

Implementations of language technology:

Learning an endangered language and facilitating translation work

Marja-Liisa Olthuis & Ciprian-Virgil Gerstenberger

University of Tromsø – The Arctic University of Norway

XLIII Finnish Conference of Linguistics
May 2016 Oulu, Finland

Overview

Aanaar Saami language community

the revitalisation of a small indigenous language

Saami Language Technology

the Machine Translation project between North Saami and Aanaar Saami

Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

Challenges

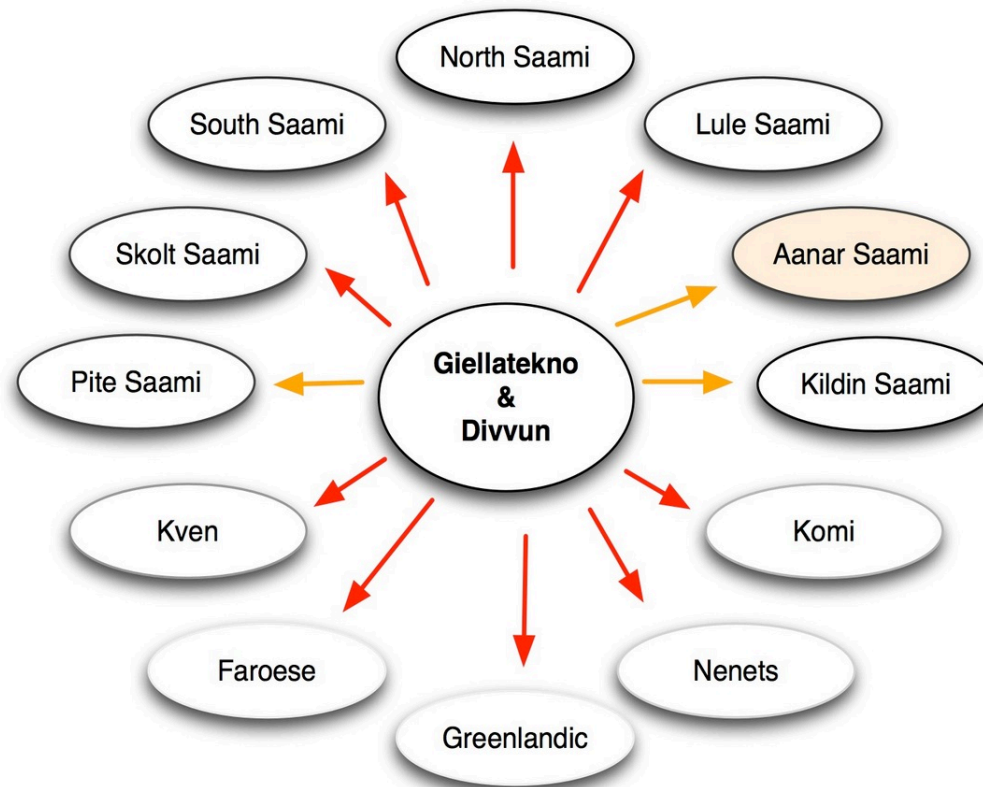
communication between researchers and language community

Worldwide use of *Giellatekno* & *Divvun* infrastructure



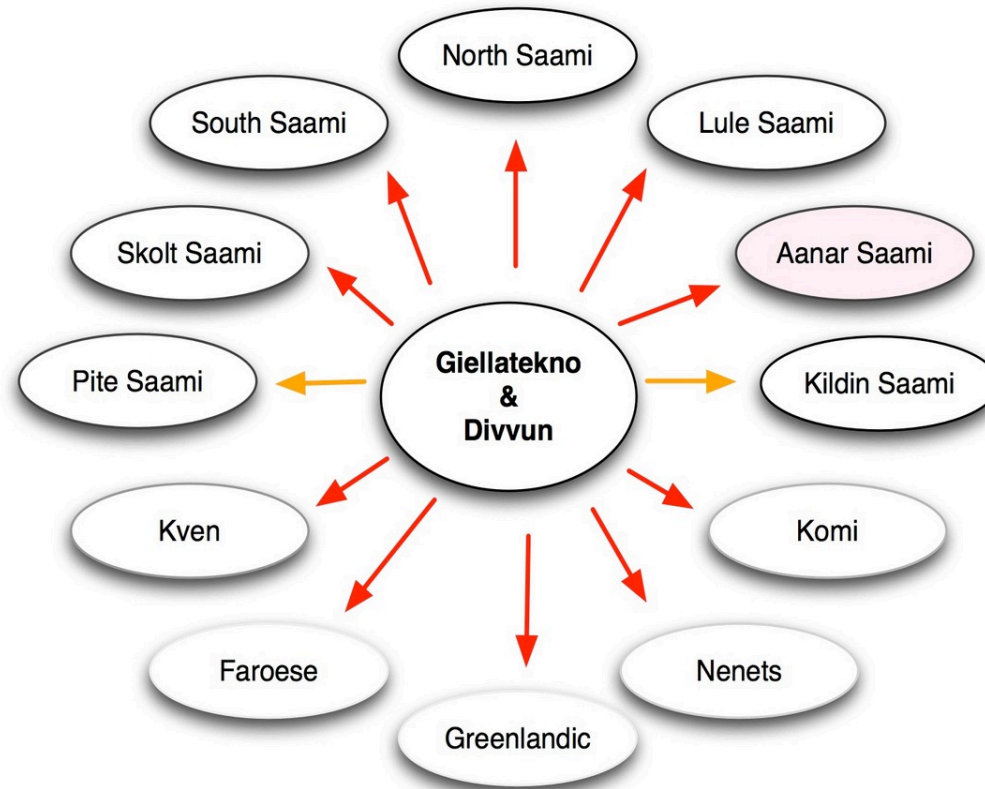
Before 2015

- Finite State Transducer (FST) ==> **morphology analysis**
- No Finite State Transducer (FST) ==> ~~morphology analysis~~

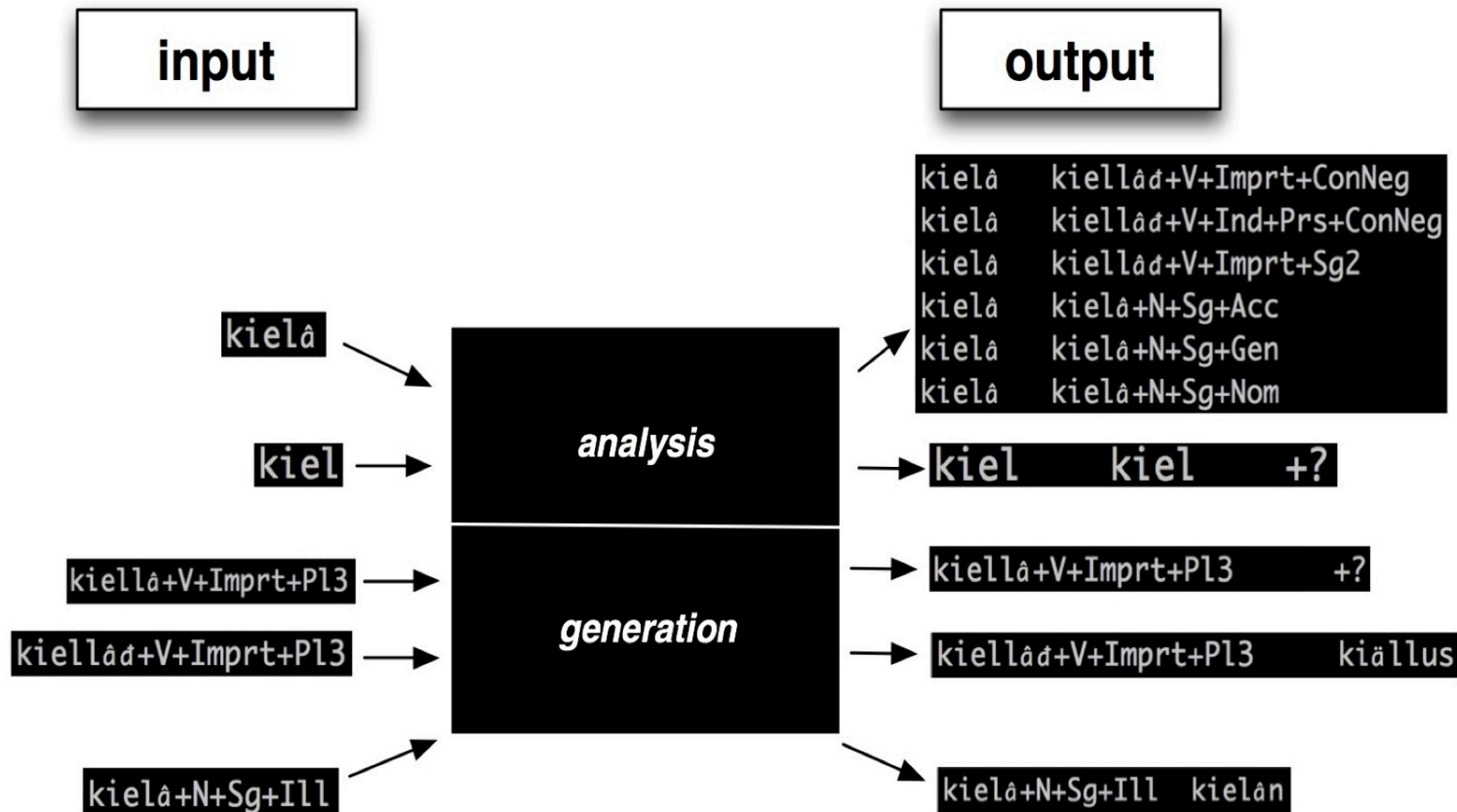


After 2015

- Finite State Transducer (FST) ==> **morphology analysis**
- No Finite State Transducer (FST) ==> ~~morphology analysis~~



What is a finite state transducer (FST)?



