Annotation, archiving and presentation of Kildin Saami lexical data with LEXUS

Michael Rießler (Berlin) & Jacquelyn Ringersma (Nijmegen)
Annotation, archiving and presentation of Kildin Saami lexical data with LEXUS

Technical and lexicographic issues connected with the creation of a digital electronic dictionary for an endangered Saami language

• Kola Saami Documentation Project (DoBeS)

• LEXUS and ViCoS
Introduction: KSDP

The Kola Saami languages
Introduction: KSDP

The Kola Saami languages

**Kildin Saami**

- highly endangered (spoken by no more than 500 speakers – number of active speakers among the young generation is zero)
- inadequately documented and described
Introduction: KSDP

The Kola Saami Documentation Project (KSDP)

- part of the DoBeS initiative
- creating a comprehensive linguistic and sociolinguistic documentation of the Kola Saami languages

(among other concrete results):
- extensive and freely available corpus of spoken language data with annotations
- digital Kildin Saami dictionary
The DoBeS Initiative (Dokumentation bedrohter Sprachen) launched in 2000 by the VolkswagenFoundation (Germany)

Aim: Documentation of languages that are potentially in danger of becoming extinct
Two pillars

1. **Documentation (40 teams):**
   Recordings as primary material
   **Audio recordings:**
   present the languages as they are spoken
   **Video recordings:**
   information about the environment in which languages are spoken as well as non-verbal communication attitudes

2. **Archiving**
   One central online digital archive
   Formatting and encoding standards
   Long term preservation and persistent accessing
   Archiving and data enrichment tools (LAT suite)
LEXUS

Web based lexicon tool

Based on the ISO recommendations for linguistic resources

**LMF**: Linguistic Markup Framework (lexicon structure)
**DCR**: Data Category Registries (concept naming)

LMF and DCR provide a modular structure for content **interoperability** between (all aspects) of lexical resources
Heterogeneity of lexica: **Structure**
Structure of lexicon depends on language, linguistic theory, purpose

Heterogeneity of lexica: **Conceptual**
Large variation in linguistic concepts (attribute) and value naming

Heterogeneity of lexica: **Format**
Large variation in formats (XML, Toolbox, Chat, Word)

---

**Data interoperability problem**
cross lexica searches, merging, comparison

**Archive requirements**
Representation format (XML)
One archive exploitation framework
LEXUS is based on the ISO recommendations

It uses:

- **LMF**: default structure is LMF compliant
- **DCR**: Data Category Registries

LEXUS advantages:

- **Format**: export in standard xml format
- **Search**: one or across lexica
- **Merge**: Merging of lexica (in development)
- **Multi-media lexica**: Lexical encyclopedia
- **Linking**: Linking to the archive resources
- **Relations**: ViCoS extension
What can LEXUS do?

• Creation of lexica from scratch, import lexica from other formats
• User defined view of the information in the lexical entries
• Linking multi-media fragments to lexical entries
• Creation of links in images
• Link to resources within the digital archive (or other external web-based resources) – interaction with other archiving tools
• Create conceptual spaces with the ViCoS extension
• Search in and cross lexica
• Export to xml and toolbox format
• Export to IMDI package (in prep.)
Two dictionaries are under creation for KSDP:

**SaRuDic** (Encyclopedic dictionary of Kildin Saami):
- Kildin Saami – Russian (– English) dictionary
- Enriched with data from KSDP’s annotated corpus of spoken Kildin Saami (via import from KSDP’s Toolbox projects) as well as with data from other existing dictionaries
- Rich structure and content – including derivations and example sentences
- Including multi-media fragments

**RuSaDic** (Russian-Kildin wordlist):
- Preliminary stage (approx. 1000 entries)
- Limited structure
- Limited content (no multi-media)
- Index for SaRuDic
Kildin Saami dictionaries in LEXUS

RuSaDic structure

SaRuDic structure
Kildin Saami dictionaries in LEXUS

**RuSaDic structure**
- Russian headword
- Additional information about the headword (if necessary)
- Kildin Saami translation (in Cyrillic orthography)
- Kildin Saami translation (in Latin transliteration)

