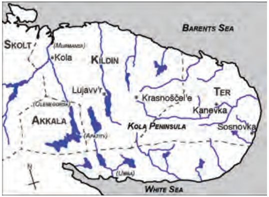
NEWS FROM SAAMI DOCUMENTARY LINGUISTICS

Michael Rießler

Ä′vv Sää′mi mu′zei Njauddâm

Budapest, 17th May 2012
Piliscsaba, 18th May 2012



INTRO

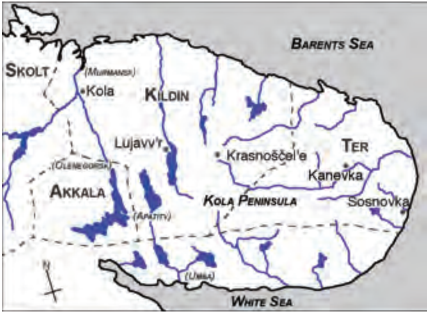
- Language documentation has a long tradition. Comprehensive language documentations of Saami languages were produced already in the beginning of the last century.
- Documentary linguistics evolved from traditional methodology in language documentation, but has become a linguistic sub-discipline of its own. The primary aim of the field is providing more and better data on the world's linguistic diversity for future research on and for endangered languages.
- The presentation focuses on the interfaces of documentary linguistics and language technology specifically from the perspective of applied research for endangered Saami languages.

BAKGROUND

- Skolt Saami museum in Neiden (currently)
- Universität Freiburg (July 2012)
- Foci: Kola Saami languages
  - Documentation
  - Description
  - Revitalization
- Other research:
  - Noun phrase syntax (history and typology) in northern European languages

Bodø, March 2010
from left: *M. Rießler, G. Lukin, D. Nathan*
(training in Saami documentary linguistics)

Kola Saami



- East-Saamic < Uralic, NW-Russia, Finland, (Norway)
  - Kildin Saami (< 800 speakers)
  - Skolt Saami (< 500 speakers)
  - Ter Saami (< 20 speakers)
  - Akkala Saami (< 5 speakers)

KOLA SAAMI DOCUMENTATION PROJECT (KSDP)

- DoBeS project (2005–2011)
  Documentation
- DFG project (2012–2015)
  Description
- Collaboration with Giellatekno (U Tromsø)
  Language technology

CONTENT

- Documentary linguistics
  - Definition
  - Short history of the field
- Kola Saami Documentation Project (KSDP)
  - Archive structure
  - Workflow practices
- Documentary linguistics as an applied discipline
  - Documentary linguistics vs. Language documentation
- Practical questions (language documentation and theoretical linguistics)
  - Metadata, equipment, data formats and archiving

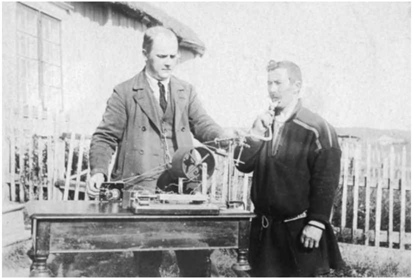
Documentary linguistics

Edison Phonograph (1877)

Documentary linguistics

- New and evolving field in linguistics
- Primarily concerned with language documentation (i.e. "a comprehensive, multi-facetted and multi-purpose record of linguistic practices characteristic of the investigated speech community"), which must be
  - comprehensive
  - multi-faceted
  - multi-purpose
- Two important purposes of language documentations are
  - being a data pool for theoretical research (in typology, anthropology, etc.)
  - being a data a pool for practical research (on revitalization, pedagogy, language technology, etc.)
- Documentary linguistics as primarily an applied discipline

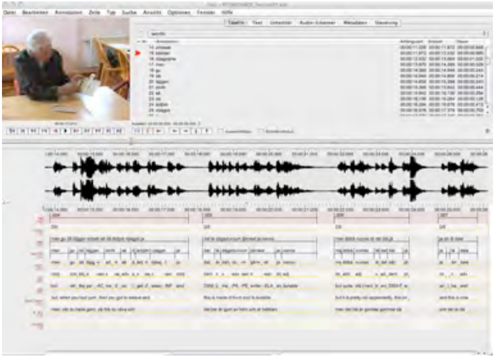
Documentary linguistics (e.g. Pite Saami)



Eliel Lagercrantz (1894–1973) documenting Pite Saami

- Classical documentary trilogy (“Boasian Trilogy”), e.g.
  - *Sprachlehre des Südlappischen* 1923.
  - *Wörterbuch des Südlappischen* 1926.
  - *West- und südlappische Texte* 1957 / *Texte aus den see-, nord-, west-, und südlappischen Dialekten* 1963.