Overview

Aanaar Saami language community

the revitalisation of a small indigenous language

Saami Language Technology

the Machine Translation project between North Saami and Aanaar Saami

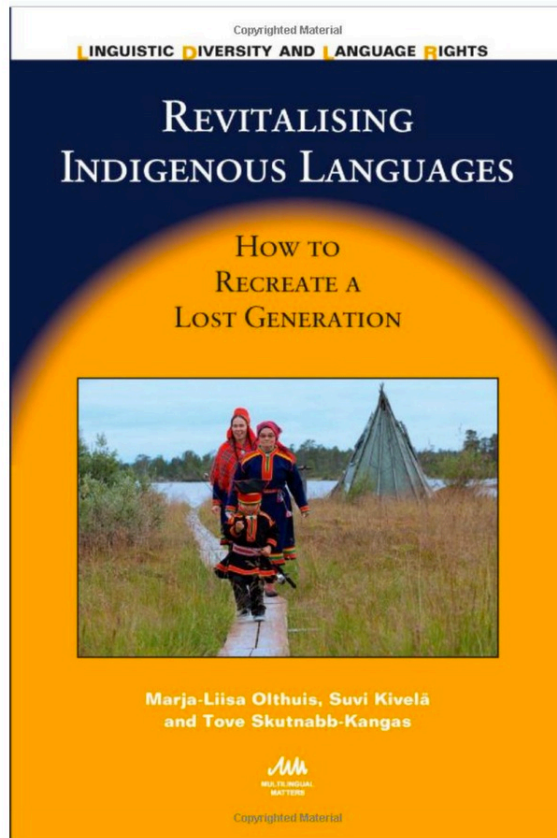
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

