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

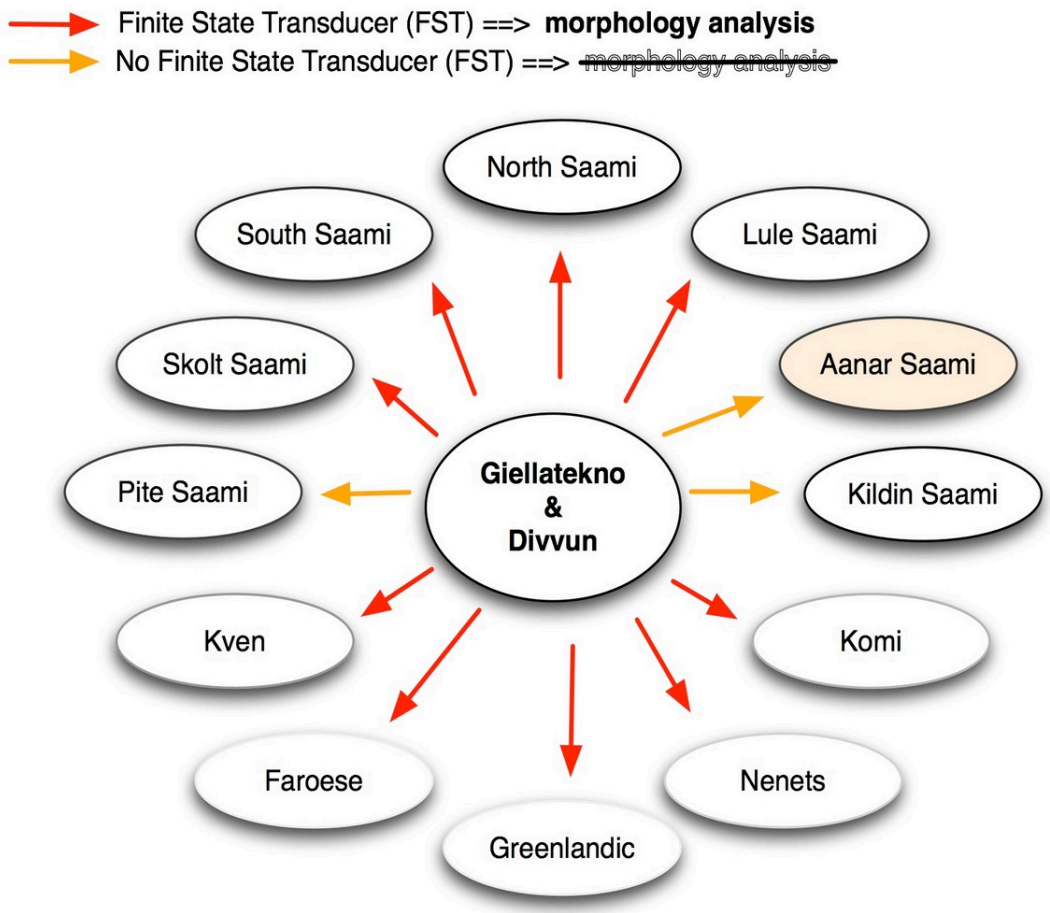
beyond Machine Translation as proof-of-concept

Challenges

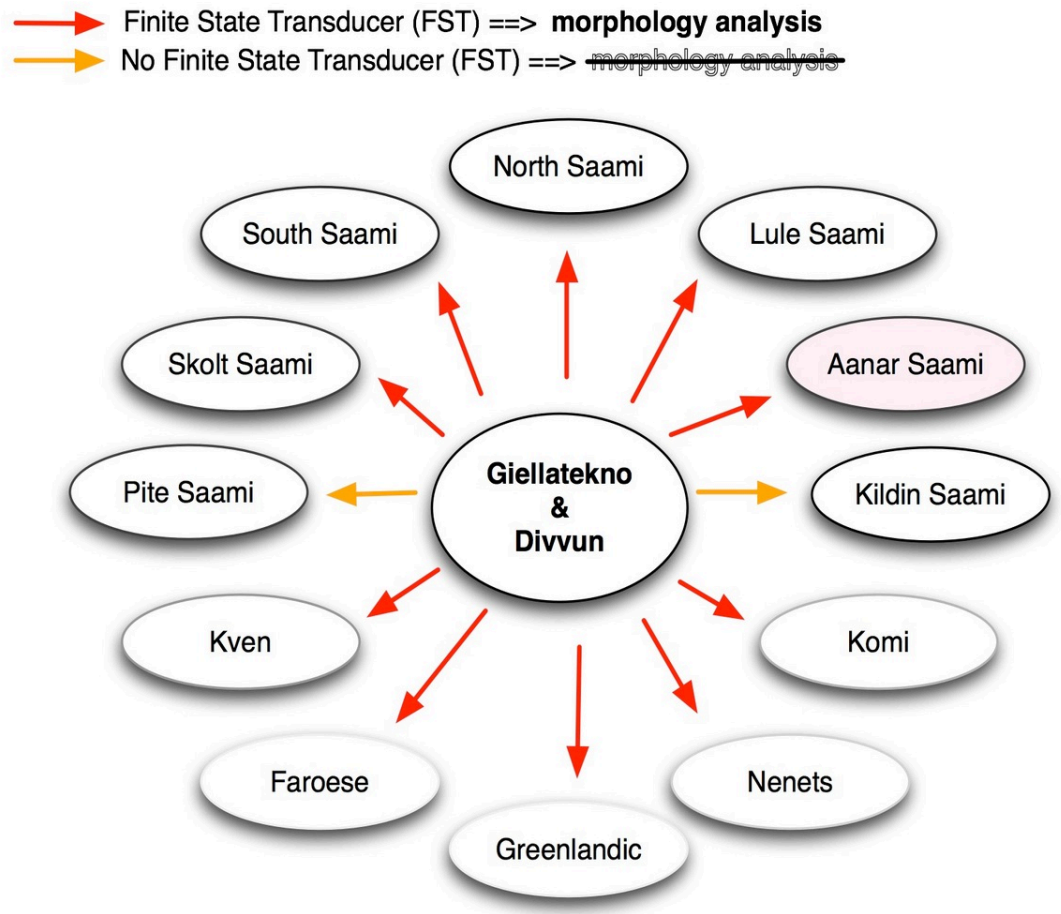
communication between researchers and language community