Documentary linguistics (e.g. Pite Saami)



Pite Saami Documentation Project (PSDP), cf. *www2.hu-berlin.de/psdp/*

- “Language documentation is a lasting, multipurpose record of a language” (Himmelmann 2006:1).

SAAMI DOCUMENTARY LINGUISTICS

- Endangered languages – here Saami – must be documented before it is too late. However,
  - Only gathering language data does not result in useful documentation.
  - Storing language data is not the same as creating an archive.
  - Digitization alone is not sufficient preservation.

SAAMI DOCUMENTARY LINGUISTICS

- With this in mind, it becomes obvious that the biggest problems with traditional methodology in language documentation are:
  - the lack of information about the existence and relevance of recordings (missing metadata);
  - their unaccessability (missing annotations and/or physical unavailability);
  - as well as the question of long-term storage and preservations (use of open, non-proprietary formats and ensuring longevity of data carriers).

LANGUAGE
DOCUMENTATION
& CONSERVATION
journal

nflrc.hawaii.edu/ldc/

KOLA SAAMI DOCUMENTATION PROJECT (KSDP)

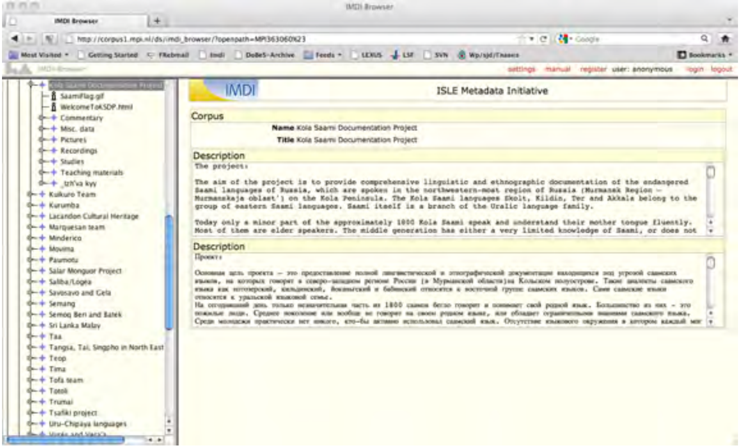
- Our main goals were (are):
  - documenting (record, annotate, archive) as much text recordings from different genres as possible, especially (formerly undocumented) conversations, spontaneous speech and procedurals
  - creating a corpus and other materials useful for future theoretical and practical research
  - working closely together with the Kola Saami communities

Elicitation session
Lujavv'r, August 2009
from left: *A. Antonova, M. Rießler*

Archive structure

- **Kola Saami Documentation Project** (introduction and general info)
- **Commentary** (access rights, bibliography, annotation conventions, project biography)
- **Misc. data** (fieldnotes, consultants' handwritten notes and wordlists, local newspaper articles, etc.)
- **Pictures**
- **Recordings**
  - —External
  - —KSDP
- **Studies** (papers and presentations on phonology, language sociology, etc.)
- **Teaching materials** (school grammar, text collection, posters, etc.)

ARCHIVE STRUCTURE



Corpus

Name Kola Saami Documentation Project

Title Kola Saami Documentation Project

Description

The project:

The aim of the project is to provide comprehensive linguistic and ethnographic documentation of the endangered Saami languages of Russia, which are spoken in the northwestern-most region of Russia (Murmansk Region – Murmanskaja oblast') on the Kola Peninsula. The Kola Saami languages Skolt, Kildin, Ter and Akkala belong to the group of eastern Saami languages. Saami itself is a branch of the Uralic language family.