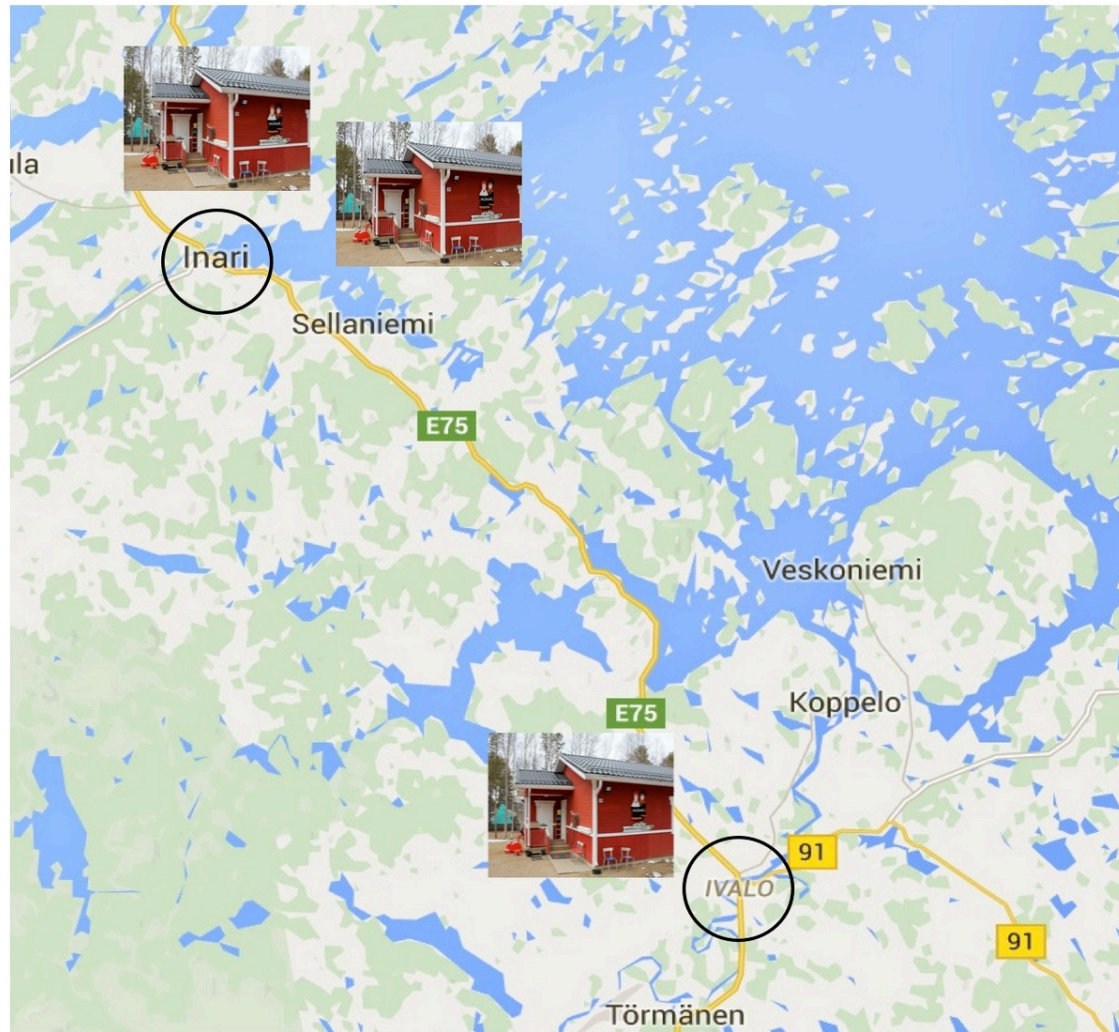
Challenges

communication between researchers and language community

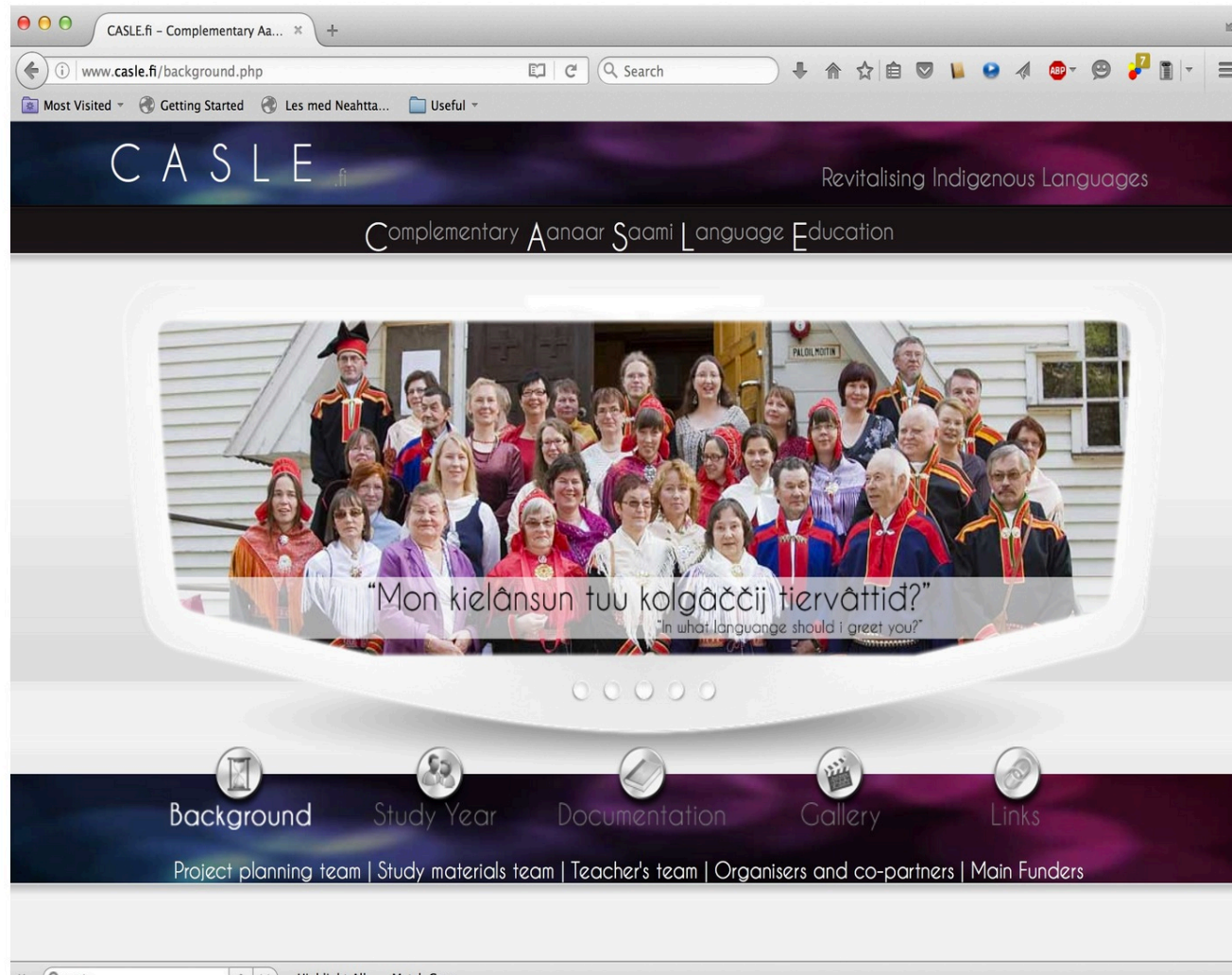
Aanaar Saami revitalisation



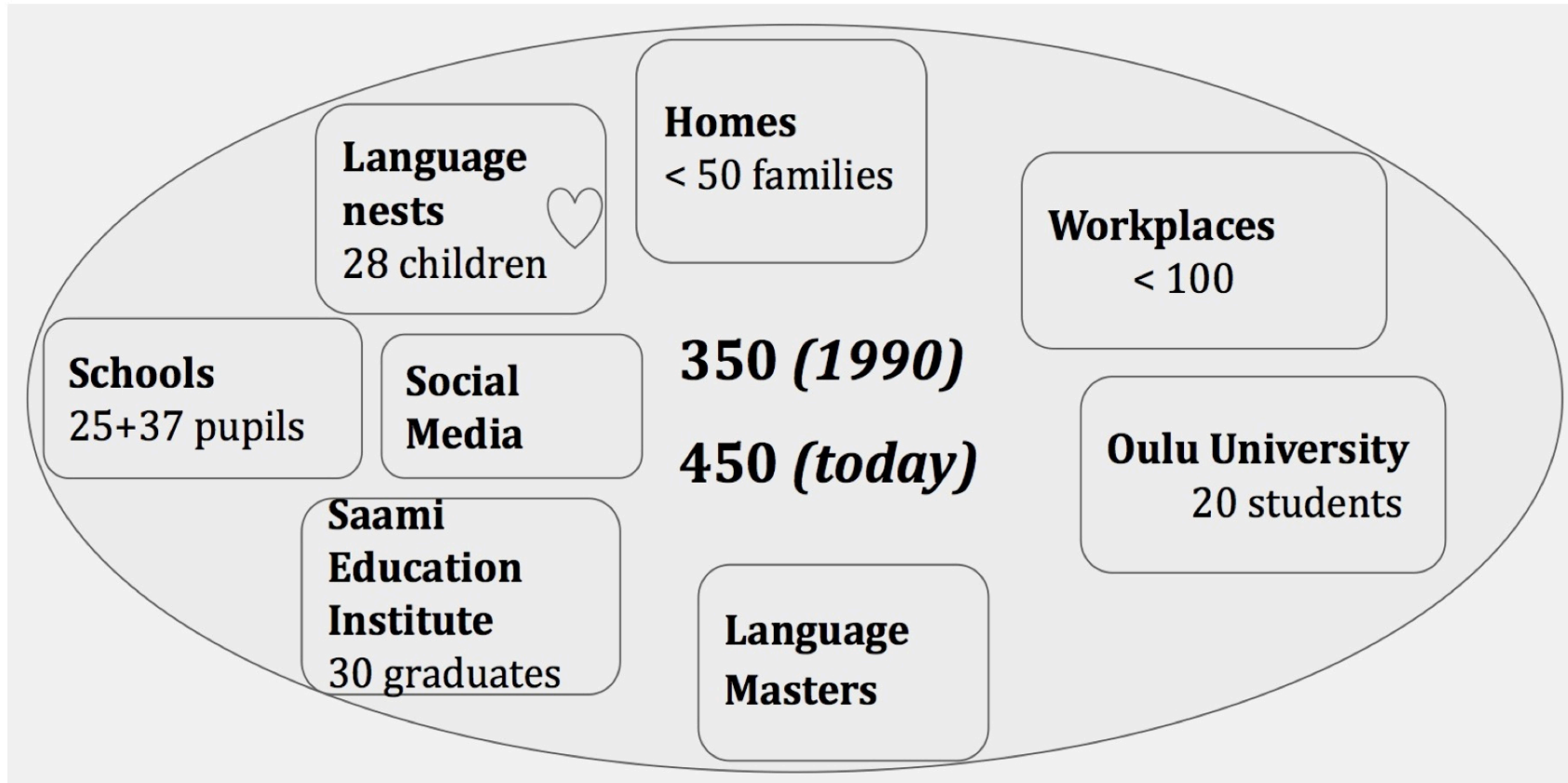
Language nests: child education



The CASLE project: adult education



Language transmission



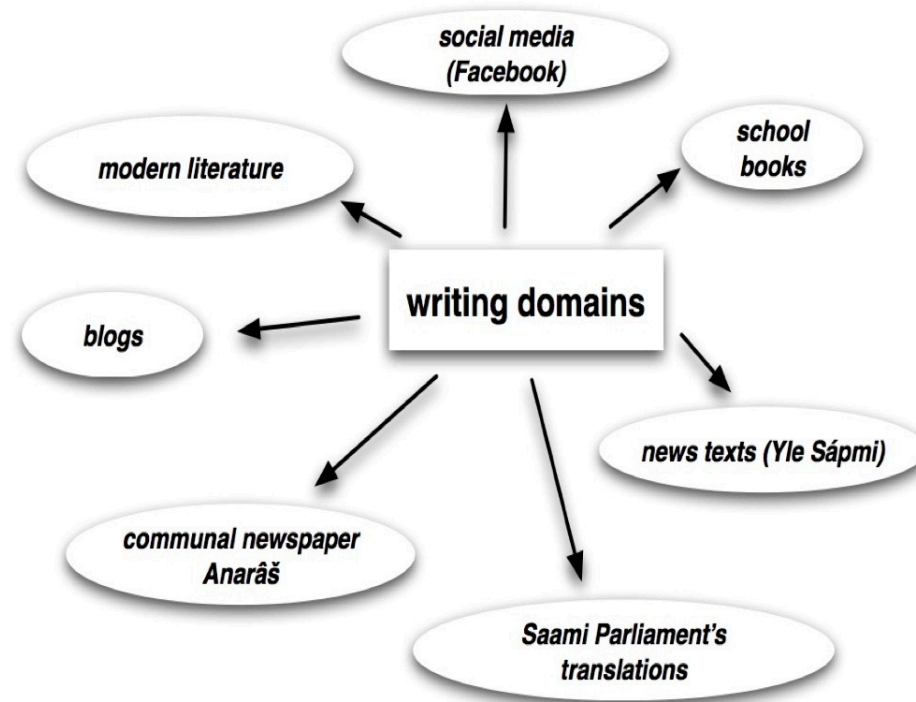
Aanaar Saami: a modern language in a modern society



Čyeti čället: 100 writers

Estimated number of writers of Aanaar Saami

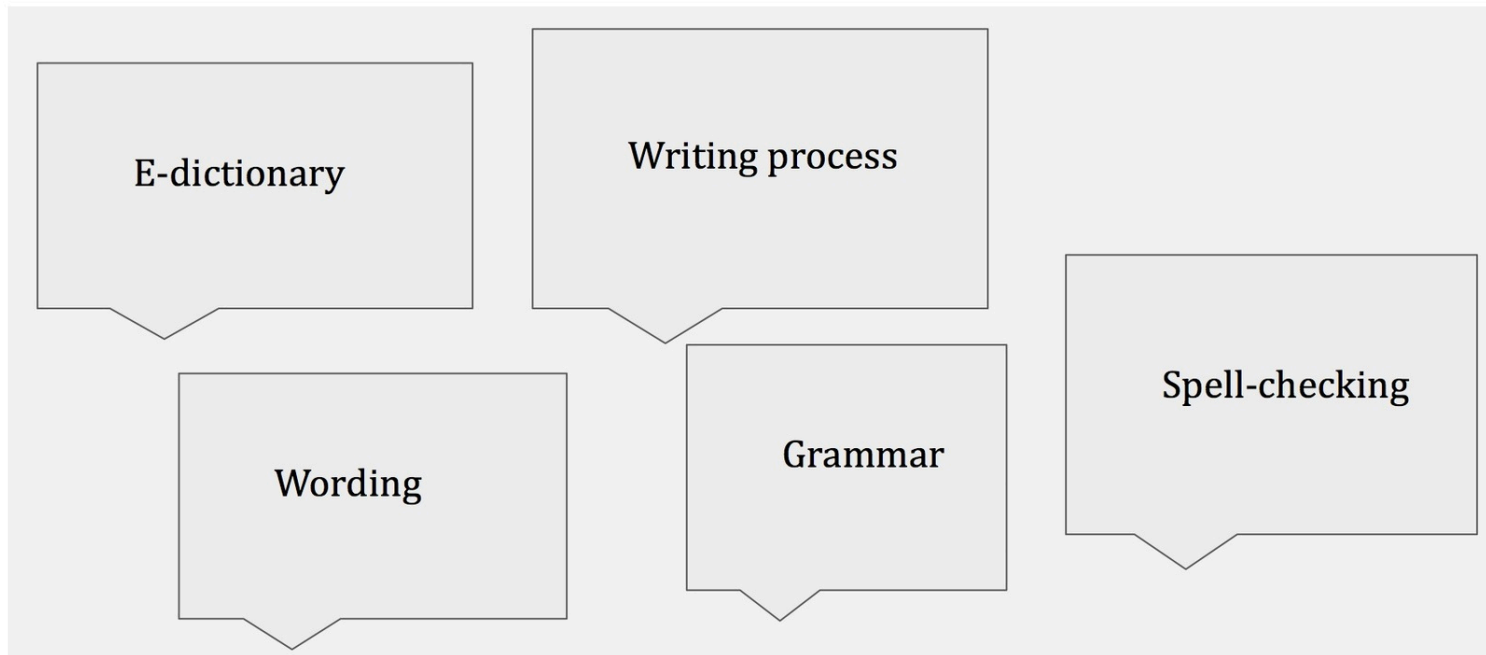
- one very active fluent writer:
the editor-in-chief of Anarâš (70 +)
- a few fluent younger writers:
writing every now and then
- 40 semi-fluent writers



"What do you write in Aanaar Saami?"

	Daily	Weekly	Monthly	Rarely	Never
Letters and postcards	0	1	3	20	11
Blogs	0	1	2	5	27
SMS	9	20	4	4	1
Social media in the Internet	12	12	4	2	8
E-mails	8	15	11	3	0
Scientific articles	0	1	7	13	16
Modern literature: poems, stories	1	3	7	18	9
Narratives about old times	0	2	2	8	26
Translations	2	8	8	14	6
TOTAL	32	63	48	87	104

Personal needs in writing (mainly for L2 speakers)



Overview

Aanaar Saami language community

the revitalisation of a small indigenous language

Saami Language Technology

the Machine Translation project between North Saami and Aanaar Saami

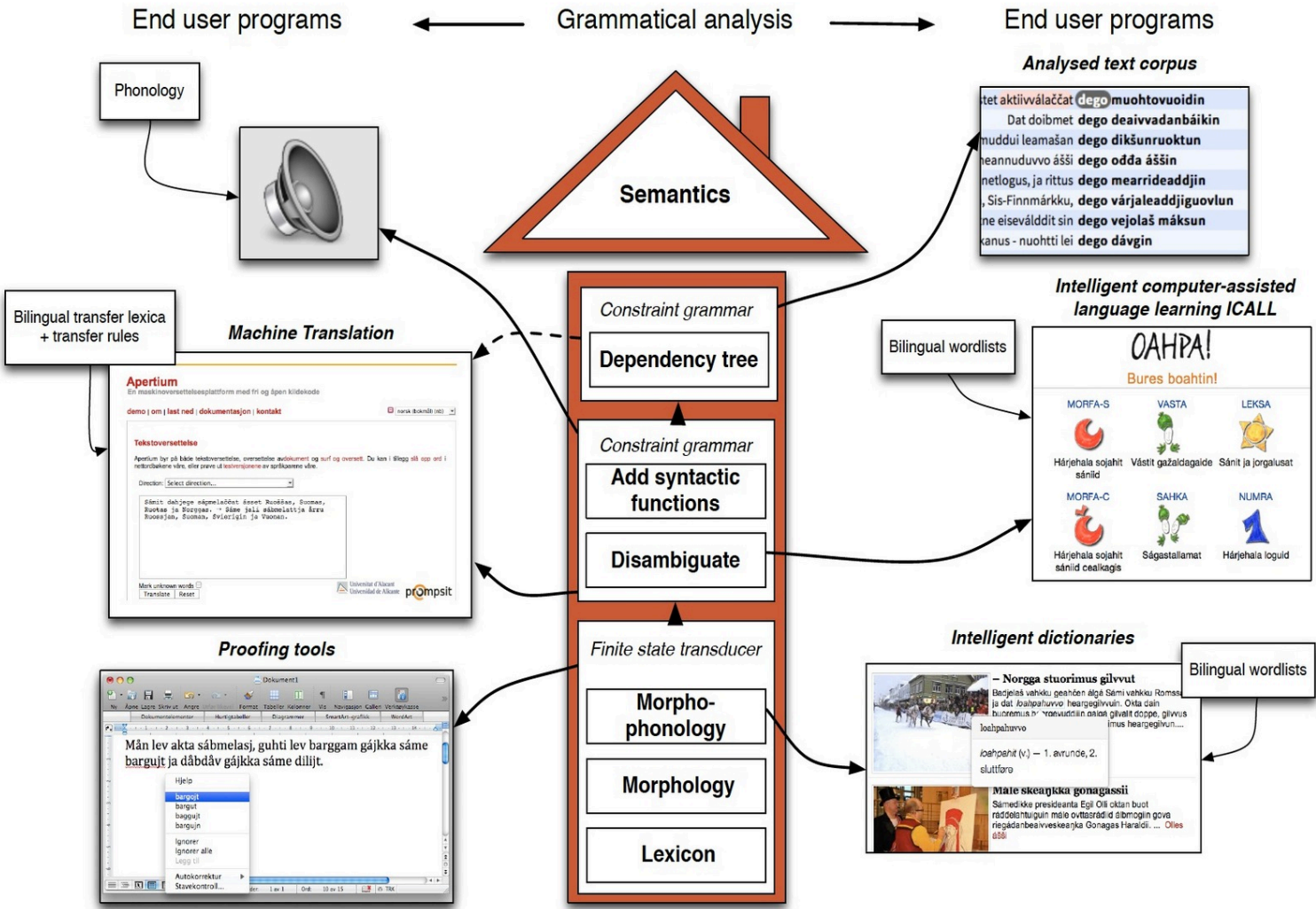
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