Worldwide use of *Giellatekno* & *Divvun* infrastructure

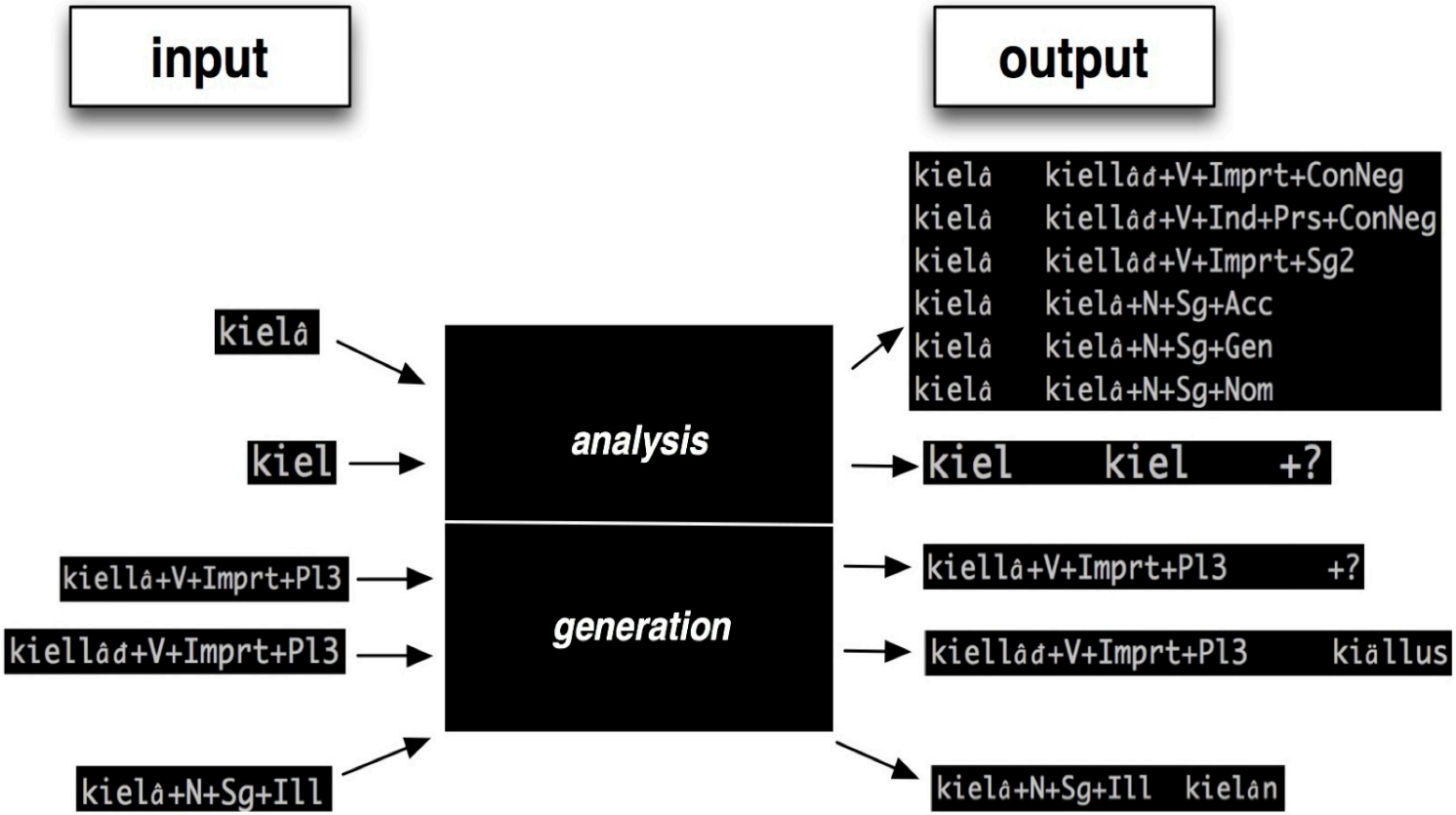
Before 2015


Finite State Transducer (FST) ==> **morphology analysis**
No Finite State Transducer (FST) ==> ~~morphology analysis~~
North Saami
South Saami
Lule Saami
Skolt Saami
Aanar Saami
Pite Saami
Giellatekno & Divvun
Kildin Saami
Kven
Komi
Faroese
Nenets
Greenlandic

After 2015


Finite State Transducer (FST) ==> **morphology analysis**
No Finite State Transducer (FST) ==> ~~morphology analysis~~
North Saami
South Saami
Lule Saami
Skolt Saami
Aanar Saami
Pite Saami
Giellatekno & Divvun
Kildin Saami
Kven
Komi
Faroese
Nenets
Greenlandic

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

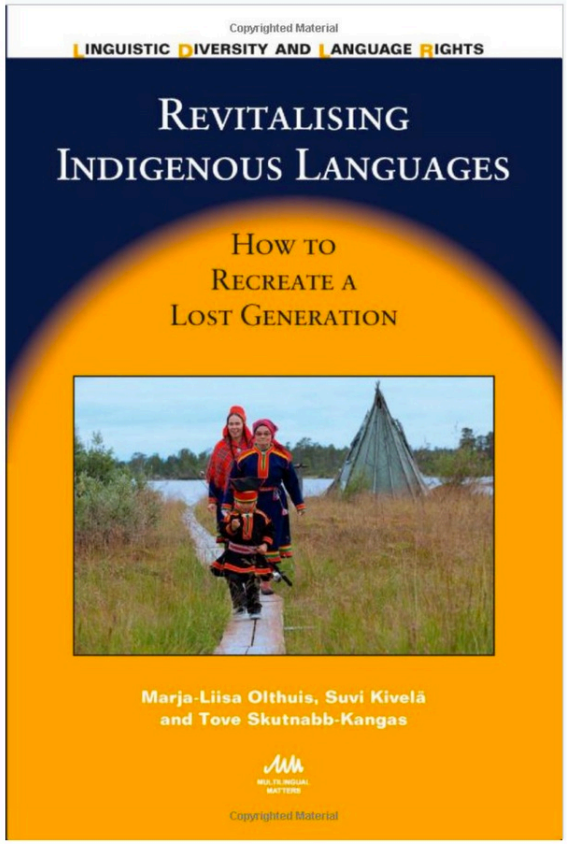
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

