

Automatic subject indexing of Sámi oral history interviews with an LLM and thesaurus

September 23, 2025
TPDL NKOS workshop

Johannes Widegren
PhD student
Library and Information Science
Department of Cultural Sciences
Linnaeus University

My project

LNU & UU: *Artificial Intelligence as a Risk and Opportunity for the Authenticity of Archives*

- My thesis: Facing Recollection or Oblivion? On the Risks and Opportunities of Using AI to Improve Searchability in Sámi Archival Collections

Testing and analyzing organization, description and search in Sámi archival collections using AI tools together with Sámi newspaper Samefolket

- Side project: pipeline for transcription and indexing of oral history interviews



Tools used

- Google Colab
 - Free cloud service coding environment (Python)
- KB-Whisper (February 2025)
 - Based on the Whisper architecture
 - Trained on 50,000+ hours of Swedish speech
 - Available for free via HuggingFace
- Gemini 2.5 Pro (June 2025)
 - “State-of-the-art thinking model, capable of reasoning over complex problems in code, math, and STEM, as well as analyzing large datasets, codebases, and documents using long context.” – ai.google.dev
 - Available for free via Gemini API
- Thesaurus for SAMLA and ISOF (kulturnav.org)
 - Developed by SAMLA and ISOF based on Uppsala Folklore Archive – trans-national Scandinavian thesaurus

...might properly be
called a taxonomy
though...

“The Uppsala catalog was chosen as the starting point because it is very comprehensive. It also originally formed the basis for the catalog at the National Folklore Collection in Ireland. Thus, it has proven to be suitable across national borders. In the work of designing the digital thesaurus for SAMLA and ISOF, the thematic catalog from the Norwegian Folklore Archives has also been used as a comparison.” – samla.no

Buildings and settlements (#A.)
Communication and trade (#C.)
Historical Traditions (#K.)
Human life (#E.)
Language (#O.)
Magic and witchcraft (#I.)
Medicine (#G.)
Minorities (#P.)
Music (#M.)
Nature (#F.)
Nature mythology poetry (#J.)
Occupations, industries, and household (#B.)
Society (#D.)
Sports, games and play (#N.)
Time and timekeeping (#H.)
Various Genres (Genre Metadata) (#L.)

- 824 subject index terms

Interview collection

- Swedish Institute for Language and Folklore (ISOF)
- 14 interviews in Swedish mixed with South Sámi
 - Freely available online (isof.se)
- Recorded between 1953 and 2006
- Range between 16 mins and 1 h & 14 mins
- Collected by Israel Ruong, Olavi Korhonen et al.
- Life stories, Sámi folklore, traditional medicine, reindeer herding, cooking, crafts, schooling, place names





ULMA Ruong00249B

1953

Tärnaby, Tärna, Lpl

Israel Ruong

Maja Ruong, Okänd man, Einar Grundström, Rut och Ida Sofia Stinnerbom (76 år)