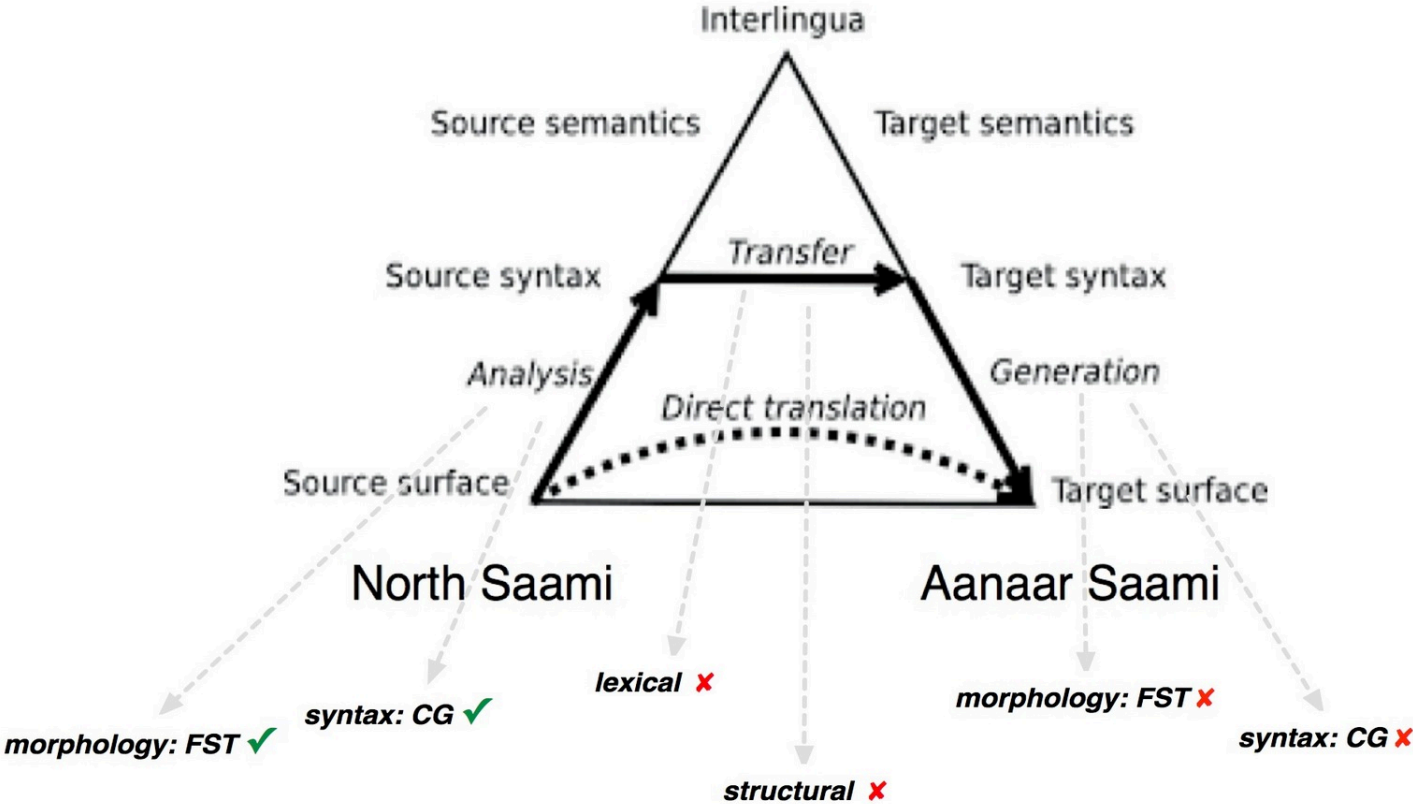
Challenges

communication between researchers and language community

Saami Language Technology: *Giellatekno* & *Divvun*



Machine Translation between minority languages



Resource building: monolingual corpus

The screenshot shows the Korp web interface for resource building. The browser address bar is `gtweb.uit.no/korp/`. The page title is "KORP". The navigation bar includes links for "North Saami texts", "Lule Saami texts", "South Saami texts", "Aanaar Saami texts", "Skolt Saami texts", "Fler", "Cite Korp", "Davvisámi", "Norsk", "Suomi", and "English".

On the left sidebar, there are options for "Simple" and "Extended" views, a search box containing "kielâ", and a "KWIC" section with "hits per page: 50".

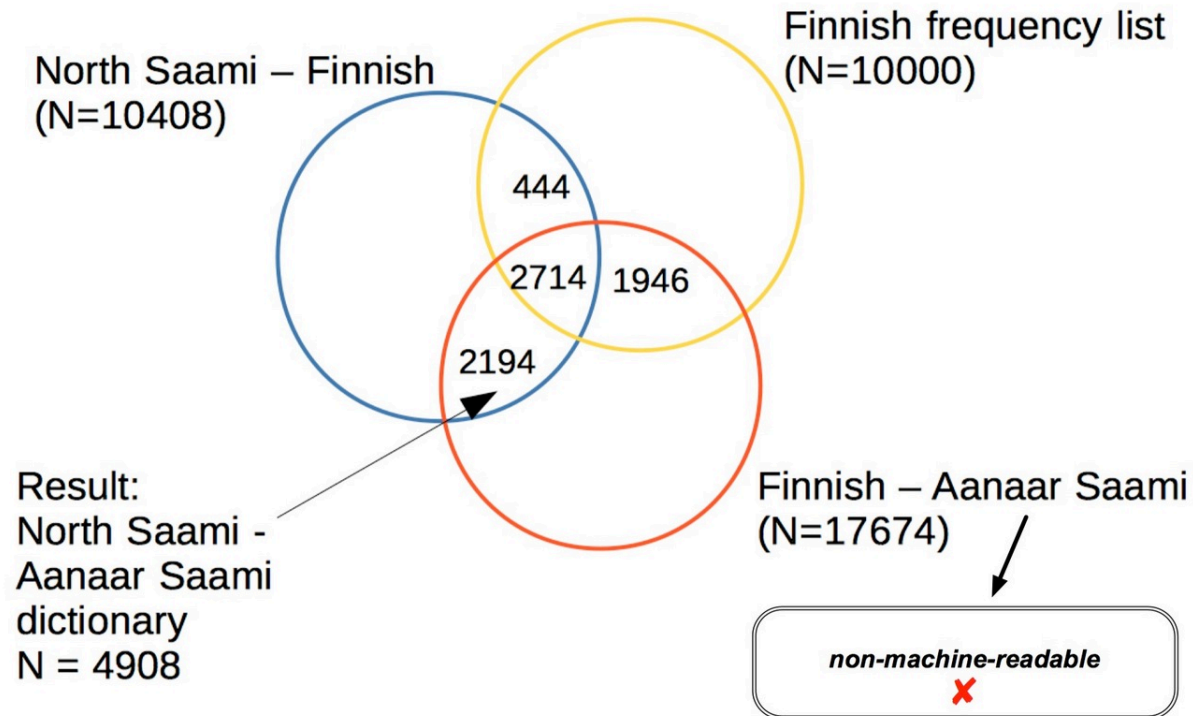
The main content area shows "4 corpora selected (all) — 1.56M of 1.56M tokens". A red circle highlights this text, with an arrow pointing to the label "number of tokens". Below this, there are "number of sentences" and "level of annotation" sections. The "level of annotation" section is expanded to show "Administrative texts" with a description: "Administrative texts, mostly from the Saami Parliament in Finland". It also lists statistics: "Number of tokens: 160,690", "Number of sentences: 14,303", and "Last update: 2015-11-05".

In the center, a list of text types is shown with checkboxes: "Administrative texts", "Religion texts", "Newspaper texts", and "Science texts". A red circle highlights the text "149,493 sentences in the selected corpora" below this list, with an arrow pointing to the label "number of sentences".

At the bottom, there is a "Text attributes" section with fields for "title", "year", "domain", and "undefined". Below that is a "Word attributes" section with fields for "part-of-speech", "grammatical analysis", and "dependency relation". A red circle highlights the "part-of-speech: noun" and "grammatical analysis: N.Sg.Nom" fields, with an arrow pointing to the label "level of annotation".

The main text area displays a snippet of text in Saami: "i iävtuid alnetoollâd já ovdedid jiešivtuvatv jiejjâs kielâ, kulttuur já iäláttástoomâ sehe tipšod já ovdedid kulttuurjieshalo...".

Resource building: bilingual dictionary



Online dictionary Nettidigisäänih (NDS)

Nettidigisäänih [Home](#) [Plugins](#) [Sänikirjeest](#)

smn→fin ▾ puáriskandâ [Uusâ](#) puáriskandâ is a possible form of ...

puáriskandâ (subst.) puáriskandâ [Texts →](#)

	ENTALL	FLERTALL
nominativ	puáriskandâ	puáriskaandah
akkusativ	puáriskaandâ	puáriskaandâid
genitiv	puáriskaandâ	puáriskandâi puáriskaandâi
illativ	puáriskaandân	puáriskandâid
lokativ	puáriskaandâst	puáriskaandâin
komitativ	puáriskandâin	
essiv	puá	
par	puá	

o vanhapoika

Heli Aikio 11 t

Uusâm [hvelkkivuotâsaanijd](#).
liččii munjin? Suomâkielâst láá
uv tiegáreh anarâškielâst? Ij vi
táárkut äijih viiljá?

hvelkkivuotâsaanijd

sääni (subst.) — sana

hvelkki (subst.) — sukulainen

sääni (subst.) — leikko

(korvamerkissä)

Resource building: parallel corpus

North Saami ← Finnish → Aanaar Saami

Size: 600 files

Genre: Saami Parliament's documents
(government + administration)

Format: Translation Memory eXchange (TMX)

```
<tu>
  <tuv xml:lang="sme">
    <seg>Senaatti-kiinteistöt västida Sajosa hukseheamis ja hálddaša giddodaga ./seg>
  </tuv>
  <tuv xml:lang="smn">
    <seg>Seenaat-kiddoduvah västid Saijoos huksiimist sehe haaldáš kiddoduv ton valmáštum maŋa ./seg>
  </tuv>
</tu>
```

