

# Korp in Tromsø, Norway

Trond Trosterud, Lene Antonsen,  
Chiara Argese, Ciprian-Virgil Gerstenberger

## **Giellatekno**

Saami Language Technology  
UiT The Arctic University of Norway

*Korp workshop 2018*  
*Stockholm, Sweden*



# Korp in Tromsø

**Instances**

**Features**

**Usage**

# Many minority languages $\Rightarrow$ many Korp instances

- korp  $\rightarrow$  North, Lule, South, Inari and Skolt Saami and parallel corpora
- f\_korp  $\rightarrow$  Kven, Meänkieli, Veps, Võro and parallel corpora
- u\_korp  $\rightarrow$  Komi-Zyrian, Komi-Permyak, Udmurt, Moksha, Erzya, Hill Mari, and Meadow Mari
- c\_korp  $\rightarrow$  Plains Cree
- tyv\_korp  $\rightarrow$  Tuvan Wikipedia

# North Saami

[North Saami texts](#) | [Lule Saami texts](#) | [South Saami texts](#) | [Aanaar Saami texts](#) | [Skolt Saami texts](#) | [More](#) ▾[Cite SIKOR](#) | [Davvisámi](#) | [Norsk](#) | [Suomi](#) | [English](#)

10 corpora selected (all) — 28.41M of 28.41M tokens

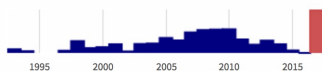
golladit

Simple

Extended

also as ☐ initial part ☐ final

KWIC: hits per page: 25 ▾



✓ Select all

✗ Select none

**Undated material**

Selected material: 4,575,949 tokens

of total: 4,575,949 tokens

- ▶ ☒ Newspapers (3)
- ☒ Administrative texts
- ☒ Religion texts
- ☒ Non-fiction texts
- ☒ Fiction texts
- ☒ Law texts
- ☒ Science texts
- ☒ An Account of the Sami

Show word picture ☐ show\_map

2,390,859 sentences in the selected corpora

# Lule Saami

[North Saami texts](#) | [Lule Saami texts](#) | [South Saami texts](#) | [Aanaar Saami texts](#) | [Skolt Saami texts](#) | [More](#) ▾[Cite SIKOR](#) | [Davvisámi](#) | [Norsk](#) | [Suomi](#) | [English](#)

Simple

Extended

also as ☐ initial part ☐ final

KWIC: hits per page: 25 ▾

7 corpora selected (all) — 1.19M of 1.19M tokens



✓ Select all

✗ Select none

**Undated material**

Selected material: 733,983 tokens  
of total: 733,983 tokens

☒ Administrative texts☒ Religion texts☒ Non-fiction texts☒ Fiction texts☒ Law texts☒ Newspaper texts☒ Science texts

110,886 sentences in the selected corpora

show\_map

# South Saami

[North Saami texts](#) | [Lule Saami texts](#) | [South Saami texts](#) | [Aanaar Saami texts](#) | [Skolt Saami texts](#) | [More](#) ⌵ [Cite SIKOR](#) [Davvisámi](#) | [Norsk](#) | [Suomi](#) | [English](#)



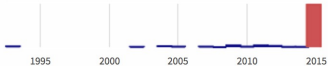
Simple

Extended

also as ☐ initial part ☐ final

KWIC:  hits per page: 25

7 corpora selected (all) — 1.10M of 1.10M tokens



✓ Select all

✗ Select none

**Undated material**

Selected material: 901,713 tokens  
of total: 901,713 tokens

☒ Administrative texts

☒ Religion texts

☒ Non-fiction texts

☒ Fiction texts

☒ Law texts

☒ Newspaper texts

☒ Science texts

114,125 sentences in the selected corpora

[show\\_map](#)

# Inari Saami

[North Saami texts](#) | [Lule Saami texts](#) | [South Saami texts](#) | [Aanaar Saami texts](#) | [Skolt Saami texts](#) | [More](#) 

[Cite SIKOR](#) [Dawisámi](#) [Norsk](#) [Suomi](#) [English](#)