Today only a minor part of the approximately 1800 Kola Saami speak and understand their mother tongue fluently. Most of them are elder speakers. The middle generation has either a very limited knowledge of Saami, or does not

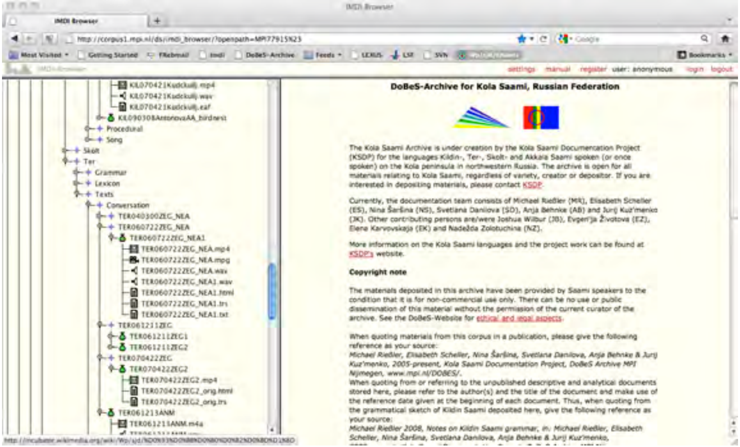
Description

Проект:

Основная цель проекта – это предоставление полной лингвистической и этнографической документации находящихся под угрозой саамских языков, на которых говорят в северо-западном регионе России (в Мурманской области)на Кольском полуострове. Такие диалекты саамского языка как нотозерский, кильдинский, йоканьгский и бабинский относятся к восточной группе саамских языков. Сами саамские языки относятся к уральской языковой семье.
На сегодняшний день только незначительная часть из 1800 саамов бегло говорит и понимает свой родной язык. Большинство из них – это пожилые люди. Среднее поколение или вообще не говорит на своем родном языке, или обладает ограниченными знаниями саамского языка. Среди молодежи практически нет никого, кто-бы активно использовал саамский язык. Отсутствие языкового окружения в котором каждый мог

- General archive structure
- `www.mpi.nl/dobes/`

ARCHIVE STRUCTURE



DoBeS-Archive for Kola Saami, Russian Federation

The Kola Saami Archive is under creation by the Kola Saami Documentation Project (KSDP) for the languages Kildin-, Ter-, Skolt- and Akkala Saami spoken (or once spoken) on the Kola peninsula in northwestern Russia. The archive is open for all materials relating to Kola Saami, regardless of variety, creator or depositor. If you are interested in depositing materials, please contact KSDP.

Currently, the documentation team consists of Michael Rießler (MR), Elisabeth Scheller (ES), Nina Šaršina (NS), Svetlana Danilova (SD), Anja Behnke (AB) and Jurij Kuz'menko (JK). Other contributing persons are/were Joshua Wilbur (JB), Evgenija Životova (EZ), Elena Karvovskaja (EK) and Nadežda Zolotuchina (NZ).

More information on the Kola Saami languages and the project work can be found at KSDP's website.

Copyright note

The materials deposited in this archive have been provided by Saami speakers to the condition that it is for non-commercial use only. There can be no use or public dissemination of this material without the permission of the current curator of the archive. See the DoBeS-Website for ethical and legal aspects.

When quoting materials from this corpus in a publication, please give the following reference as your source:
Michael Rießler, Elisabeth Scheller, Nina Šaršina, Svetlana Danilova, Anja Behnke & Jurij Kuz'menko, 2005-present, Kola Saami Documentation Project, DoBeS Archive MPI Nijmegen, www.mpi.nl/DOBES/.
When quoting from or referring to the unpublished descriptive and analytical documents stored here, please refer to the author(s) and the title of the document and make use of the reference date given at the beginning of each document. Thus, when quoting from the grammatical sketch of Kildin Saami deposited here, give the following reference as your source:
Michael Rießler 2008, Notes on Kildin Saami grammar, in: Michael Rießler, Elisabeth Scheller, Nina Šaršina, Svetlana Danilova, Anja Behnke & Jurij Kuz'menko,

- Example of a session with different files
- www.mpi.nl/dobes/

WORKFLOW PRACTICES

- Annotations include minimally catalogue metadata, preliminary orthographic transcriptions and a translation into Russian.
- First transcriptions and translations are done by native speaker assistants (teachers, other interested language workers).