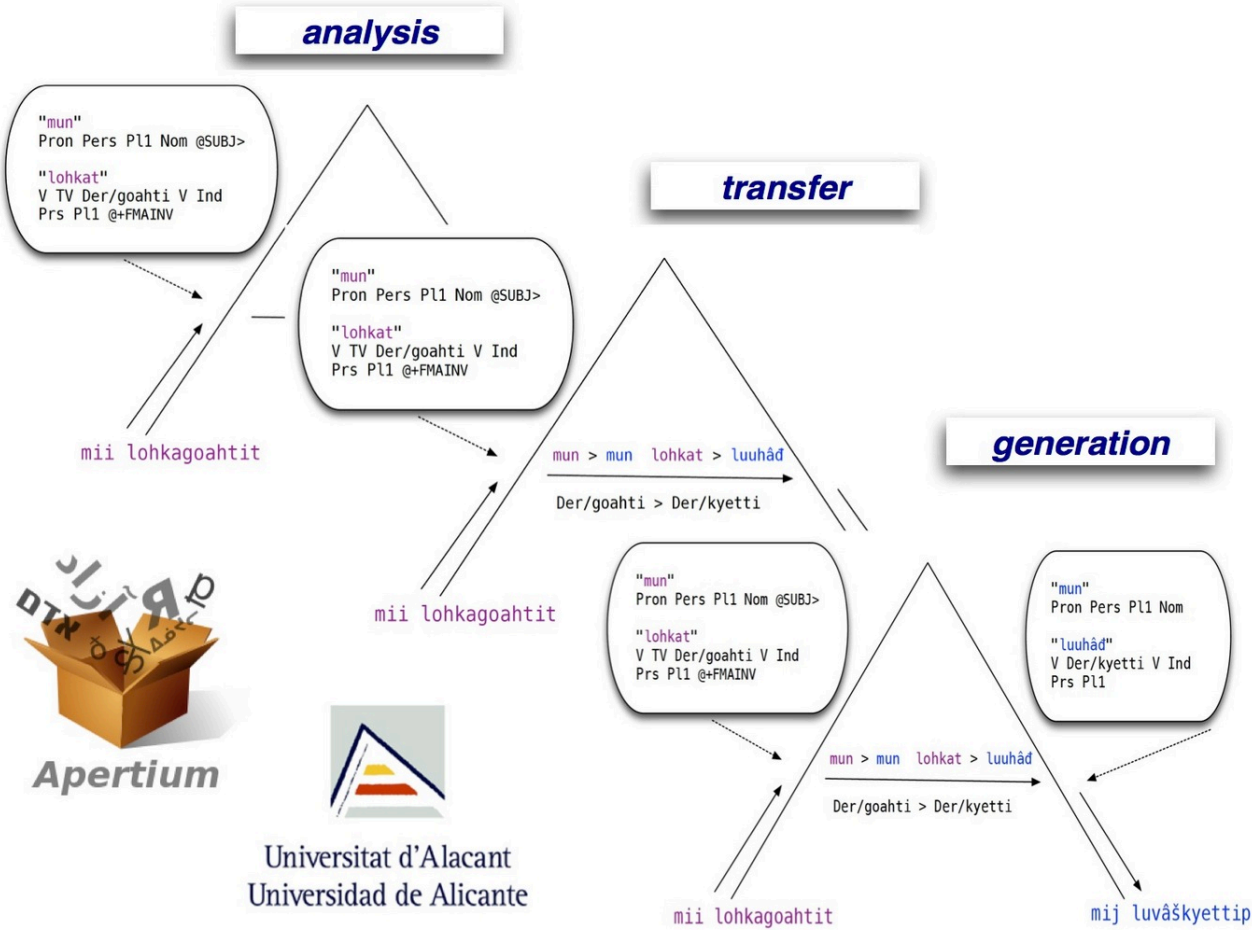
Usage: Evaluation of the Machine Translation output

Senaatti-kiinteistöt västida Sajosa hukseheamis ja hálddaša giddodaga .

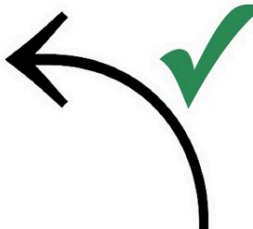
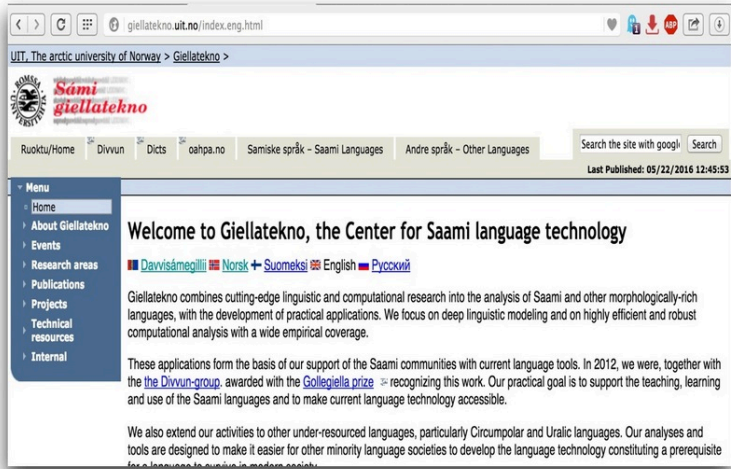
Seenaat-kiddoduvah västid Saijoos huksiimist sehe haaldáš kiddoduv ton valmáštum maŋa .

*Senaatti-*kiinteistöt västid Sajos #huksiđ<vblex><der_tt><der_nomact><n><sg><loc> já haaldáš kiddoduv .

Machine Translation process



After 2016?



Overview

Aanaar Saami language community

the revitalisation of a small indigenous language

Saami Language Technology

the Machine Translation project between North Saami and Aanaar Saami

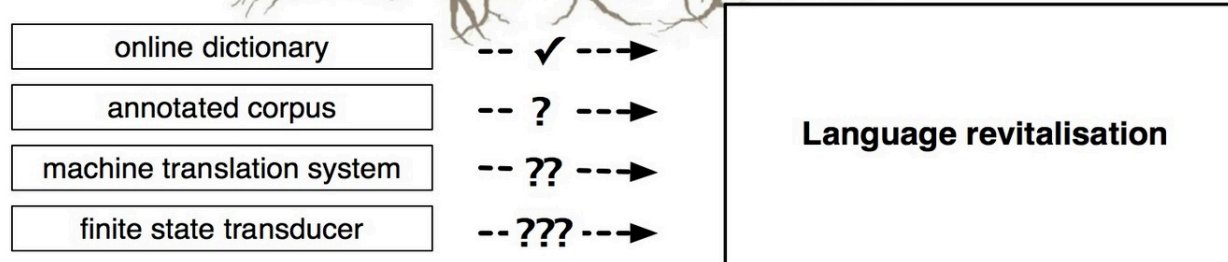
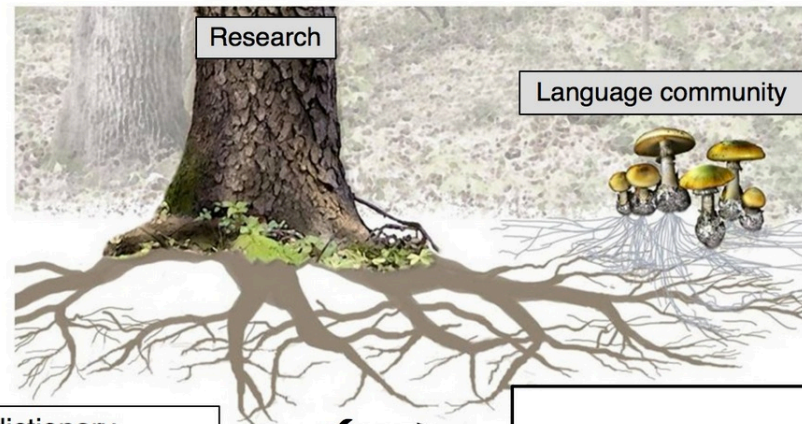
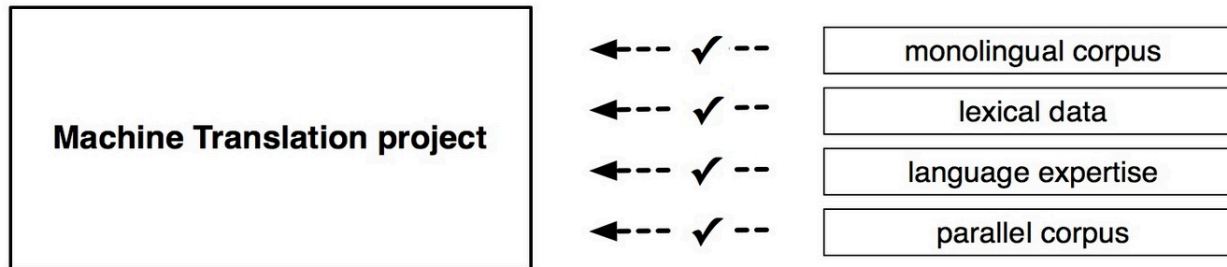
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

