheme is never final and may change when more work is done on the corpus.

3.6. The Language Bank of Finland version

The version stored in the Language Bank of Finland will be made openly available in 2022, but as described, personally identifiable information has been filtered out. There are references to the original archival materials, and it is possible to request access to them through the Saami Culture Archive. The material is organized so that searches are possible by all the parameters we have annotated in the previous stages, with the exception of the cultural comments.

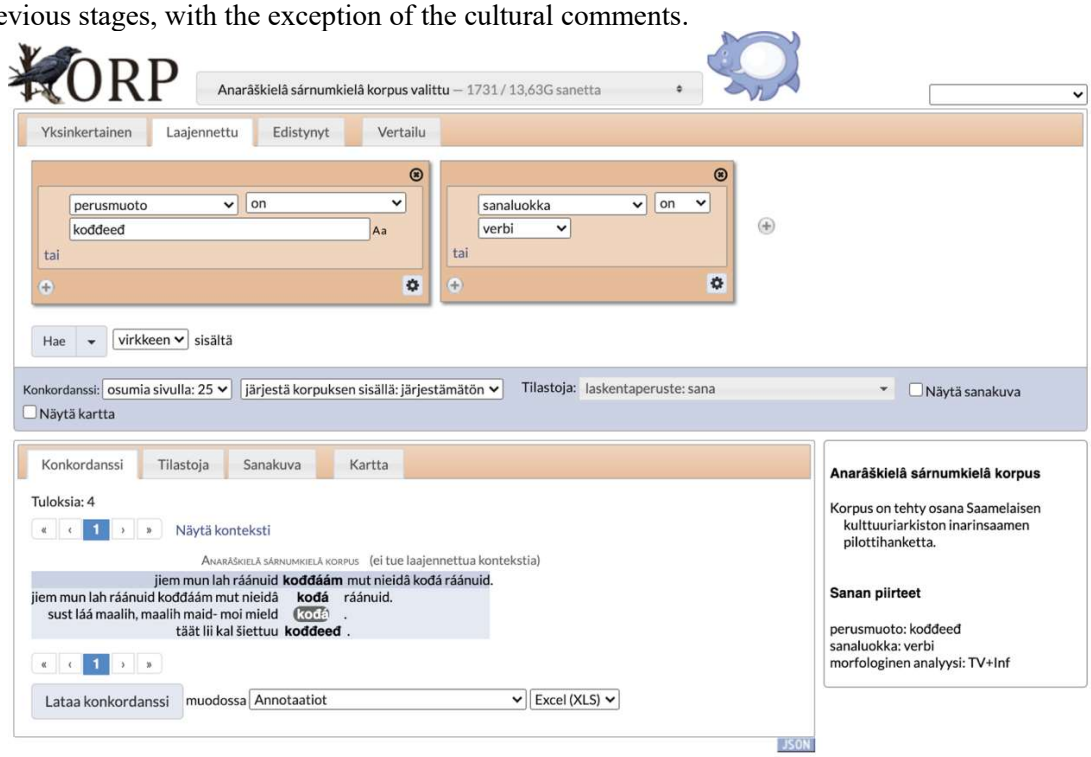

Figure 1: Upcoming Korp-interface view in the Language Bank of Finland[6]. The final version is intended to be as fully translated into Aanaar Saami as possible; note that the draft version in the screenshot also contains Finnish.

In the Korp interface [14] there are three possible search methods: Simple, Extended and Advanced. Figure 1 above displays the Extended query mode, where combinations of word sequences and annotations can be searched in a relatively simple way: horizontal elements are sequential words, and vertical elements represent annotations, which would allow searching for a specific part of speech and

[6] https://korp.csc.fi/

morphological category, for example. At the same time, the search query currently used always becomes visible in the Advanced view using CQP query language, or Corpus Query Protocol [15]. This is a well standardized and widely used search format, which many users may already be familiar with.

The search results are displayed in a typical keyword-in-context view, also known as KWIC. This displays all results vertically below one another with the matching word highlighted. By clicking individual words, it is possible to see further information about the line in question on the right side. This includes the original analysis, but also information about the pilot project, corpus and the details needed to request the original audio and ELAN files from the Saami Culture Archive.

If the segment has been flagged in a previous work phase, the content of those utterances will not be shown at all in this view. It could be possible to create a more fine-grained filtering here, for example, by taking into account that the flag for place names should only filter those tokens tagged as places. This, however, is already error prone, both from technical and human perspectives, as one would need to ascertain that the place is always tagged correctly and systematically. Again, this could also be verified by an additional manual check.

4. Evaluation of time spent in different tasks

Some of the work in the pilot project was technical and some linguistic. Evaluation of the time spent in different work phases is not trivial, but, in Table 1, we provide a careful evaluation based on our experience. For the technical work, we must emphasize that we used an already very functional software[7] which the authors have made available under an open license. The version used in our pilot project was adjusted to the tier structure in the project, but the foundation is the same with a rule-based analyser and processing the XML files with Python.