5 corpora selected (all) — 1.58M of 1.58M tokens

golladit

### Simple

Extended

also as ☐ initial part ☐ final

KWIC: hits per page: 25 

✓ Select all

✕ Select none

**Undated material**

Selected material: 1,482,365 tokens

of total: 1,482,365 tokens

show\_map

☒ Administrative texts☒ Religion texts☒ Non-fiction texts☒ Newspaper texts☒ Science texts

149,691 sentences in the selected corpora

# Skolt Saami

[North Saami texts](#) | [Lule Saami texts](#) | [South Saami texts](#) | [Aanaar Saami texts](#) | [Skolt Saami texts](#) | [More](#) ▼[Cite SIKOR](#) | [Davvisámi](#) | [Norsk](#) | [Suomi](#) | [English](#)

Simple

Extended

also as ☐ initial part ☐ finalKWIC:  hits per page: 25 

5 corpora selected (all) — 214.88K of 214.88K tokens



✓ Select all

✗ Select none

☒ Administrative texts☒ Non-fiction texts☒ Law texts☒ Literary texts☒ Science texts

16,855 sentences in the selected corpora

**Undated material**

Selected material: 145,496 tokens

of total: 145,496 tokens

golladit

show\_map



# Next SIKOR update

- North Saami: 34 057 981 tokens
- Lule Saami: 1 304 787 tokens
- South Saami: 1 755 984 tokens
- Inari Saami: 1 798 969 tokens
- Skolt Saami: 22 8961 tokens

## Usr-frndly abbr pres

### Word attributes

part-of-speech: noun  
baseform: Ipml  
grammatical analysis: N.Prop.Sg.Nom  
dependency relation: SUP<sub>PL</sub>

[Show Dependency Tree](#)

noun  
proper  
noun  
singular  
nominative

sanan ominaisuudet

sanaluokka: substantiivi  
perusmuoto: lpmil  
morfologinen analyysi: N.Prop.Sg.Nom  
dependenssisuhde: SUBJ→

Näytä dependenssipuu

- substantiivi
- erisnimi
- yksikön
- nominatiivi

## ordattributt

ordklasse: substantiv  
grunnform: lpmil  
grammatisk analyse: N.Prop.Sg.Nom  
dependensrelasjon: SUB<sup>1s</sup>

### Vis dependenstre

substantiv  
egennavn  
entall  
nominativ

## sáni iešvuohta

sátneluohkká: substantiiva  
vuoddohápmi: lpmil  
sojahananalysa: N.Prop.Sg.Nom  
dependeansa: SUBJ

### Čájet dependeansa

- substantiiva
- namma
- ovttaidlogu
- nominatiiva

## NDS ⇒ Korp

← → ↺ 🏠  ⌵ 🔍 Search      ☰

 Inbox (38) - ciprian... ⚙️ Most Visited 🌐 Getting Started 📁 Useful 🌐 NDS

Neahttdigisánit [Home](#) [Click-in-text](#) [About](#) 

## Dictionaries

[North Saami → Finnish](#)[North Saami → Norwegian](#)

## WRITTEN VARIANT

[Standard \(áčđŋštž\)](#)[Tolerant \(incl. acđnstž\)](#)[Norwegian → North Saami](#)[Finnish → North Saami](#)[Other North Saami dictionaries](#)[Other dictionaries](#)

## North Saami (Standard) → Norwegian (⇌ Swap)

á č đ ŋ š t ž

á 

Search

Search texts

*materiála* is a possible form of ...*materiála**materiála* (subst.)○ (subst.) *materiale*

*materiála* [Texts →](#)  
*materiála* subst. ent. akkusativ  
*materiála* subst. ent. genitiv  
*materiála* subst. ent. nominativ

# NDS ⇒ Korp

Davvisámi teavsttat | Julevsámi teavsttat | Lullisámi teavsttat | Anársámi teavsttat | Nuortalaš teavsttat | Eanet

Čujuhitt SIKORii | Davvisámi | Norsk | Suomeksi | English



10 vällejuuvon čaokkáldaga (buot) — 28,41M/28,41M token

[lemma = "materiála"]