Sydsamiska/Svenska

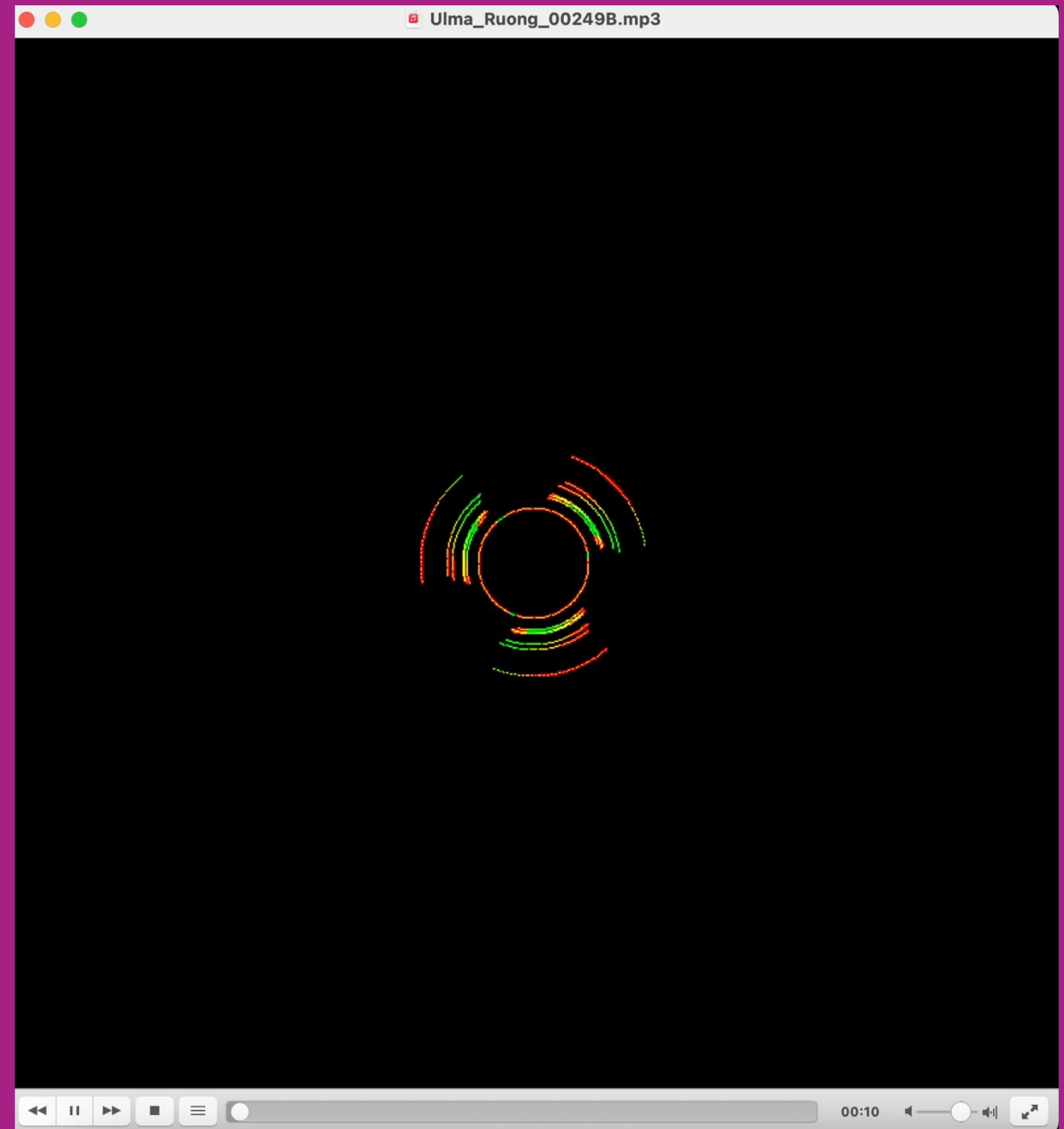
Maja sjunger en visa. Okänd sagesman berättar om samiska benämningar på renen: ålder, kön, färger och horn. Om tänkbara sagespersoner området. Einar berättar om skolan och tänkbara sagespersoner i i Tärnaby och dess omnejd. Rut och Ida Sofia berättar om ålderdomshemmet. Vidare om livet och sin uppväxt, uppväxtort och föräldrarna. Att hon inte arbetet med renar, men lärde sig samiska av sina föräldrar. Hon har en dotter som bor i Lövlund, Tärnaby. Om Skarmos där hon växte upp. Jon Jonsson Goruk var den första som bodde i Skarmos [Skalmodal?]. Pratar lite grann om jojk. Israel försöker få henne att jojka och jojkar. Hon har glömt alla jojkar.

Transcription

- KB-Whisper model is downloaded to device (or Colab)
- Took about 15 mins for 30 min audio file
- Outputs text + time stamp for audio chunks
- Result can be saved as .srt subtitle file
- Output is very decent overall, but...
 - Can't handle Sámi language of course