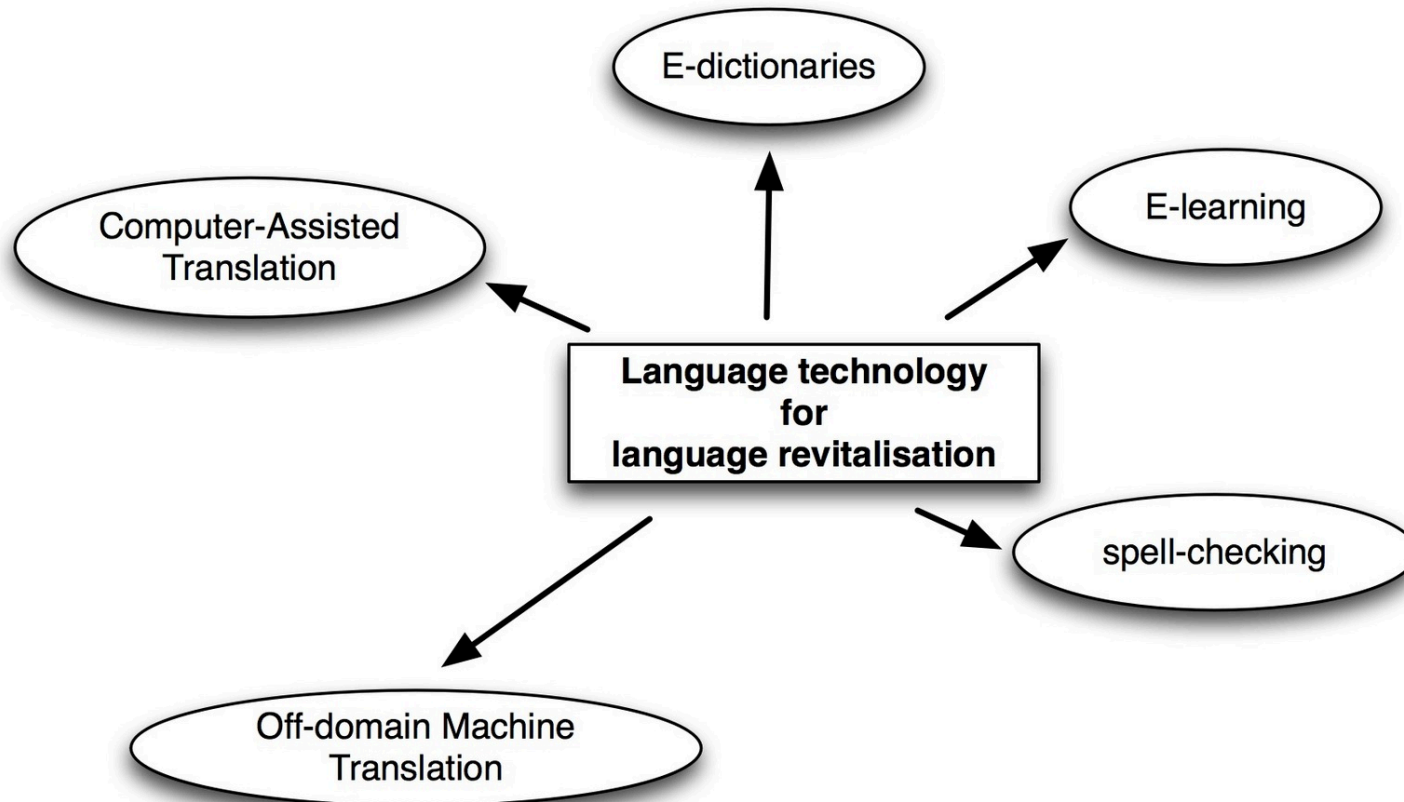
Challenges

communication between researchers and language community

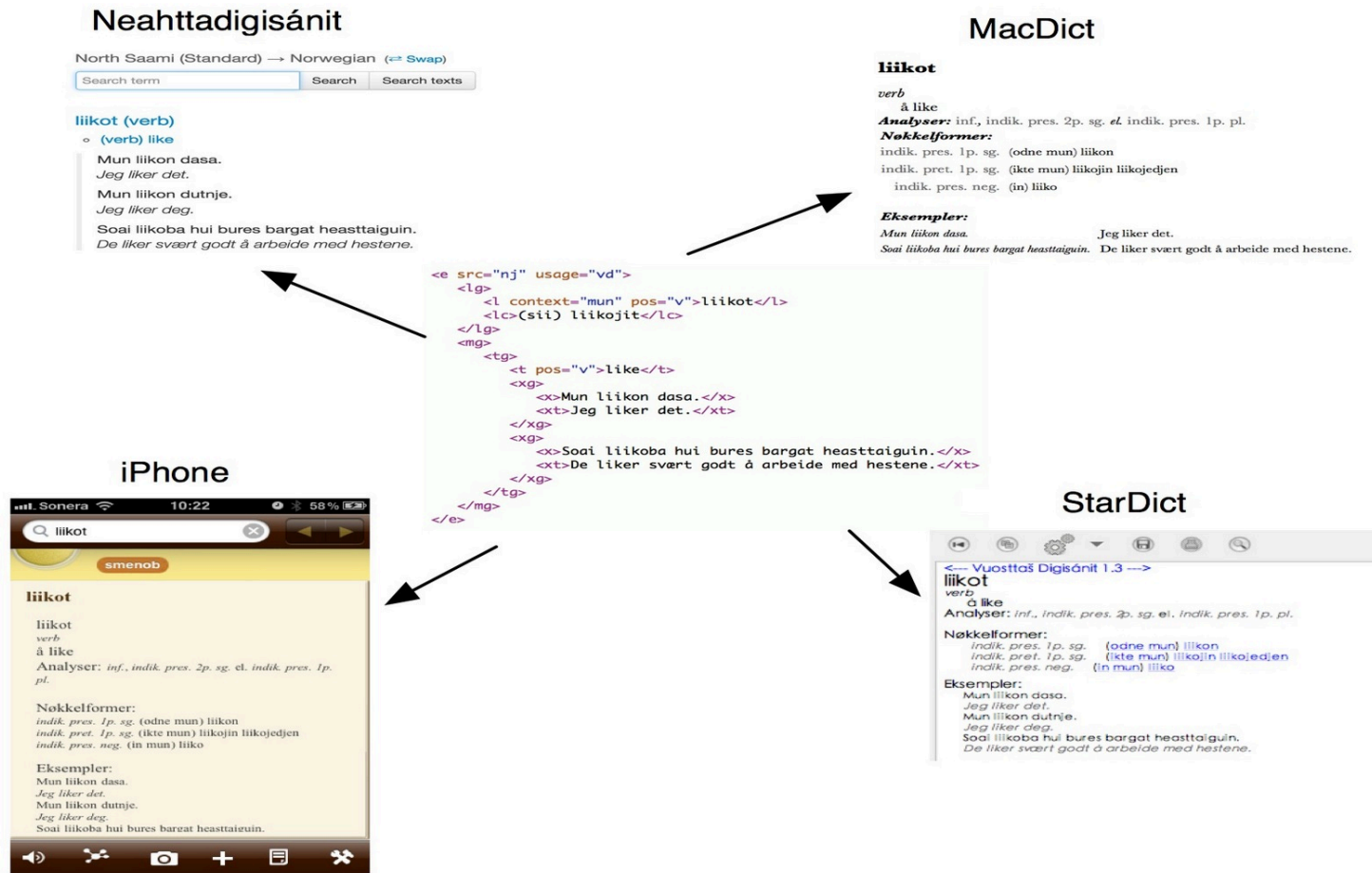
Research-Revitalisation symbiosis



Applications for language revitalisation



Cross-platform online/offline intelligent dictionaries



ICALL application: Oahpa

HJELP


OAHPA!

Bures boahтин!

Veahkkegiella


Suopman

MORFA-S




Hárjehala sojahit
sániid

VASTA




Vástit gažaldagaide

LEKSA




Sárit ja jorgalusat

MORFA-C




Hárjehala sojahit
sániid cealkagis

SAHKA



Ságastallamat

NUMRA



Hárjehala loguid

OAHPA lea interneahhtaprográmma nuoraide ja rávesolbmuid geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja ođđa bargobihát ráhkaduvvojit automáhtalaččat.

[Bagadus](#)
[Davvisámegiella-dárogiella neahhtasátnegirji](#)
[Davvisámegiela grammatihkka](#)

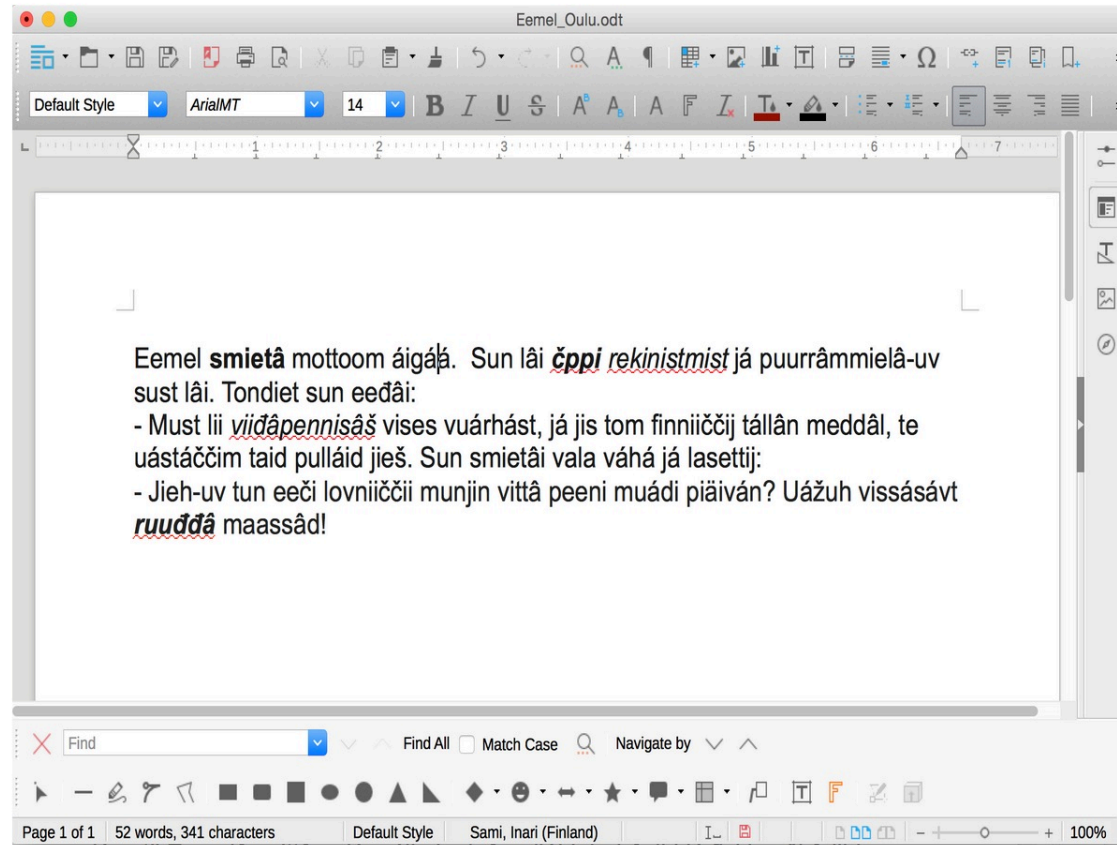
Leatgo dás kurssa oktavuodas? [Logge sisa dás](#)

bilingual dictionary ✓

morphology: FST ✓

syntax: CG ✗

Spellchecker: non-detectable word



→ need for a grammar checker

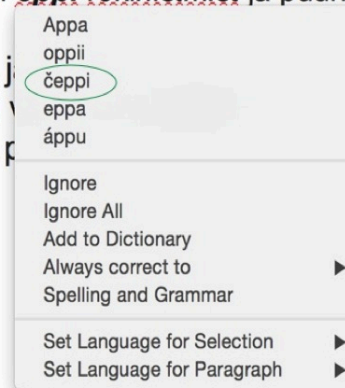
Spellchecker: correcting a typo

Eemel **smietâi** mottoom áigáá. Sun lâi čppi rekinismist já puurrâmmielâ-uv sust lâi. Tondiet sun eedâi:

- Must lii viidâpennisâš vises vuárhást, já jis tom finniičij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniičcii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ruudđâ maassâd!