Annotation, Lujavv'r, February 2006
from top: *A. Mozolevskaya, E. Scheller, N. Zolotuchina*

WORKFLOW PRACTICES

- Recording
  1. Recording on video and/or audio
  2. Written notes: catalogue metadata, preliminary annotations, elicitation stimuli, etc.
- Processing of data
  1. Digitizing video (if applicable)
  2. Preparing data for further processing (collecting session data, session cutting, labeling, storing on local server)
  3. Original transcription (by native speaker assistants) in a preliminary orthography and Russian translation
  4. Edited transcription (of selected recordings) in normative orthography, English translation, morphology, phonology, etc.
- Archiving

Documentary linguistics and applied research

Is *documentary linguistics*

- a new evolving subfield of applied linguistics?
  documentary linguistics $\geq$ *language documentation*

or is it

- just a (newly improved) empirical method for linguistic data collection?
  documentary linguistics $=$ *language documentation*

Annotation depth in KSDP

Still the same question…

How deeply should we annotate our Kola Saami corpus?

… like Lagercrantz

pu͔aᴰzª ji͔all vā͕r̜es̜t

Annotation depth in KSDP

Still the same question…

How deeply should we annotate our Kola Saami corpus?

… like contemporary typologists

puaz	jāll	vār'-es't
n	v	n
deer	live\3SG:PRS	fjell(WK)-LOC.SG

‘the reindeer lives on the fjell’

Annotation depth in KSDP

Our answer…

пуаз я̄лл ва̄ресьт

‘олень живет в тундре’

‘the reindeer lives on the fjell’

- Orthography and translation is adequate for the documentary linguist
- The theoretical linguist can do further analyses, annotations, descriptions, etc.
    - because sufficient phonological/grammatical and lexical description exists as the result of our and earlier research

DOCUMENTARY LINGUISTICS AND APPLIED RESEARCH

For the *documentation of Kola Saami* an orthographic representation is even preferable, because

- faster work provides more text annotations for —revitalization *and* research
- merging our spoken language data with written text corpora results in a relatively large corpus, which helps producing tools for — revitalization *and* research, e.g.
    - Digital infrastructure: spell-checkers, machine translation, teaching aids, etc.
    - Corpus linguistic tools: morphological and syntactic analysers, etc.
- North Saami (15,000 speakers, corpus of 500,000 words)
- *Boazu eallá váris*

Documentary linguistics

- We see *documentary linguistics* as a discipline concerned with language documentation (i.e. "a comprehensive, multi-facetted and multi-purpose record of linguistic practices characteristic of the investigated speech community")
- Two important purposes of language documentations are
  - being a data pool for theoretical research (in typology, anthropology, etc.)
  - being a data a pool for practical research (on revitalization, pedagogy, language technology, etc.)
- Documentary linguistics as primarily an applied discipline

Research ethics

- When working with speakers of endangered languages, you should always consider making available your data (and research results)
    - Archiving!
- If you believe in linguistics as a *science*, you should also consider making available your data (and research results) – and making your analyses provable (and falsifyable)
    - Archiving!

Legacy data

- The empirical linguist (of whatever framework) often also records data useful for the documentation of endangered languages.
- Important points to consider when creating legacy data:
    - Accessability (physically and intellectually)
    - Long-term storage and preservation (data formats and carriers)

METADATA

- Minimal cataloging metadata
  - Session label
  - Actors (consultant, collector, annotator, etc.)
  - Date and place
  - Research background
  - …
- Annotations
  - Elicitation materials (questionnaire)
  - Transcription, translation
  - Research notes
  - …

Data formats

- Audio
  - Preferably .wav
- Text (Metadata, Annotations)
  - Preferably open text formats (e.g. simple .txt)
  - Alternatively .pdf versions of your word document

Equipment

- Audio is normally most important
- Example of mid-range semi-professional equipment
    - Microphone Sony ECM-MS 957 (approx. € 250)
    - Recorder Edirol R-09 (approx. € 350)
- Software
    - ELAN (freeware)
    - Texteditor (freeware)