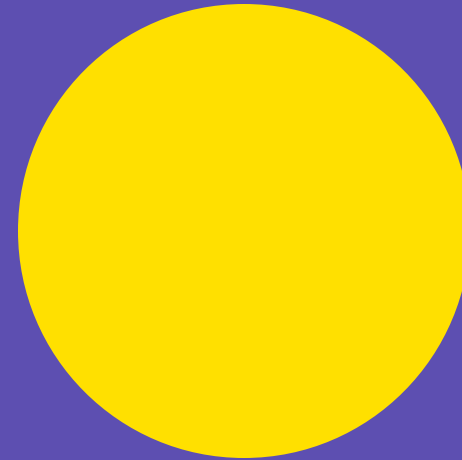


```
4  
00:00:10,000 --> 00:00:13,000  
Så där, nu kan du sjunga.  
  
5  
00:00:13,000 --> 00:00:19,000  
Strålande solljus från blånande höjder  
  
6  
00:00:19,000 --> 00:00:25,000  
Lärka som driver mot himmelens sky  
  
7  
00:00:25,000 --> 00:00:32,000  
Hälsen ohändig med jublande fröjder  
  
8
```



Indexing

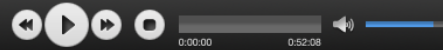
- Oral history indexing (OHI), several practices developed by different institutions (Lambert, 2023)
- I follow the leading example set by the Louie B. Nunn Center for Oral History at the University of Kentucky (Boyd, 2013)
- OHMS (Oral History Metadata Synchronizer) – open-source platform
- Metadata scheme: time stamp, partial transcript, segment title, subjects, keywords, and segment synopsis (+GPS and hyperlink data)





Interview with Martin Luther King, Jr., March 18, 1964

Louie B. Nunn Center for Oral History, University of Kentucky Libraries
Robert Penn Warren, Interviewer | 2002oh108_rpwcr003
[Who Speaks for the Negro? The Robert Penn Warren Civil Rights Oral History Project](#)



00:00:07 - Civil rights over generations

00:04:16 - Responsibility of the "negroes"

00:07:09 - Dr. King, a sell out?

[Play segment](#)



Partial Transcript: Dr. Kenneth Clark has remarked in print, that your appeal to many white people is because you lull them into some sense of security. . .

Segment Synopsis: Dr. King tries to explain his actions in regard to accusations from other "negro" leaders.

Keywords: "Pandering to the white man"; Accusations; Bitterness; Hatred; Love; Love in action; Morality; Resistance; Violence

Subjects: African Americans; Civil rights movements; Nonviolence; Whites

00:10:12 - Reparations for slavery

00:15:22 - Revolution and a single leader

00:19:43 - Integration of public schools

00:26:44 - Splitting the psyche of the "negro" / African or American?

00:32:44 - Slogan: "freedom now"

00:39:20 - The revolution

00:42:43 - Assaults in Harlem

Transcript ☒ Index

[Switch to Spanish](#)

Search this Index ▼

Keyword

Go

[View Rights Statement](#)

[View Usage Statement](#)

[Acknowledgment](#)

Series Link: [Who Speaks for the Negro? The Robert Penn Warren Civil Rights Oral History Project](#)

powered by
OHMS

© 2025
Louie B. Nunn Center for Oral History
University of Kentucky Libraries

Indexing

Text from .srt file is passed to Gemini API

System instruction: “You are an archivist tasked with improving the searchability in transcribed interviews by subject indexing segments of interviews with subject headings from a controlled vocabulary.”

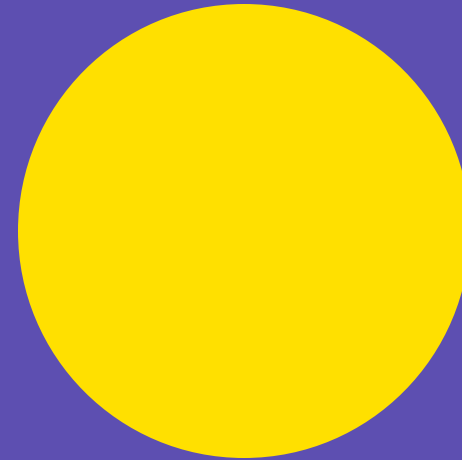
Temperature = 0

Thinking budget = 32768 (max)



Tasks

1. Split the interview into thematic segments
2. Specify the start time of each segment
3. Mark segments which are unclear or seem to be in a language other than Swedish
4. Give each segment a brief title
5. Summarize each segment
6. Describe each segment with free keywords
7. Subject index each segment using terms from a controlled vocabulary