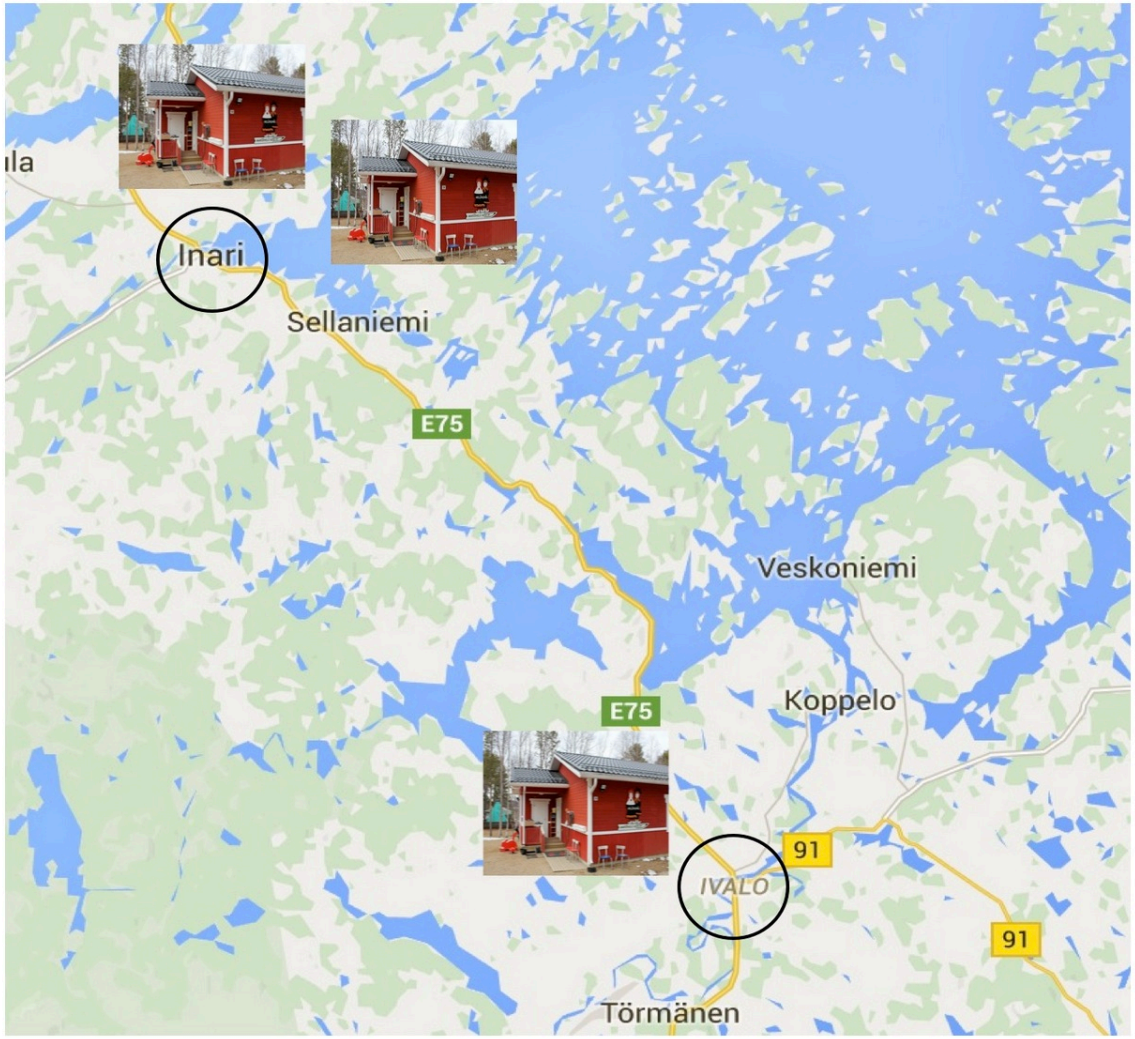
Challenges

communication between researchers and language community

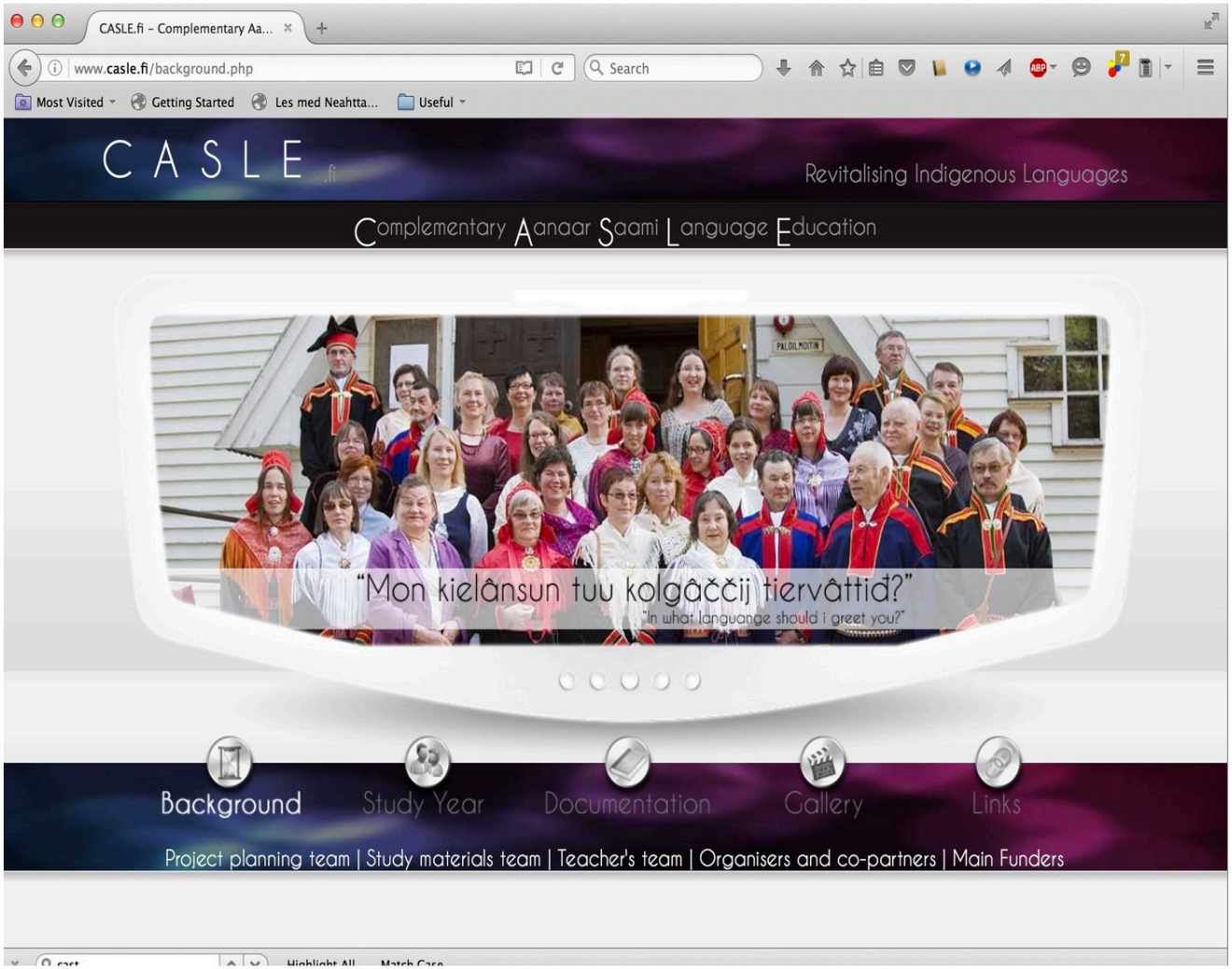