**SaRuDic structure**
- Saami headword
- Grammatical information (stem alternation, inflection class)
- Russian translation or explanation
- Saami example phrase(s)/sentence(s)
- Russian translation of example(s)
- Derived word forms (in alphabetical order)
  - Grammatical information (stem alternation, inflection class)
  - Russian translation or explanation
  - Saami example phrase(s)/sentence(s)
  - Russian translation of example(s)
Самъ́ (съ́мъ) |säm' (sâm')| I, 3 Saami person (masc., fem. and pl.); съ́мъ
пудъ́ |säm' pud'e| a Saami arrived; съ́мъ пудтёнъ́ |säm' pudt'en'| Saami (pl.)
arrived; съ́мъ къ́лл |säm' kil| Saami language; сённ ли съ́мъ олма |sönn lî
säm' olma| s/he is Saami
съ́мас |sämas| Saami (adv.), съ́рнэ съ́мас |särre sämä| to speak Saami
съ́мъленьч (съ́мълъя) |sämml'enč (sämml'a)| IV Saami person (masc., fem.
and pl.; originally a diminutive); сённ ли съ́мъленьч |sönn lî sämml'enč| s/he
is Saami; съ́йъ лёт съ́мълъя |sijj liev sämml'a| they are Saami
съ́мълуввэ (съ́мълувэ) |sämmluvve (sämmluve)| I to become like a Saami
 esp. according to one's way of life); сённ съ́мълуввэ |sönn sämmluvve| s/he
became like a Saami; сённ ялл съ́мас |sönn jäll sämä| s/he lives (like)
Saami (i.e. the Saami way)
Kildin Saami dictionaries in LEXUS

САММЬ (самь) | sāmm' (sām') | I, 3 Saami person (masc., fem. and pl.); cāmmь
пудэ | sāmm' pud'e | a Saami arrived; cāmmь пудтень | sām' pūdt'en' | Saami (pl.)
arrived; cāmmь килл | sām' kill | Saami language; sōnn lī cāmmь олма | sōnn lī
sām' olma | s/he is Saami

Saami (person) cāmmь | sāmm' | speak Saami
самье | sām' | n (masc., fem.
and pl.; orić)

Saami (person) cāmmьленч | sām'lenč | s/he
is Saami; сыий лев самьлья | sīj lēv sāmm'la | they are Saami
cамьльувэ (cамьльувэ) | sāmm'luve (sāmm'luve) | I to become like a Saami
(esp. according to one's way of life); sōnn cāmmьльувэ | sōnn sāmm'luve | s/he
became like a Saami; сөнн ялл самьас | sōnn jāll sāmas | s/he lives (like)
Saami (i.e. the Saami way)
Kildin Saami dictionaries in LEXUS

SaRuDic Aug08: Word list

**LEMMMA**

АББЫР

мэр

**MEDIA**

**TRANSLATION**

дождь

**PHRASES**

аббэр арр (мын, арт)

элааэ эрпэна
Why ViCoS?

Paper dictionaries’ limited usefulness in language maintenance & language revival (Manning et al., 2000)

Speech community may prefer *explicit* semantic links, possibly of a wide variety of types

→ introducing *Conceptual Spaces*, where concepts are
  → related to others
  → anchored in language
  → illustrated with multimedia

→ **Vi**sualizing **Co**nceptual **S**paces
Existing lexical resource is starting point

Words offer key access

ViCoS is technology to link words – and the associations they evoke – to other words

 Ease-of-use to empower community members to actively describe their language & culture and to learn from such resources resources targeted for human consumption

Resulting Conceptual Spaces = sort of informal ontology of fuzzily-defined concepts and relationships but where concept nodes are anchored in corresponding formal resources (lexica)
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Connecting RuSaDic and SaRuDic using ViCoS
Using ViCoS links can be made from ViCoS concepts to external URL sources (new feature in testing phase)
Unidecode: LEXUS and ViCoS are online tools
   → sometimes the correct display of characters is problematic

Offline versions of dictionaries
   → LEXUS and ViCoS need database (storage) so CD options are limited.
   → PDF paper output is foreseen (but excludes the video options)
   → Conversion from XML to formatted text is easy

Community Based continuation
   → current version: read only (login with username and password)
   → editing: restricted to selected users
   → data curation is an issue
Michael Rießler
Department for Northern European Studies, Humboldt-Universität zu Berlin/Germany
michael.riessler@staff.hu-berlin.de

Jacquelijn Ringersma
Max-Planck-Institute for Psycholinguistics, Nijmegen/The Netherlands
jacquelijn.ringersma@mpi.nl

DoBeS and KSDP are funded by the Volkswagen Foundation. Additional funding for the Kildin Saami dictionary project was provided by the Saami Council.

KSDP’s Kildin Saami dictionary project also benefits (or benefitted) from help provided by: Aleksandr Jur’jev, Aleksandra Antonova, Anastasija Nielsen, Anja Behnke, Elena Karvovskaja, Nina Afanas’jeva, Jevgen’ja Životova
References

• Documentation of Endangered Languages (DoBeS) [http://www.mpi.nl/dobes/]


• Kola Saami Documentation Project (KSDP) [http://www2.hu-berlin.de/ksdp/]


• Language Archiving Technology (LAT) [http://www.mpi.nl/lat/]