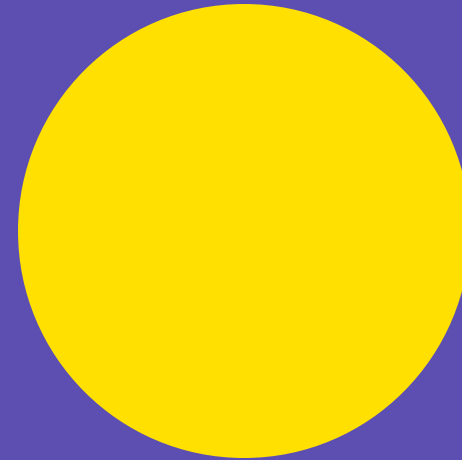


Prompt

“Split the following interview into thematic segments, where each segment deals with a defined topic.

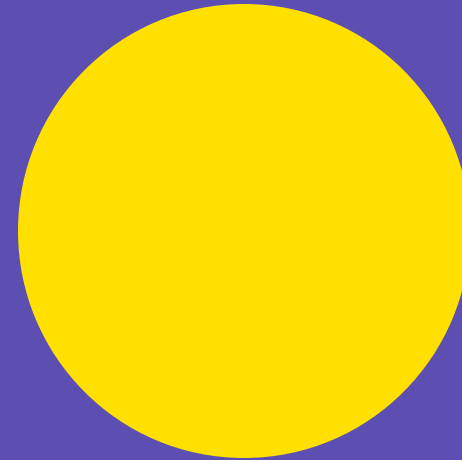
Then go through the segments one by one and follow these steps:

1. Specify the exact start time for each segment in hours, minutes and seconds according to the transcription
2. Give the segment a short title, up to 10 words. If a segment is unclear or cannot be understood, give it only the title [unclear/other language] and nothing else



Prompt

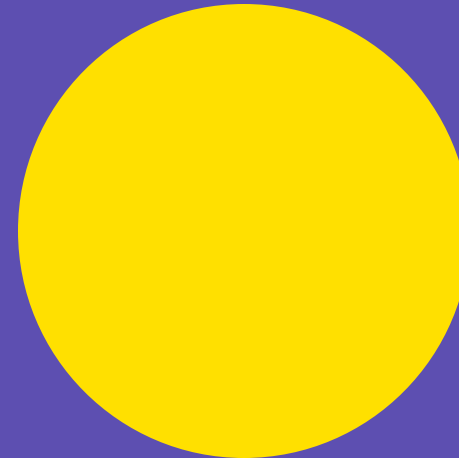
3. Summarize the segment with up to 100 words. If the segment has been given the title [unclear/other language], do not summarize it
4. Describe the segment with up to 10 keywords. If the segment has been given the title [unclear/other language], do not assign keywords
5. Describe the segment with up to 10 subject terms from the controlled vocabulary below. If the segment has been given the title [unclear/other language], do not assign subject headings



Prompt

Each subject term is followed by its superordinate term, e.g. Birds (#F.IV.2.) < Wild animals (#F.IV.) means Wild animals is the superordinate term for Birds

- Use specific subject terms that correspond as closely as possible to the content of the work.
- Normally, at least 20 percent of the section should address a topic for a subject term to be applicable.
- Normally, both a narrower and a broader subject term are not used for the same work. However, if the work contains at least 20 percent on both the narrower and the broader topic, then both subject terms may be used.
- You may only use subject terms from the controlled vocabulary. State them exactly as they appear in the controlled vocabulary.



Prompt

Format the result as a CSV file with five columns and semicolon as the separator.

The columns should be named Starttid, Titel, Sammanfattning, Nyckelord, Amnesord.

Here is an example:

{example}

Here comes the transcription:

{transcription}

Controlled vocabulary:

{vocabulary}



Result



00:00:00	Sound test and song about the homeland	The recording begins with a sound test, followed by a person singing a song. The lyrics describe natural scenery such as sunlight, larks, forest winds, and flowers, and concern love for the homeland, childhood, and youth.
00:02:08	[unclear/other language]	
00:04:23	Sámi terms for reindeer by sex and age	The conversation is about different Sámi terms for reindeer based on their sex and age. They discuss terms such as sarva for a reindeer bull, renoxe, girka and altan for a cow, as well as oängla for a one-year-old reindeer. They go through designations for reindeer at different ages, from one to four years.
00:05:46	Sámi terms for reindeer by color	The discussion continues with Sámi terms for reindeer, now focused on their color. Different colors and patterns are described with specific terms, such as klos for white reindeer, kölle for gray, bläse for spotted (especially around the nose), and chips for black reindeer.
00:06:33	Sámi terms for reindeer antlers	The conversation shifts to terms for reindeer antlers. One participant mentions that there are specific terms but that they do not know them all. Examples such as karatjorvi and ramsor (long, curved antlers) are given. It is suggested that the use of these terms may vary geographically.
00:07:26	Reindeer diet and different types of lichen	The discussion is about what reindeer eat. They speak about lichen, with specific names such as bushvia and jordgvinet (lichen that grows on trees). Summer food is also mentioned, mainly grasses called kräsi. However, the participant cannot specify the names of the different grass species.
00:08:26	Identification of elder knowledge holders of Sámi culture	The conversation focuses on identifying older people in the area who can tell about old times and Sámi culture. Brita Mårtens son, who lived her whole life in the mountains with reindeer herding, is mentioned as an important knowledge holder. Other people are also identified as potential interviewees. The importance of documenting their knowledge before it disappears is emphasized.
00:11:41	Short sound test and technical adjustment	A short technical break in the recording where a sound test is carried out with counting (“one, two, three”) to adjust the equipment.
00:11:51	Conversation about spring, birdwatching, and weather	The conversation begins with comments about the late spring weather. A teacher tells about a school outing to observe birds and signs of spring. He lists eleven bird species they observed, including dunnoek, brambling, and rough-legged buzzard. They then discuss the snowy winter and its causes.
00:13:47	Planning interviews with more local knowledge holders	The conversation returns to identifying and planning interviews with more elderly knowledge holders in the Tärnaby area. Several people are mentioned by name, such as Emma Larsson (yoik), Nils Axelsson, and Axel Nilsson Vinka (terrain terms). They discuss the people’s backgrounds, where they live, and how to get in contact with them.
00:17:12	Sound test	A short sound test is carried out with counting (“one, two, three”) to check the recording equipment.
00:17:22	Visit to the nursing home in Tärnaby	An interview begins at a nursing home in Tärnaby. The head expresses dissatisfaction with the home’s remote location from the village center, which makes it difficult for the elderly to visit the church and shops. She explains that the priest holds a service at the home every Saturday and that there are currently nine residents.
00:18:42	Conversation in Sámi about life in the past and taking bark	The interview shifts to a conversation, mainly in Sámi, with the residents Anders August Mattsson and Margareta Israelsdotter. They tell where they come from and speak about how life used to be. A central theme is the practice of taking bark from trees (boti), which they confirm Sámi people did in the past, possibly as food.
00:21:12	Conversation in Sámi about traditional cooking	The conversation continues in Sámi and revolves around traditional cooking and food handling. They speak about milking and boiling, and mention specific terms such as skalmåsen and björkunjaure, which suggest discussions about particular dishes or vessels. The process of dipping food into a pot is described.
00:24:30	Conversation in Sámi about colostrum and “skrovväxter”	The Sámi-language discussion deepens and is about specific foods and storage methods. They speak about nålmjölke (colostrum) and storage in a bortsebod. A central part of the conversation concerns skrovväxter, edible plants (possibly angelica), and how these were prepared and smelled in traditional Sámi food culture.
00:27:09	Conversation in Sámi about milk processing, bears, and wolves	The Sámi conversation touches on milk processing, with terms such as sötēs mjölke, which may refer to cheese-making. The discussion then turns to wild animals such as bears and wolves. The word jaurin (lake) is mentioned in connection with these animals, suggesting that they are sharing memories of encounters with predators in the landscape.
00:29:41	[unclear/other language]	