Aanaar Saami revitalisation



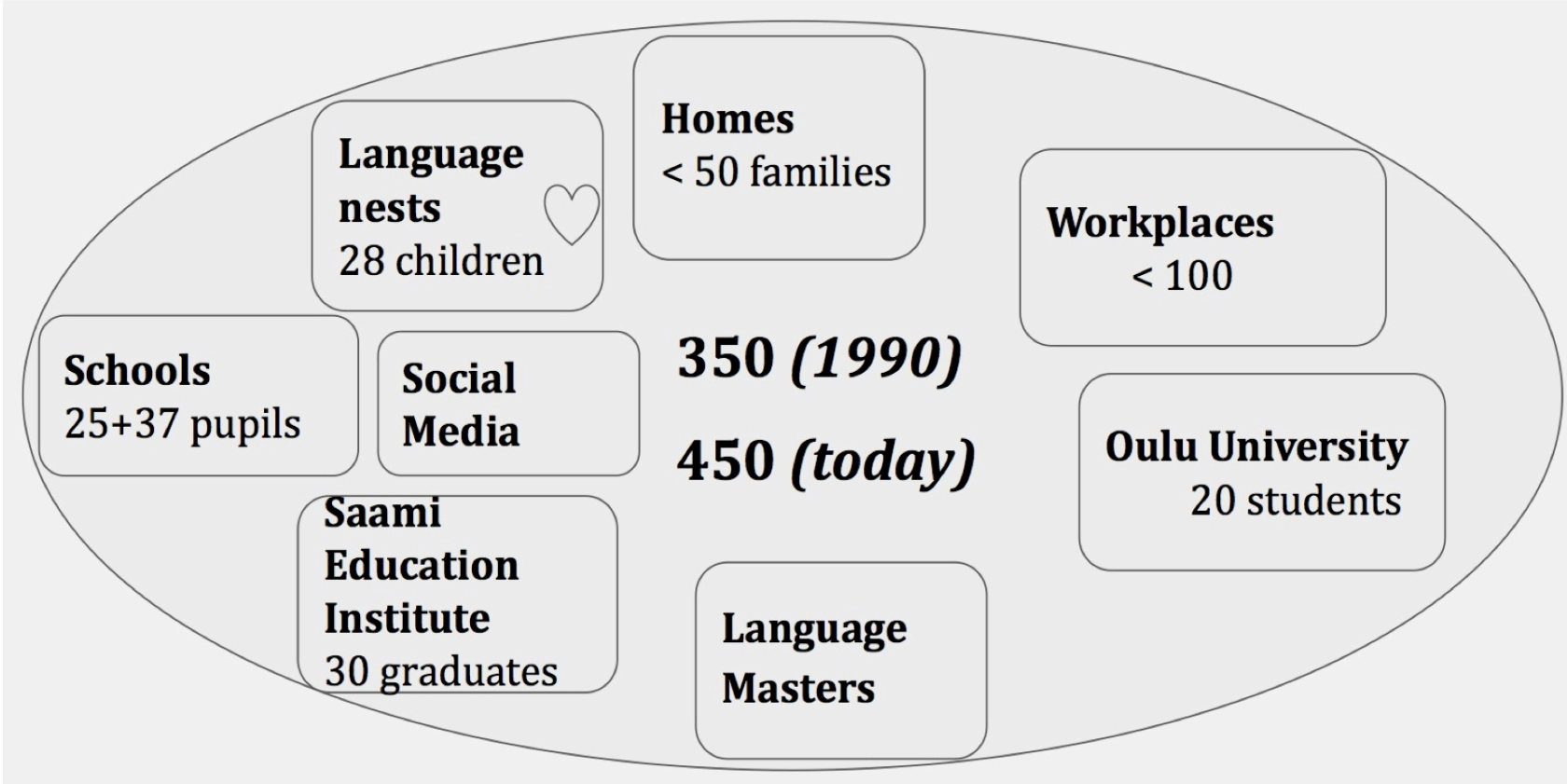
Language nests: child education



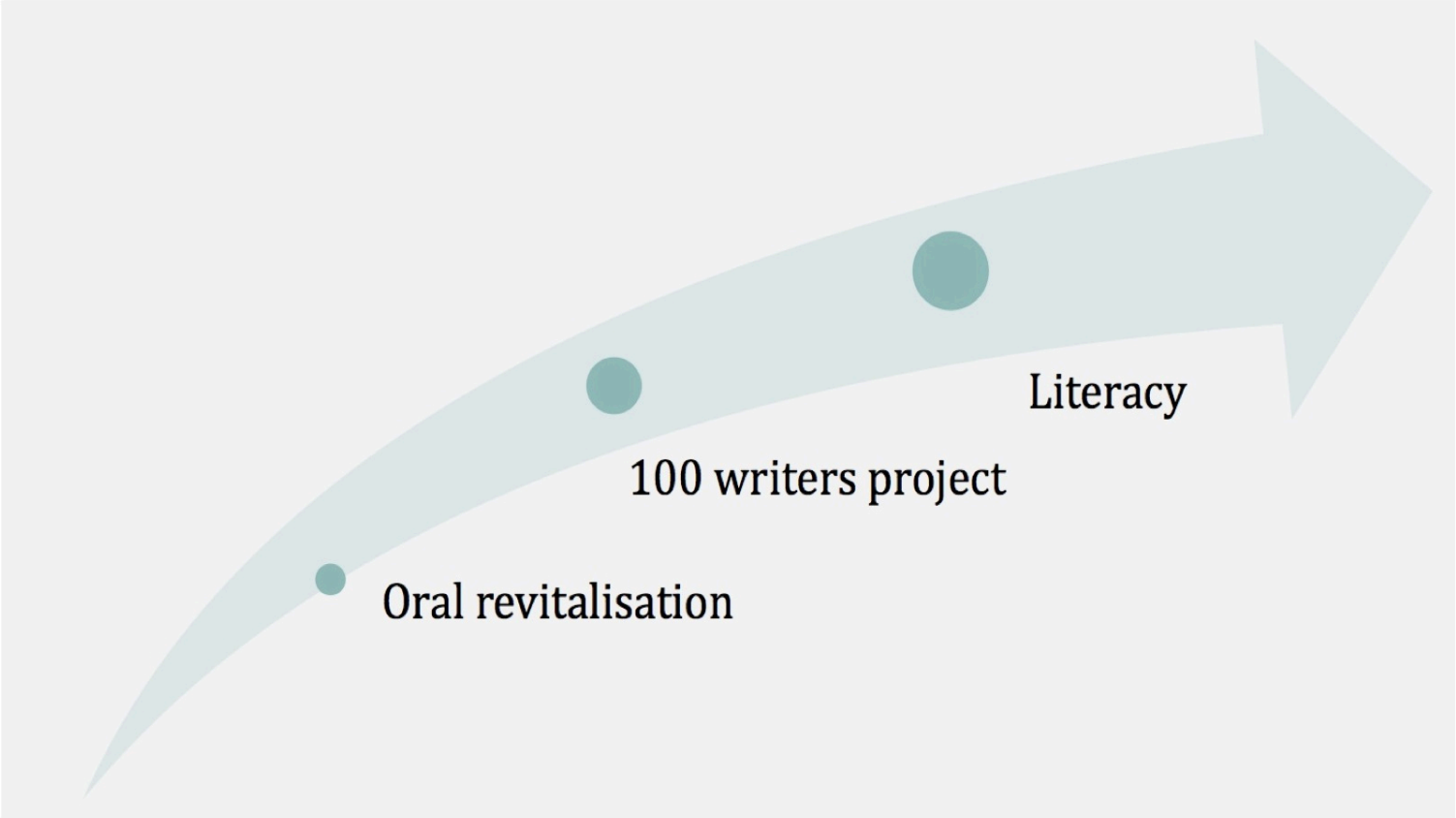