Oktageardán ohcan | Viiddiduvvon ohcan | CQP ohcan | Buohtastahttin

Oza

maiddái ☐ sáni álgun ☐ sáni loahppan ja ☐ beroškeahhtá unna/stuorra bustávain

KWIC:   Statistihkka:  ☐ Čájjet sátnegova ☐ show\_map

KWIC | Statistihkka | Sátnegovva | map

Boadus: 1 546

« « 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... » »

goto\_page of 62

Cealkkáčájeheapmi

ADMINISTRATIVE TEXTS (-)

|            |                                                              |                   |                                                                 |
|------------|--------------------------------------------------------------|-------------------|-----------------------------------------------------------------|
|            | Eará                                                         | <b>materiála</b>  | maid fágaoahpeaddji <b>juohká</b> .                             |
| ii         | dasa, leago oppalohkáii vejolaš čadahit dutkama vällejuuvon  | <b>materiála</b>  | vuodul.                                                         |
|            | Gánniha jearrat: Basttango hálddašit                         | <b>materiála</b>  | , leago dat mendo viiddis?                                      |
|            | Leago                                                        | <b>materiála</b>  | mendo njárbat ja vástidago dat gažaldagaide maid áiggun         |
|            |                                                              | <b>materiála</b>  | .                                                               |
| lea        | čállii torehtalaš duogáš, mo čállii dovda dutkanárbevieru ja | <b>materiála</b>  | systemáhtalašvuoda ja áigeguovdilvuoda.                         |
|            | Áigodat sáhtttá leat                                         | <b>materiála</b>  | almustahttinbeaivemearri dahje manjimus geardi, go sisdoall     |
| allanbáddi | sealluhuvvo dahje priváhta olmmoš, gean hálddus              | <b>materiála</b>  | lea.                                                            |
|            |                                                              | <b>materiálan</b> | , de daid gánniha merket sierra joavkun gáldologahallamii.      |
|            | Mielddusin biddjo                                            | <b>materiála</b>  | , mii ii dárbbas leat ášši áddema geažil teavsttas, muhto masa  |
|            | Dákkár lea ovdamearkan dutkanmateriála háhkamii guoski       | <b>materiála</b>  |                                                                 |
|            | Dán                                                          | <b>materiálas</b> | eat leat vuhtiváldán golladeami mii lea heajudan kvalitehta, ee |
|            | Dasa lassan lea                                              | <b>materiála</b>  | pieđahallojuuvon nu ahte puohtunkiláiat earuhuvvoit dan miel    |

Teakstačaokkáldat

Administrative texts

teavstta birra

tihttel: *[guorus]*

domeana: administration

sáni iešvuotta

sátneluohká: substantiiva

vuoddohápmi: materiála

sojahananalysa: [N.Sg.Gen](#)

dependeansa: [→N](#)

Čájjet dependeansamuora

Navigation icons: back, forward, search, etc.

# TODO list

- lemgram
- word picture
- map (NEs)
- login
- word aligned corpora

# Usage

- Linguists use Korp for getting an empirical foundation to their work (scientific articles)
- Translators checking technical terms (mostly admin domain) in the parallel corpora
- Language activists and students use Korp to get usage examples
- Dictionary users (<http://sanit.oahpa.no/> and 10 other dictionaries) use Korp to get examples from their dictionary lookups

# Some figures

## Server log for 2018

| Corpus                 | Speakers | MWords | Searches |
|------------------------|----------|--------|----------|
| North Saami            | 21000    | 28.4   | 6 965    |
| Lule Saami             | 1000     | 1.2    | 494      |
| Inari Saami            | 400      | 1.6    | 2 519    |
| South Saami            | 400      | 1.1    | 1 844    |
| Skolt Saami            | 300      | 0.2    | 199      |
| Parallel Norw-North    | -        | 2.1    | 2207     |
| Parallel Norw-South    | -        | 0.001  | 1136     |
| Parallel Finnish-Inari | -        | 0.09   | 31       |

# Conclusions

- ▶ easy installation
- ▶ better configuration
- ▶ some docu needed (word alignment, lemgram, word picture)
- ▶ better presentation of parallel data
- ▶ multi-modal data (audio, video)

**Tromsø likes Korp a lot!**