Eemel **smietâi** mottoom áigáá. Sun lâi čppi rekinismist já puurrâmmielâ-uv sust lâi. Tondiet sun eedâi:

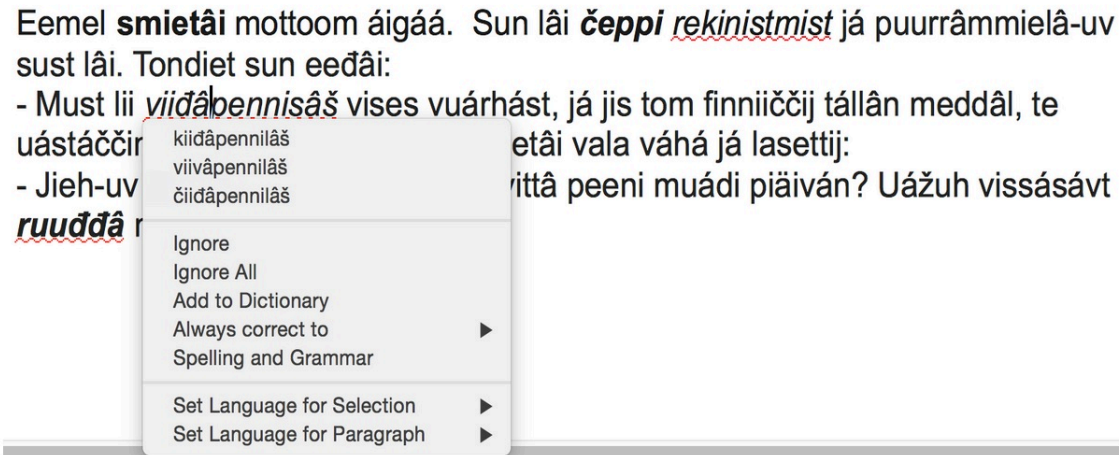
- Must lii viidâpennisâš vises vuárhást, já jis tom finniičij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniičcii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ruudđâ maassâd!



Eemel **smietâi** mottoom áigáá. Sun lâi čeppi rekinismist já puurrâmmielâ-uv sust lâi. Tondiet sun eedâi:

- Must lii viidâpennisâš vises vuárhást, já jis tom finniičij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniičcii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ruudđâ maassâd!

Spellchecker: unknown, yet correct word form



→ add lemma to the source files!

→ recompile the spellchecker plugin!

→ upload the plugin to the editing program!

Extending the "knowledge" of the tool

==> add missing lemma to the source files

```
sekkâkuámmirâš:sekkâ#kuámmir 4LAS_NOUN "pussikämmekkä" ; !
vorrâkuámmirâš:vorrâ#kuámmir 4LAS_NOUN "verikämmekkä" ; !
oivârâš:oi4vâr 4LAS_NOUN "hakkuupölkky" ; !
viidâpennisâš:viidâ#pennis 4LAS_NOUN "viisipenninen" ; !
sajasâš:sajas 4LAS_NOUN "viransijainen" ; !
čiähásiemmânsâš:čiähâ#siemmâns 4LAS_NOUN "koppisiemeninen" ; !
oovtâitosâš:oovtâ#itos 4LAS_NOUN "tasavertainen" ; !
```

==> configure the process accordingly

```
tf4-hsl-m0019:smn marjaliisa$ ./configure --with-hfst --enable-spellers
checking for a BSD-compatible install... /usr/bin/install -c
checking whether build environment is sane... yes
checking for a thread-safe mkdir -p... build-aux/install-sh -c -d
checking for gawk... gawk
```

==> recompile the language tools

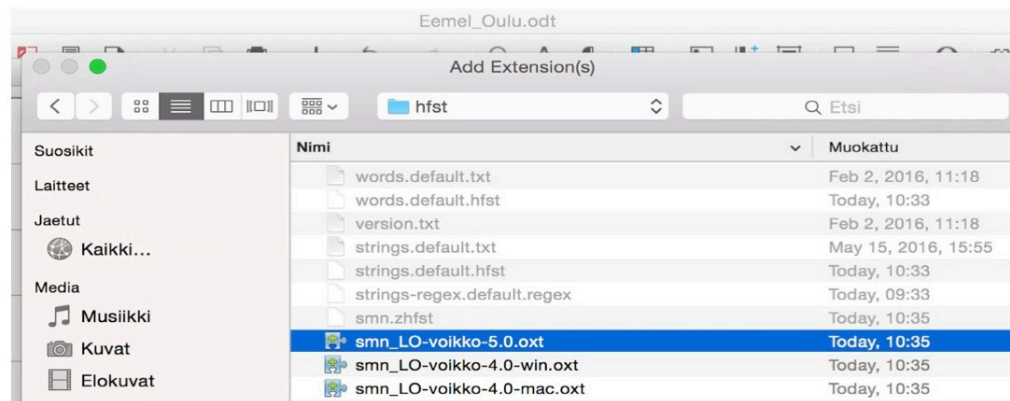
```
tf4-hsl-m0019:smn marjaliisa$ make
*** Compiling the smn language. ***
CDPATH="${ZSH_VERSION+}:" && cd . && /bin/sh /Users/marjaliisa/main/langs/smn/build-aux/missing auto
conf
Making all in .
*** Compiling the smn language. ***
CDPATH="${ZSH_VERSION+}:" && cd . && /bin/sh /Users/marjaliisa/main/langs/smn/build-aux/missing auto
conf
make[1]: Nothing to be done for `all-am'.
Making all in src
Making all in phonology
make[2]: Nothing to be done for `all'.
Making all in morphology
GEN      lexicon.tmp.lexc
XFST     lexicon.tmp.xfst
```

Updating the spellchecker plugin

==> check the result of the tool generation

```
tf4-hsl-m0019:smn marjaliisa$ ls tools/spellcheckers/fstbased/desktop/hfst/*.oxt
tools/spellcheckers/fstbased/desktop/hfst/smn_LO-voikko-4.0-mac.oxt
tools/spellcheckers/fstbased/desktop/hfst/smn_LO-voikko-4.0-win.oxt
tools/spellcheckers/fstbased/desktop/hfst/smn_LO-voikko-5.0.oxt
tf4-hsl-m0019:smn marjaliisa$
```

==> add the plugin to the editing tool



==> reopen the text file

Eemel **smietâi** mottoom áigáá. Sun lâi **čepi rekinismist** já puurrâmmielâ-uv sust lâi. Tondiet sun eedâi:

- Must lii **viidâpennisâš** vises vuárhást, já jis tom finniičij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:

- Jieh-uv tun eeči lovniičii munjin vittâ peeni muádi piäiván? Uážuh vissásávt **ruudđâ** maassâd!

Off-domain Machine Translation test: children's books

Beaivelanjas eadni geige sutnje stuorra ránes konvoluhta.	←	Source: North Saami
Orroomvisteest enni keigee sunjin stuorrâ ränis kooveert.	←	Target: Aanar Saami (human)
Orroomvisteest enni keigee sunjin stuorrâ ränis kooveert.	←	Target: Aanaar Saami (machine)
"Áigin Lávra" lea dasa čálistuvvon stuorra, roanci bustávaiguin.	←	
"Äigee Lavrá" lii toos čaallum stuorrâ, ruánzáás puustavijgijn.	←	
"Áigin Lávra" lii tos čäálistum stuorrâ, *roanci pustiguin.	←	
Muđui ii oidno das mihkkege. li leat čállojuvvon geaidnonamma konvoluhttii, ii báikenammage.	←	
Mudoï tast ij oinuu mihheen. Ij kiäinunommâ ige päikkinommâgin lah čaallum kovertân.	←	
Mudoï ij oinuu tast mihheen. Ij lah čaallum kiäinunommâ kovertân, ij päikkinommâgin.	←	
Áigin jorgala konvoluhta, muhto ii maŋábealdege leat čállon mihkkege.	←	
Äigee jorgeet kooveert, mut ij ton tuáhágin lah čaallum mihheen.	←	
#Áigin jurgâl kooveert, mut ij maŋabelngin lah čaallum mihheen.	←	

= word form not in the morphology * = lemma not in the lexicon

Online use of the Machine Translation system

Giellatekno's translation systems (work in progress!)

[cat](#) | [eng](#) | [eus](#) | [nno](#) | [nob](#) | [sme](#)

Beaivelanjas eadni geige sutnje stuorra ránes konvoluhta.
 "Áigin Lávra" lea dasa čálistuvvon stuorra, roanci bustávaiguin.
 Muđui ii oidno das mihkkege. Ii leat čállojuvvon geaidnonamma konvoluhttii, ii báikenammage.
 Áigin jorgala konvoluhta, muhto ii maŋábealdege leat čállon mihkkege.

Translate North Sámi → Inari Sámi

Orroomvisteest enni keigee sunjin stuorrâ ränis kooveert. "Áigin Lávra" lii tos čäälístum stuorrâ, *roanci pustuiguin. Mudoij oinuu tast mihheen. Ij lah čaallum kiäinunommâ kovertân, ij päikkinommâgin. #Áigin jurgâl kooveert, mut ij maŋabelngin lah čaallum mihheen.

Timestamp of sme-smn binary directory /usr/share/apertium//apertium-sme-smn:
 2016-05-07 20:00:22

Computer-Assisted Translation (CAT)

The screenshot displays the OmegaT 3.5.4 software interface. The main window is titled "Editor - about_omegat.txt" and contains the following text:

Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.
Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä. <segment 0001>

Esitys sopii teemamateriaaliksi historian opetukseen.

Saamelainen parlamentaarinen neuvosto vaati toimenpidekieltoa YK:n ihmisoikeusneuvostolta siidan tokkakunnan porojen pakkoteurastusten estämiseksi.

The interface also features three panels on the right side:

- Fuzzy Matches:** An empty panel for displaying fuzzy matches.
- Glossary:** A panel containing the following entries:
 - osoittaa = čujottid
 - teksti = tekstâ
 - toimi = toimâ
- Dictionary:** An empty panel for displaying dictionary entries.

At the bottom of the interface, there are four buttons: "Multiple Translations", "Notes", "Comments", and "Machine Translation". In the bottom right corner, there are two status indicators: "0/3 (0/3, 3)" and "87/87".

CAT: total string match

The screenshot displays the OmegaT 3.5.4 interface with two overlapping windows. The top window shows a source segment in the editor and a corresponding target segment in the 'Fuzzy Matches' panel. The bottom window shows the same source segment with a different target segment, also showing a 100% match in the 'Fuzzy Matches' panel. A black arrow points from the match percentage in the bottom window to the match percentage in the top window.

Top Window (Source and Target):

- Editor - about_omegat.txt:**

Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.

Esitys sopii teemamateriaaliksi historian opetukseen.
Esitys sopii teemamateriaaliksi historian opetukseen.<segment 0002>
- Fuzzy Matches:**

1. Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdänpyehtim heivee teemamateriaalân historjá mättäättäsân.
<100/100/100%
/Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www.samediggi.fi/index2.