The CASLE project: adult education



Language transmission



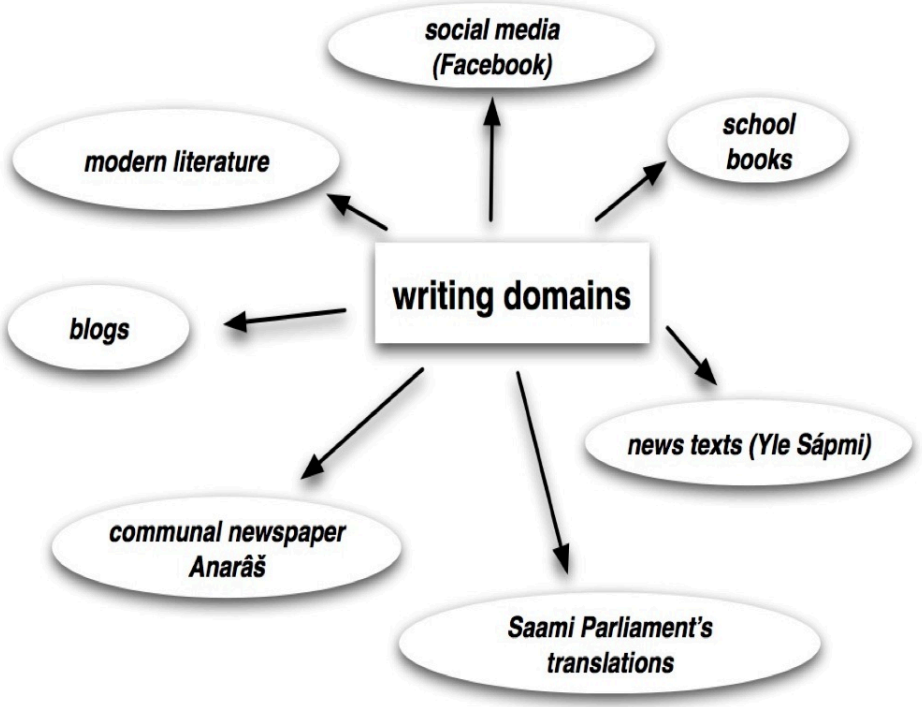
Aanaar Saami: a modern language in a modern society