Initial adjustment of the script was thereby relatively fast, including approximately one week of work time. As the project progressed, we adjusted the software for the new needs that emerged. This involved adaptations for new transcription conventions and the processing of flagging information. We also added a new output option for VRT files used by the backend of the Korp interface. All this code is available in GitHub[8]. Although the CSC's Allas service functioned very well, there was an occasional need to check the files, organize them and verify in which steps they currently are to be found. All in all, this resulted in approximately one month of software-engineering related work. If the underlying analyser had been modified during the project, the amount of work would have probably increased by several months, and this would have also then involved highly specialized work where both linguistic and technical skills are crucial.

As shown in Table 1, the technical work is still minor when compared to the workload of even one hour of transcription. To set this into scale, the whole Aanaar Saami collection in the Saami Culture Archive is 92 hours, as described above. Applying the whole linguistic workflow to one hour of recording would amount to approximately 35 hours of work.

Table 1
Estimation of the time spent in different work phases

Work phase	Time (for 1 h of transcription)
Adjusting the script initially	One week
Modifying the script during the project	Two weeks
Maintaining the data repository	One week
Non-standard transcription (and flagging)	10 h
Orthographic transcription	5 h
Morphological checking	10 h
Syntactic checking	10 h

[7] https://github.com/langdoc/elan-fst
[8] https://github.com/nikopartanen/giellagas

The details vary from recording to recording, as the quality of audio and possible original transcription, the amount of overlapping speech, hesitations, code-switching and pauses all influence how much time is needed. Naturally, also the experience of the transcribers and annotators matters. In our pilot project, both transcribers, one native speaker and one advanced learner, were highly qualified in the language.

One figure that is currently missing concerns the speed of annotations if they were done entirely manually. We believe, however, that this would be many times slower than what we can currently present. The morphological and syntactic checking is also a task that could become faster if the analyser were systematically improved, but the improvement in speed would not be linear, as the checking would still be important and done for each token.

5. Further work

Transcribing work of Aanaar Saami materials will continue gradually by volunteer Aanaar Saami students who will benefit from the newly established conventions, which bring steadiness and continuity to the work. When untranscribed Aanaar Saami recordings are used and transcribed, for example, in BA and MA level theses, the resulting transcriptions will be archived and integrated into the corpus through the conventions described in this study.

The Saami Culture Archive will continue to maintain the infrastructure which has been set up during the pilot project. Further collaboration with the Language Bank of Finland and UiT – The Arctic University of Norway will also be fostered, as their services and tools are essential for our work. The Language Bank of Finland offers the corpus search interface in a setting that is already becoming familiar for different user groups, and the GiellaLT language technology is the backbone of the morphosyntactic analysis which our pipeline depends on. Nonetheless, even from the point of view of the Saami Culture Archive, there are several issues where local infrastructure could still be improved. Most of these components involve local data storage and metadata handling.

Currently, only very general metadata is included in the corpus, as the openly available version is anonymized. The Saami Cultural Archive, however, has much more detailed information about the recordings. At the moment, the metadata is not directly searchable, as the archival data management infrastructure still has to be fully established, and most of the information is stored in spreadsheets. One solution could be to use open-source archive repository management tools, such as AtoM[9]. At the same time, there is national Finnish infrastructure that is being developed, and the development of which the Saami Culture Archive is also actively monitoring.

Regardless of which management system is used with the repository, the metadata, ideally, and at least on some level, would be published so that the users could more easily judge whether there are materials useful for them in the archive. At the same time, however, the Language Bank of Finland version already gives the users a new kind of impression of the materials in the archive, which can make it better known what kind of content there is. Additionally, the metadata in the archive could be expanded, for example, based on the new transcriptions. For example, the fact that some recordings do contain sensitive information is something that should be added from the transcription level to the item description level.

When the corpus is available to the users, we should also build upon the feedback we receive. The current setup aims to be broadly useful and to have a low barrier of entry, but it is still a complex search system with its own query language. This means it is possible that the user instructions and even some parts of the pipeline could be redesigned so that they are in practice more useful. We may not need to build a complete user feedback system, but it is important to acknowledge that we do not yet know who all the users will be and what kind of backgrounds and needs they have, especially in a longer time frame.

As the pilot project has proven to work well, the same model will be used for organizing and expanding the spoken corpora of North Saami and Skolt Saami. Hopefully, in the near future, we will find a solution to add the audio parts in word queries. This has both technical and ethical considerations. First, hosting the audio files and linking the relevant segments is not trivial. Second, the audio would

[9] https://www.accesstomemory.org/en

be identifiable, and it should, somehow, need to account for the removed segments that contain personal information, as the intention is not to make those publicly available.