Bottom Window (Source and Target):

- Editor - about_omegat.txt:**

Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.

Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdänpyehtim heivee teemamateriaalân historjá mättäättäsân.<segment 0002>

Saamelainen parlamentaarinen neuvosto vaatii toimenpidekieltoa YK:n ihmisoikeusneuvostolta siidan tokkakunnan porojen pakkourastusten estämiseksi.
- Fuzzy Matches:**

1. Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdänpyehtim heivee teemamateriaalân historjá mättäättäsân.
<100/100/100%
/Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www.samediggi.fi/index2.
php_option=com_content_task=view_id=140_lang=finnish.html.tmx>

Bottom Panel: Multiple Translations | Notes | Comments | Machine Translation

Status Bar: 0/3 (0/3, 3) 53/60

CAT: partial string match

The screenshot shows the OmegaT 3.5.4 interface with the following components:

- Editor - about_omegat.txt:**
 - Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.
 - Oovdånpyehtim heivee teemamateriaalån historjá máttááttåsån.
 - Saamelainen parlamentaarinen neuvosto vaati toimenpidekieltoa YK:n ihmisoikeusneuvostolta siidan tokkakunnan porojen pakkoteurastusten estämiseksi.**
 - Saamelainen parlamentaarinen neuvosto vaati toimenpidekieltoa YK:n ihmisoikeusneuvostolta siidan tokkakunnan porojen pakkoteurastusten estämiseksi. <segment 0003>
- Fuzzy Matches:**
 - 1. Saamelainen parlamentaarinen neuvosto vaati **vetoomuksessaan 23.9.2011 välitöntä** toimenpidekieltoa YK:n ihmisoikeusneuvostolta **Nellimin siidan eli** tokkakunnan porojen pakkoteurastusten estämiseksi.
 - Säämi Parlamentaarláš Räädi váátá avžuuttásásástis 23.9.2011 tállån tábáhtuvvee olášuttemkiäldu OA olmoošvuogådvuodárááđist Njeellim siijdá poccui pággunjuovámij estimån.
 - <75/75/72%
 - /Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www.samediggi.fi/index2.php_option=com_content_task=view_id=519_lang=finnish.html.tmx>
 - 2. Saamelainen Parlamentaarin Neuvosto vaatii Nellimin pakkoteurastusten pysäyttämistä
 - Säämi Parlamentaarláš Räädi Váátá Njeellim pággunjuovámij orostitten
 - <41/33/38%
 - /Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www.samediaai.fi/index2.
- Glossary:** (Empty)
- Dictionary:** (Empty)

At the bottom of the interface, there are buttons for "Multiple Translations", "Notes", "Comments", and "Machine Translation". The status bar shows "Project autosaved on 22:35" and "1/3 (1/3, 3) 147/147".

Overview

Aanaar Saami language community

the revitalisation of a small indigenous language

Saami Language Technology

the Machine Translation project between North Saami and Aanaar Saami

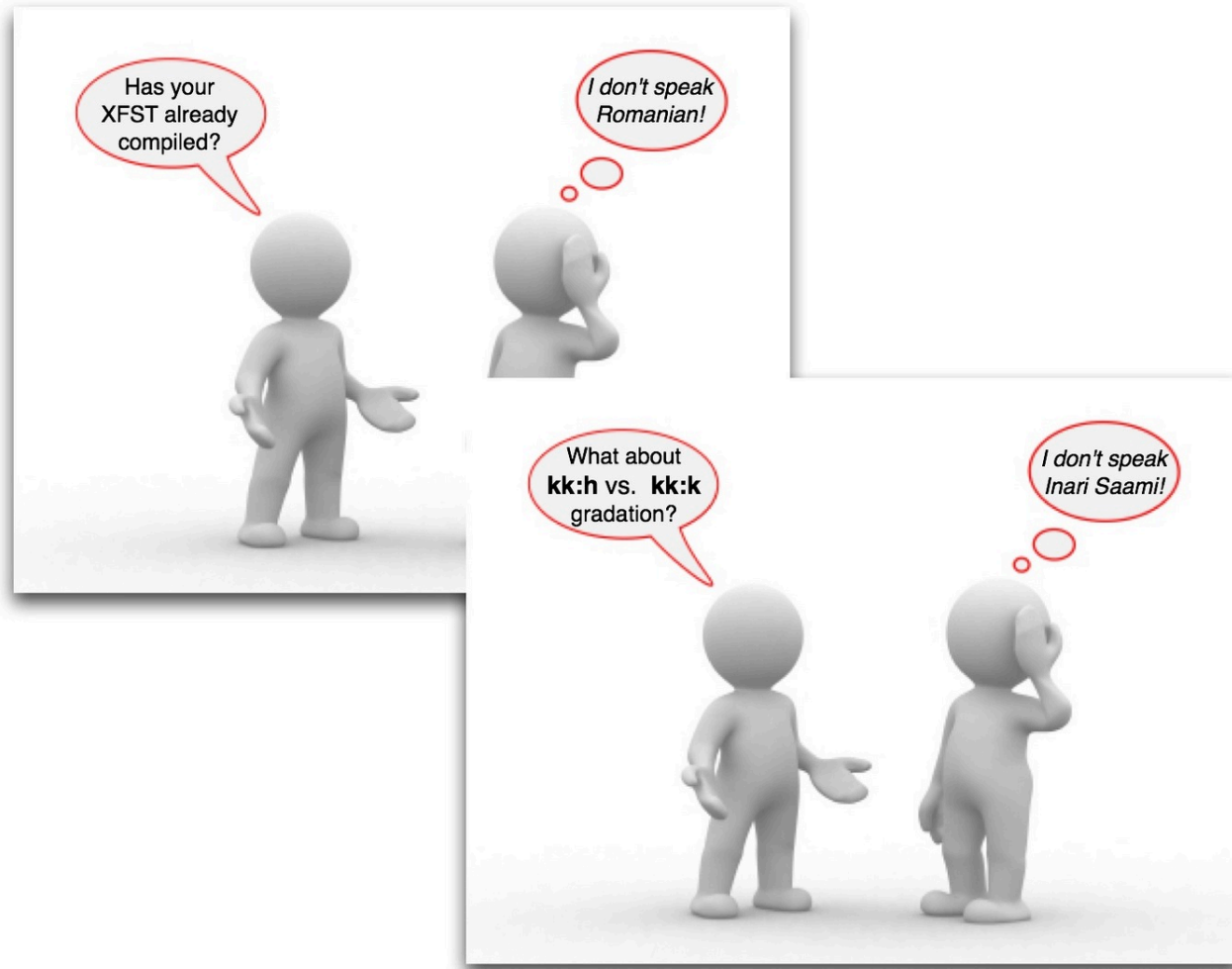
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

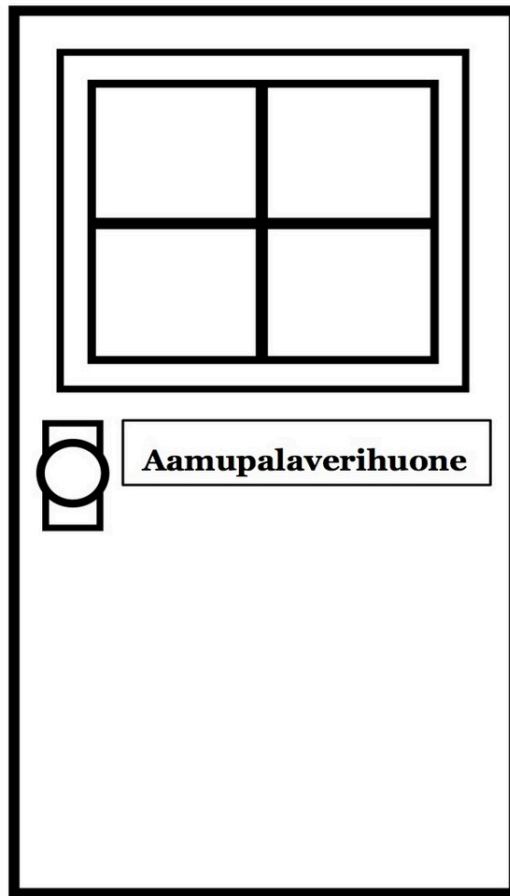
Challenges

communication between researchers and language community

Do we understand each other?



I understand Finnish perfectly!



Machine-readability

aakkostaa ornið puustavij mild
ahdistaa atâštid atâštâm aatâšt

[eng. "to alphabetise"; lit. "to organise alphabets accordingly"]

[eng. "to haunt"; lit. "to haunt", "I haunt", "(s)he haunts"]

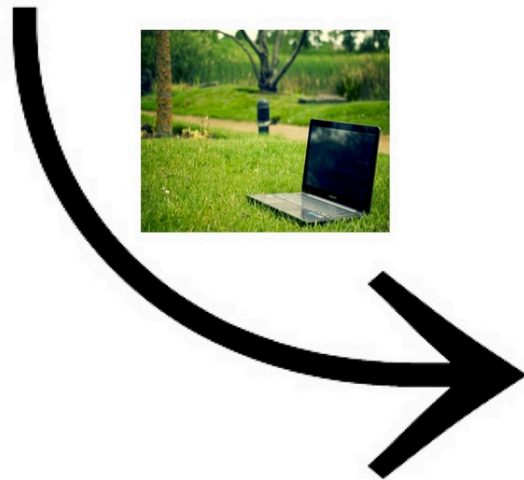
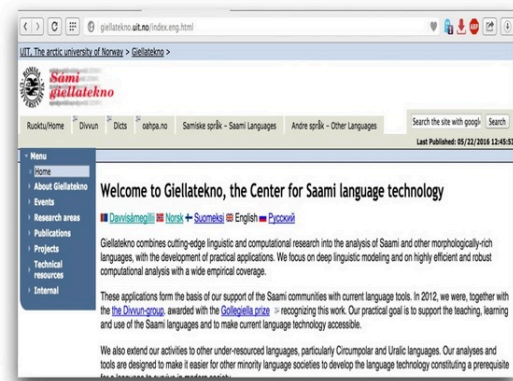
whitespace ambiguity

disambiguate

label items appropriately

```
<e id="3">
  <lg>
    <l pos="V">aakkostaa</l>
  </lg>
  <mg id="1">
    <tg xml:lang="smn">
      <t type="multiword expression">ornið puustavij mild</t>
    </tg>
  </mg>
</e>
<e id="4">
  <lg>
    <l pos="V">ahdistaa</l>
  </lg>
  <mg id="1">
    <tg xml:lang="smn">
      <t pos="V" morph_1s="atâštâm" morph_3s="aatâšt">atâštid</t>
    </tg>
  </mg>
</e>
```

Teaching how language tools work



Thank you for your attention!

We thank our colleagues *Trond Trosterud, Lene Antonsen, Sjur Moshagen, and Erika Sarivaara* for the permission of re-using some of their material!