Čyeti čälled: 100 writers

Estimated number of writers of Aanaar Saami

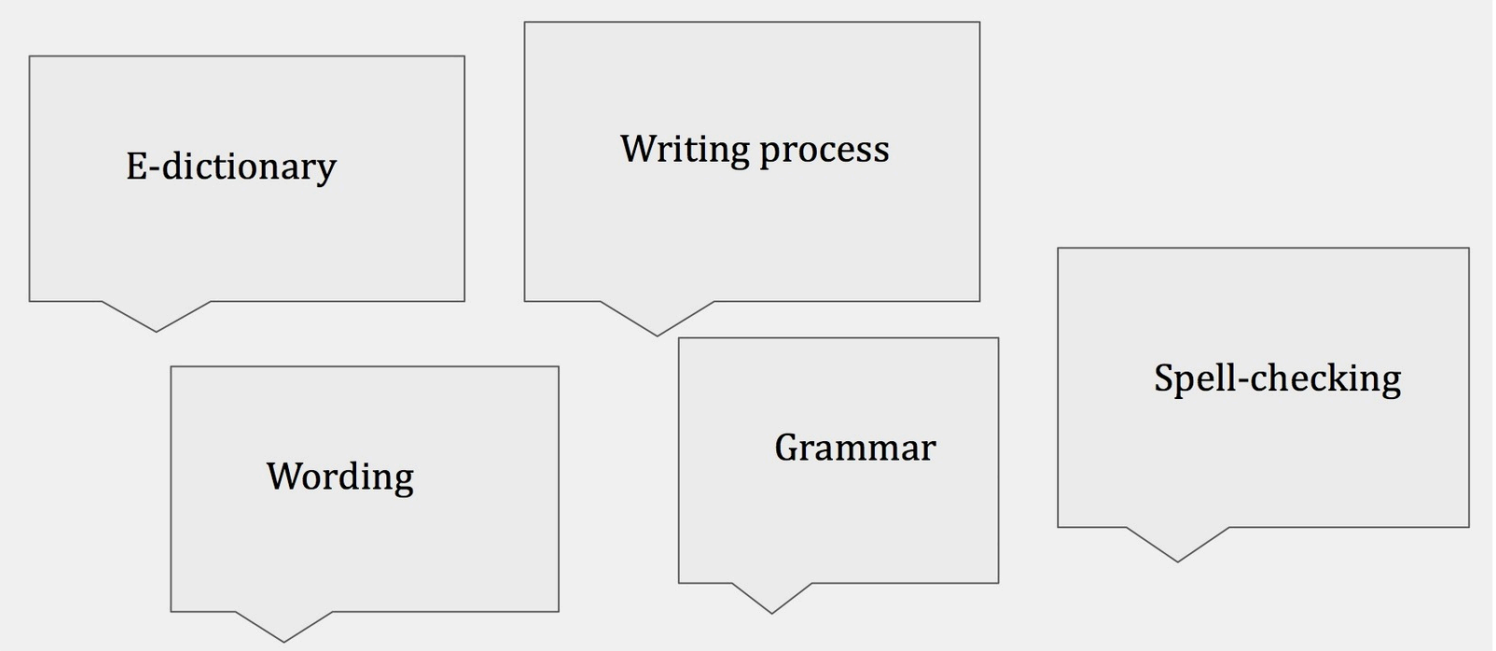
- **one very active fluent writer:** the editor-in-chief of Anarâš (70 +)
- **a few fluent younger writers:** writing every now and then
- **40 semi-fluent writers**



"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

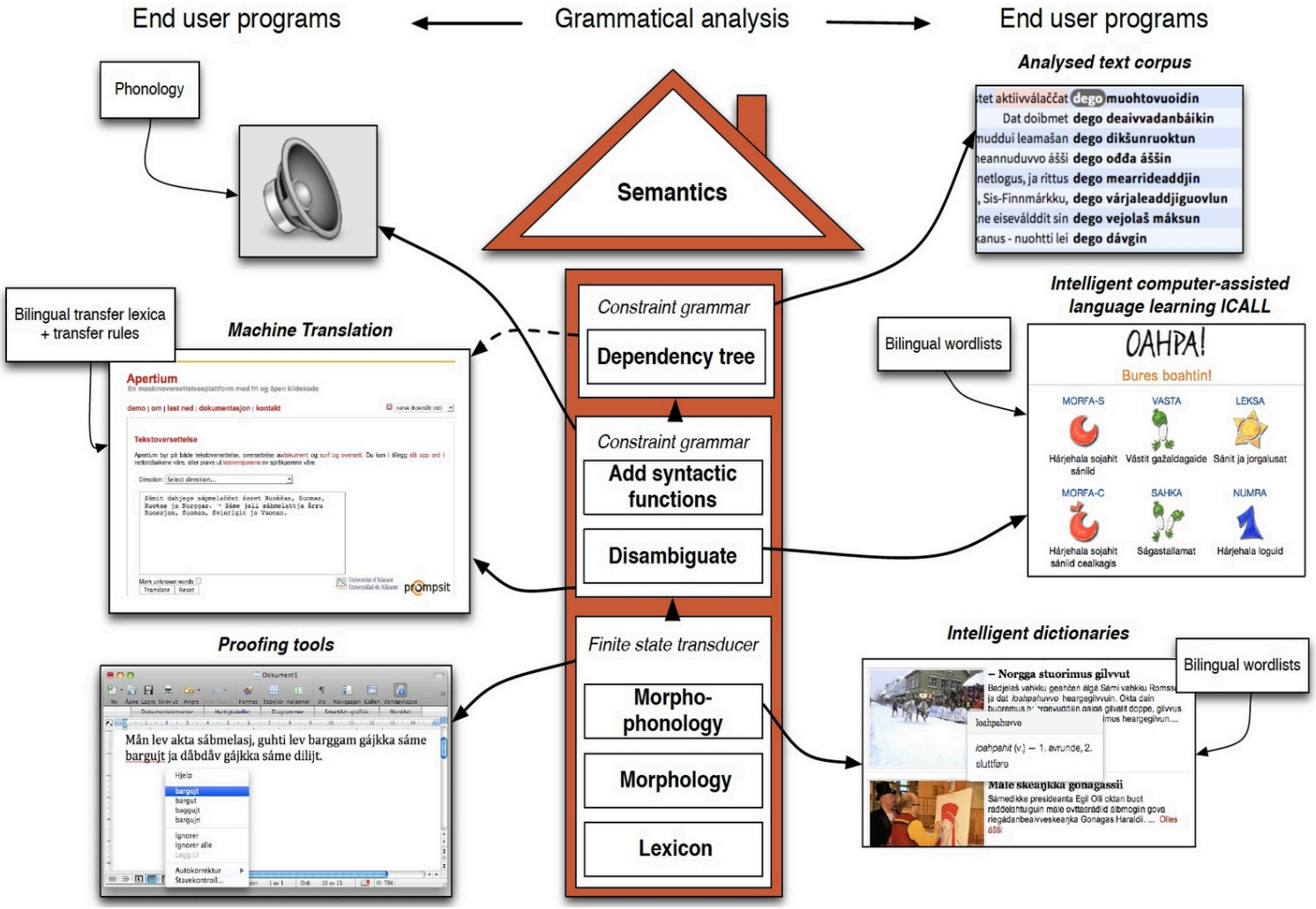
Language technology for language revitalisation