The materials could be extended, with the emphasis on modern language (e.g. Yle archives and possible brand new materials that will be recorded to be openly published from the start). In the future, the gathered data of parallel audio and text will advance the development of voice recognition and speech synthesis for Aanaar Saami. This way the workflow described in the paper would be even more effective.

6. Conclusion

In this paper, we have given a detailed description of the conventions of developing an Aanaar Saami spoken language corpus in the Saami Culture Archive in the University of Oulu. The pilot project provides a way to make archived materials easily accessible and searchable while respecting the privacy and anonymity of individual participants in the recordings. Language technology is used to make the annotation phase more efficient, but the whole data curation still depends primarily on the linguistic and cultural knowledge of the specialists working with the materials.

Although the chosen structures are influenced by the conventions often seen in linguistics and language documentation, we do believe that the work can also benefit other adjacent fields, especially in the context of digital humanities. The recordings and their transcriptions store a wealth of cultural information both about Aanaar Saami life and history in the 20th century and are also important documents about topics such as toponymy. The main audience of our work is the Aanaar Saami language community, primarily the language learners and users, but the conventions we have used are not restricted in their usability to just a few specific purposes.

7. References

[1] M-L. Olthuis, S. Kivelä, and T. Skutnabb-Kangas, Revitalising indigenous languages: How to recreate a lost generation. Multilingual matters, Bristol, 2013.

[2] A. Pasanen, Kuávsui já peeivičuovâ. 'Sarastus ja päivänvalo': Inarinsaamen kielen revitalisaatio. Uralica Helsingiensia 9, Finno-Ugrian Society, Helsinki, 2015.

[3] A. Pasanen, “This Work is Not for Pessimists”: Revitalization of Inari Sámi Language, in: L. Hinton, L. Huss and G. Roche (Eds.), The Routledge handbook of language revitalization, Routledge, New York, NY, 2018, pp. 364–372.

[4] G. Holton, Y. Wesley, P. Leonard and L. Pulsifer, Indigenous Peoples, Ethics, and Linguistic Data, in: A. L. Berez-Kroeker, B. McDonnell, E. Koller, L. B. Collister (Eds.), The Open Handbook of Linguistic Data Management, The MIT Press, 2022.

[5] L. M. Dobrin & S. Schwartz, The social lives of linguistic legacy materials. Language Documentation and Description 21 (2021) 1–36.

[6] C. O’Meara, and J. Good, Ethical issues in legacy language resources. Language & Communication 30.3 (2010) 162–170.

[7] M. Lukaniec, Managing Data from Archival Documentation for Language Reclamation, in: A. L. Berez-Kroeker, B. McDonnell, E. Koller, L. B. Collister (Eds), The Open Handbook of Linguistic Data Management, The MIT Press, 2022.

[8] M. Jouste, Tullâčalmaaš kirdâččij 'tulisilmillä lenteli' - Inarinsaamelainen 1900-luvun alun musiikkikulttuuri paikallisen perinteen ja ympäröivien kulttuurien vuorovaikutuksessa. [The One Who Flew with the Fire eyes - The Musical Culture of the Aanaar Sámi People in the Interaction of the Local Tradition and the Neighbouring Cultures]. Acta Universitatis Tamperensis 1650, Tampere University Press, 2011. urn:isbn:978-951-44-8551-0.

[9] N. Partanen, R. Blokland & M. Rießler, A pseudonymization method for language documentation corpora: an experiment with spoken Komi, in: 6th International Workshop on Computational Linguistics of Uralic Languages, January 10–11 2020, Vienna, Austria, 2020, pp. 1–8.

[10] ELAN (Version 6.3), Max Planck Institute for Psycholinguistics, Nijmegen, The Language Archive, 2022. URL: https://archive.mpi.nl/tla/elan.

[11] C. Gerstenberger, N. Partanen, M. Rießler & J. Wilbur, Instant Annotations: Applying NLP Methods to the Annotation of Spoken Language Documentation Corpora, in: International Workshop for Computational Linguistics of Uralic Languages. The Association for Computational Linguistics, 2017, pp. 25–36.
[12] J. Wilbur, ELAN as a search engine for hierarchically structured, tagged corpora, in: Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, 2019, pp. 90–103.
[13] J. Rueter, N. Partanen, M. Hämäläinen & T. Trosterud, Overview of open-source morphology development for the Komi-Zyrian language: Past and future, in: Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages, The Association for Computational Linguistics, 2021.
[14] L. Borin, M. Forsberg, & J. Roxendal, Korp – the corpus infrastructure of Språkbanken, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 474–478.
[15] S. Evert, The CQP query language tutorial, IMS Stuttgart, CWB version, 2, b90, 2005.