sound test, song, ballad, homeland, nature, childhood, memories, poetry	Songs and vocal music (#M.I.) < Music (#P.I.)/Singing (#M.I.23.) < Songs and vocal music (#M.I.)/Childhood memories (#E.X.1.) < Life story (E.X)
reindeer, reindeer herding, Sámi, terminology, language, animals, age, sex	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Mammals (#F.IV.1.) < Wild animals (F.IV)/Dialects (#O.I.) < Language (O)/Reindeer (#B.III.10.) < Livestock and Animal Husbandry (#B.III.)
reindeer, reindeer herding, Sámi, terminology, color, animals, description	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Mammals (#F.IV.1.) < Wild animals (F.IV)/Dialects (#O.I.) < Language (O)/Reindeer (#B.III.10.) < Livestock and Animal Husbandry (#B.III.)
reindeer antlers, reindeer, reindeer herding, Sámi, terminology, Västerbotten, dialect	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Mammals (#F.IV.1.) < Wild animals (F.IV)/Dialects (#O.I.) < Language (O)/Reindeer (#B.III.10.) < Livestock and Animal Husbandry (#B.III.)
reindeer, food, diet, reindeer lichen, lichens, grass, plants, reindeer herding	Mammals (#F.IV.1.) < Wild animals (F.IV)/Plants (#F.III.) < Nature (#F.)/Wild plants (#F.III.1.) < Plants (#F.III.)/Reindeer (#B.III.10.) < Livestock and Animal Husbandry (#B.III.)
knowledge holder, elderly, samisk kultur, reindeer herding, oral tradition, documentation, Våådalen, life history	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Life story (#E.X.) < Human life (E)/Old age (#E.VIII.3.) < Life cycle (E.VIII)/Historical Traditions (#K.)
sound test, technology, recording, adjustment, pause	
spring, birdwatching, school, excursion, birds, weather, snow, Tärnaby	Birds (#F.IV.2.) < Wild animals (F.IV)/The weather (#F.I.9.) < The sky and the weather (#F.I.)/Seasons (#H.I.4.) < The course of the year (#H.I.)/School (#D.VI.) < Society (#D.)
interview, planning, knowledge holder, yoik, samisk kultur, oral tradition, Tärnaby, place names	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Life story (#E.X.) < Human life (E)/Old age (#E.VIII.3.) < Life cycle (E.VIII)/Joik (#M.I.24.) < Songs and vocal music (M.I)/Place names (#O.III.) < Language (#O.)
sound test, technology, recording	
nursing home, elderly care, Tärnaby, church, service, elderly, community services	Old age (#E.VIII.3.) < Life cycle (E.VIII)/Elderly care (#E.VIII.3.c.) < Old age (#E.VIII.3.)/Church and congregation (#D.V.1.) < Religious communities (D.V)
Sámi, Sámi people, life history, bark, food tradition, Skalmodal, oral tradition, in the past	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Life story (#E.X.) < Human life (E)/Dialects (#O.I.) < Language (O)/Bark (#F.III.2.a.) < Trees (F.III.2)/Plants used as human food and drink (#F.III.4.) < Plants (#F.III.)
Sámi, cooking, tradition, milk, boiling, food culture, foodstuffs, recipes	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Cooking (#B.VIII.12.) < Household and ways of life (#B.VIII.)/Dialects (#O.I.) < Language (O)/Dairy farm (#B.VIII.3.) < Household and ways of life (#B.VIII.)
Sámi, cooking, colostrum, edible plants, angelica, storage, tradition, food culture	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Cooking (#B.VIII.12.) < Household and ways of life (#B.VIII.)/Wild plants (#F.III.1.) < Plants (#F.III.)/Plants used as human food and drink (#F.III.4.) < Plants (#F.III.)/Dialects (#O.I.) < Language (O)
Sámi, milk, cheese, bear, wolf, predators, reindeer herding, memories	Sami culture (#P.I.1.) < Minority culture (#P.I.)/Dairy farm (#B.VIII.3.) < Household and ways of life (#B.VIII.)/Mammals (#F.IV.1.) < Wild animals (F.IV)/Bear (#F.IV.1.a.) < Mammals (F.IV.1)/Wolf (#F.IV.1.b.) < Mammals (#F.IV.1.)/Dialects (#O.I.) < Language (O)

Task 1: Split each interview into thematic segments

Result:

Very good. Each segment contains a well-defined topic, some interviews have many segments (40+) and some fewer (10-)

Two segments mixed



Task 2: Specify the start time of each segment

Result:

A few time stamps slightly off, but only by a few seconds



Task 3: Mark segments which seem to be in a language other than Swedish

Result:

Of a total of 41 segments in Sámi -

14 correctly marked as [unclear/other language]

27 not identified, but – are these really thematic segments?