beyond Machine Translation as proof-of-concept

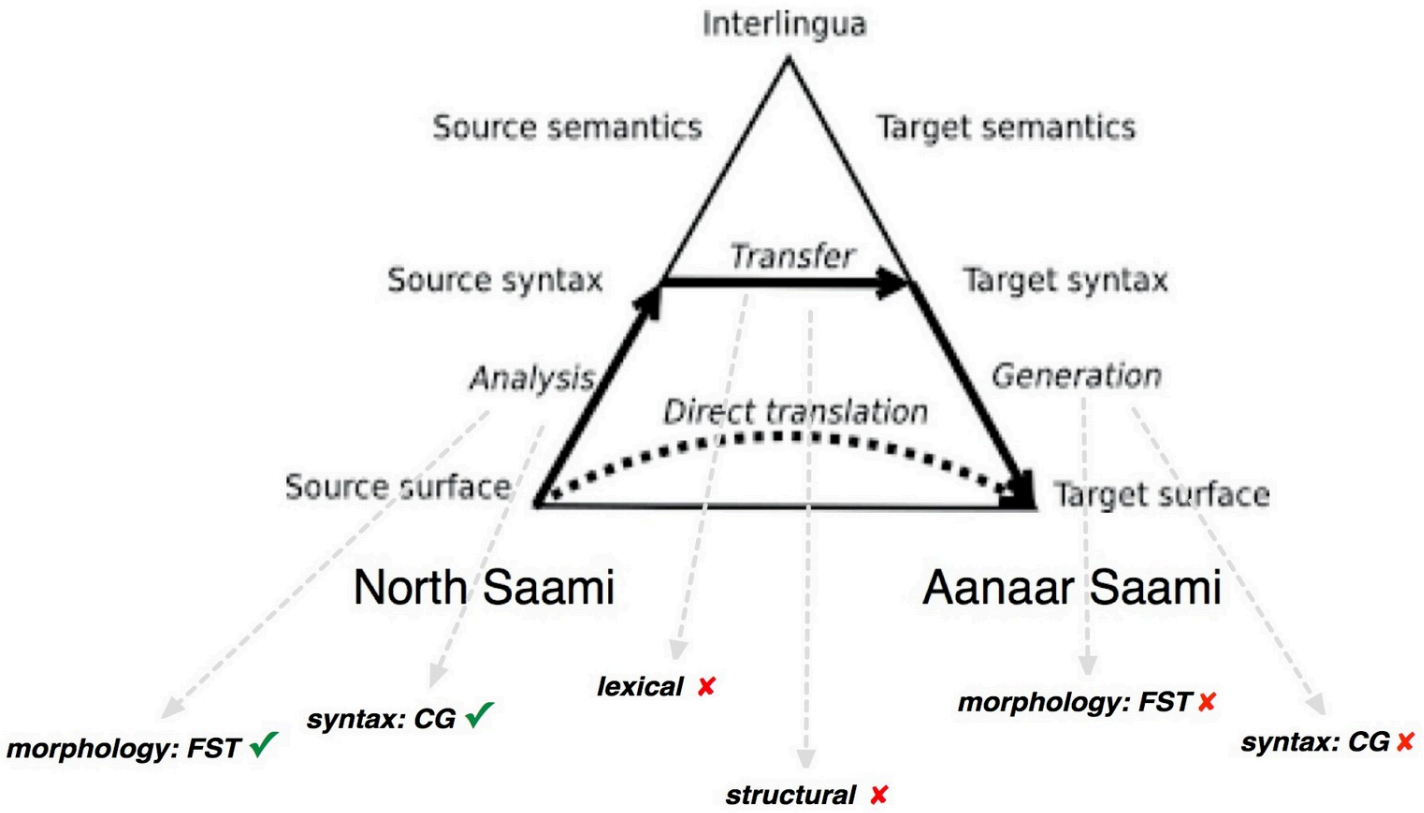
Challenges

communication between researchers and language community