Accuracy: 34%



Task 4: Give each segment a brief title

Result:

Great.

(except for misclassified Sámi segments)

Example:

Conversation about different flies and their impact on reindeer

Task 5: Summarize each segment

Result:

Very good. Only lacking when subject expertise is needed.

(except for misclassified Sámi segments)

Example:

The conversation continues about trade and how goods were paid for, such as with silver. Reindeer meat is mentioned as a barter item. The discussion shifts to storage huts ("bor"). Two types are described: one on four legs (njalla or herebre) and one on a single post (julkiaburia). These were built high up to protect the contents from animals. A ladder (raitas), often a split log with carved notches, was used to climb up.

November -> novemberkaktus, nässla -> palettblad

Hen

Task 6: Describe each segment with free keywords

Result:

Ethnobotany

Toponymy

1 perhaps inaccurate

2 missed letters



Task 7: Subject index each segment using terms from a controlled vocabulary

Result:

Some subject terms too general – Historical traditions, Society

Lack of Sámi-specific subject terms – reindeer migration and Sámi village

Including misclassified Sámi segments: 89%

Excluding misclassified Sámi segments: **97%**

Example:

Spell and conjure (#I.X.1.b.) < Wise (#I.X.1.)/

Wise (#I.X.1.) < People with supernatural abilities (I.X)/

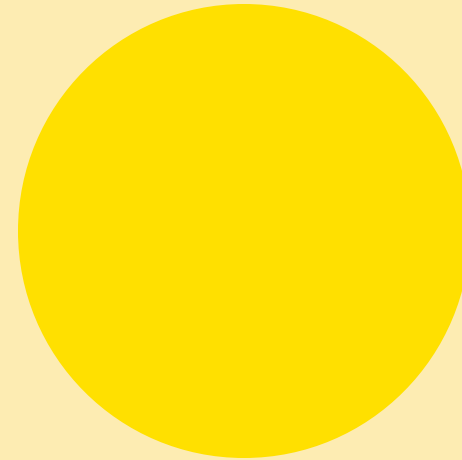
Cure with magical actions (#G.II.1.) < Cure (#G.II.)/

Personal historical legends (#K.VI.) < Historical Traditions (#K.)/

Blood (#E.I.3.) < The body (E.I)

Next steps

- Try an offline pipeline for processing sensitive interview data (new GPT-OSS?)
- Try incorporating more detailed instructions in the prompt – when are certain terms meant to be used?
- Try a larger controlled vocabulary with more Sámi-specific terms
- Output in xml-format, hook up with OHMS



References:

Boyd, D. (2013). OHMS: Enhancing Access to Oral History for Free. *The Oral History Review*, 40(1), 95–106.
<https://doi.org/10.1093/ohr/oht031>

Gemini models. (2025). <https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro>

Google Colaboratory. (2025). <https://colab.google/>

KB-Whisper Large. (2025). <https://huggingface.co/KBLab/kb-whisper-large>

Lambert, Douglas. “Oral History Indexing.” *The Oral History Review*, 50, no. 2 (2023): 169–92.
<https://doi.org/10.1080/00940798.2023.2235000>.

Project: Artificial Intelligence as a risk and opportunity for the authenticity of archives. (2025). <https://lnu.se/en/research/research-projects/project-artificial-intelligence-as-a-risk-and-opportunity-for-the-authenticity-of-archives/>

Samiska ljudfiler i Örnsköldsvik. (2022). <https://www.isof.se/folkminnen/vara-folkminnesarkiv/minoriteter/samiskt-material/det-samiska-i-ornskoldsvik/samiska-ljudfiler-i-ornskoldsvik>

Thesaurus for SAML A and ISOF (Universitetet i Bergen). (2025). <https://kulturnav.org/de3d9f18-1ab4-4421-b7fd-e5ce3d09bf81>

Linnæus University



**This work was partially supported by the Wallenberg
AI, Autonomous Systems and Software Program –
Humanities and Society (WASP-HS) funded by the
Marianne and Marcus Wallenberg Foundation and the
Marcus and Amalia Wallenberg Foundation**