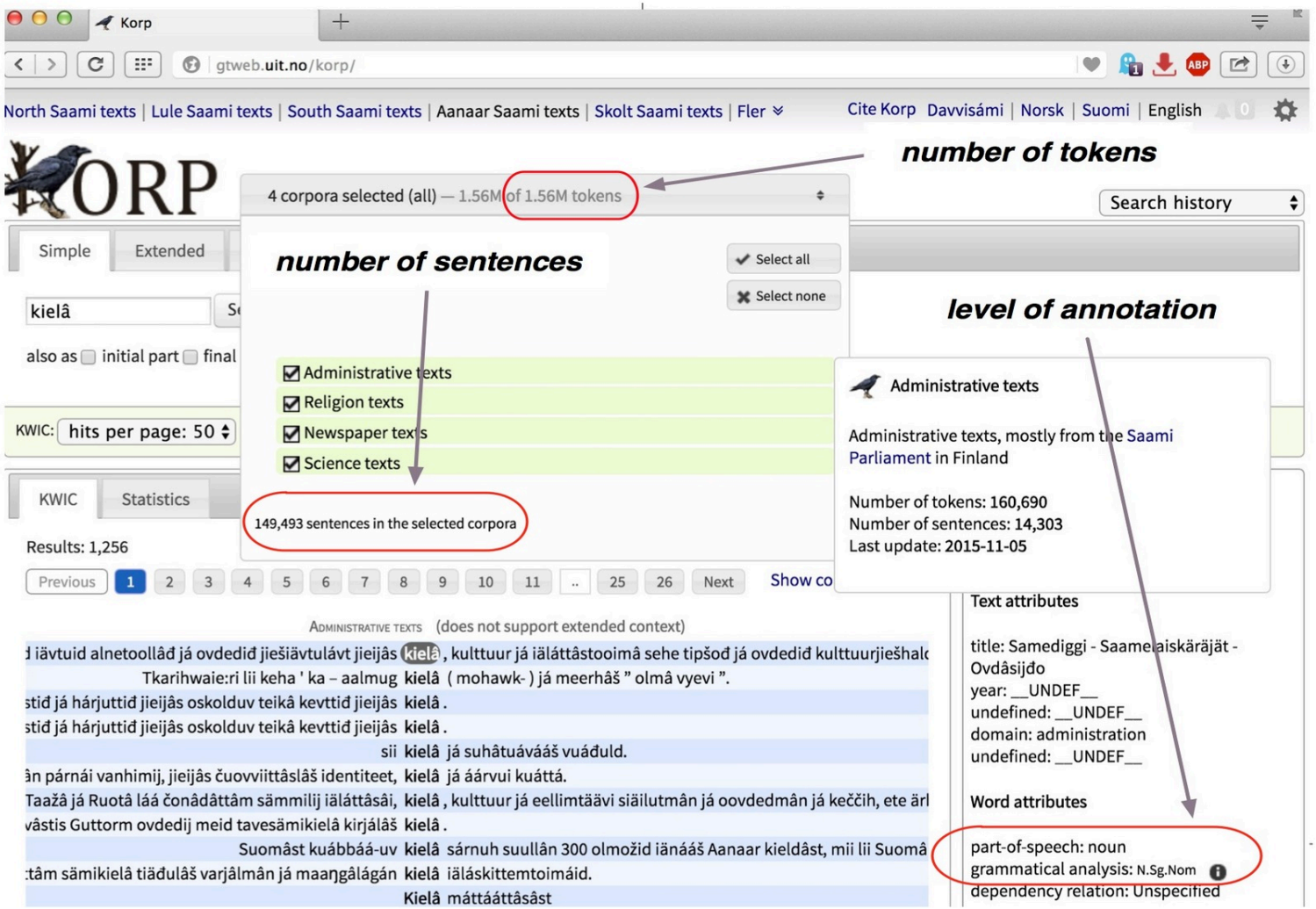
Saami Language Technology: *Giellatekno* & *Divvun*



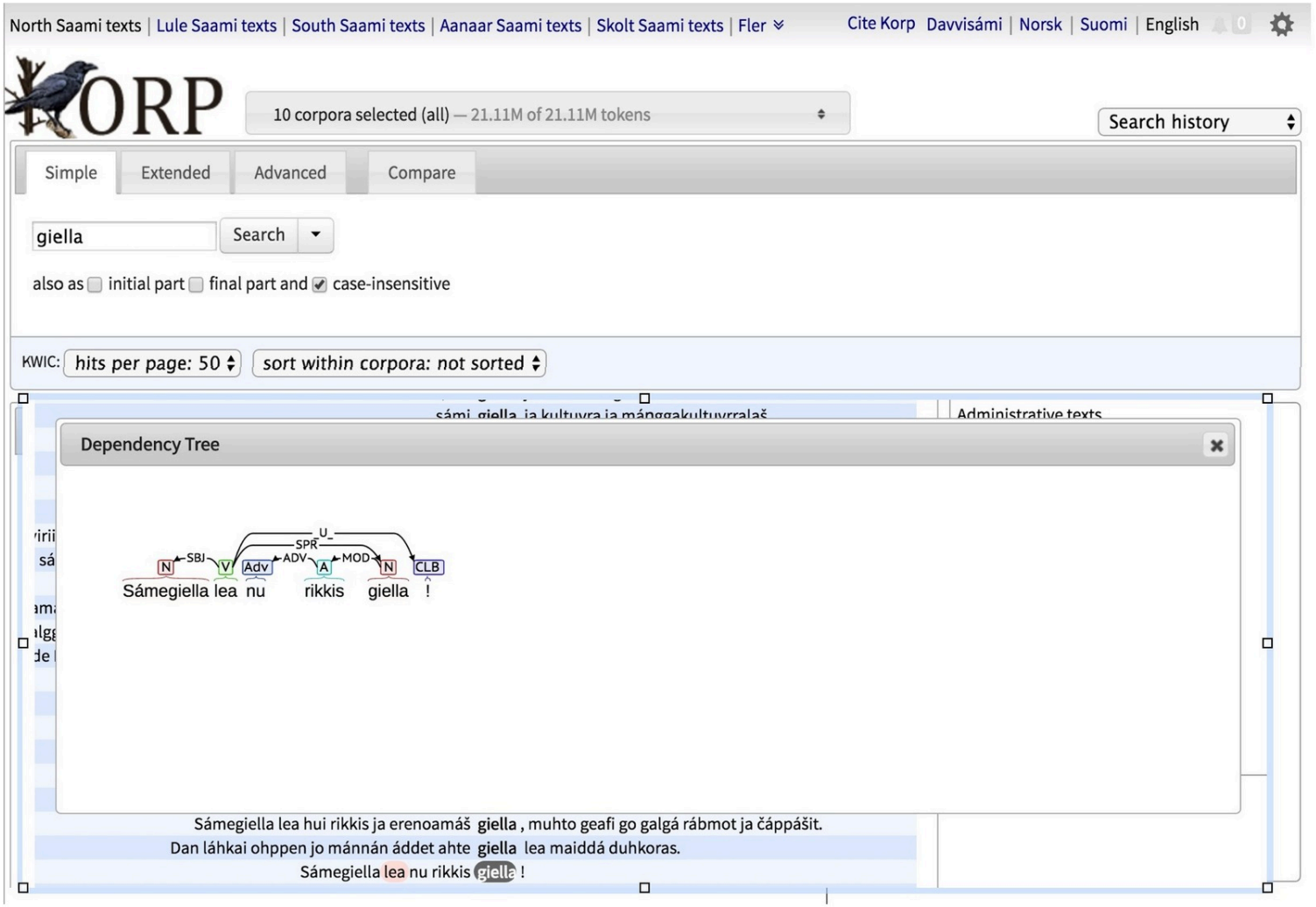
Machine Translation between minority languages



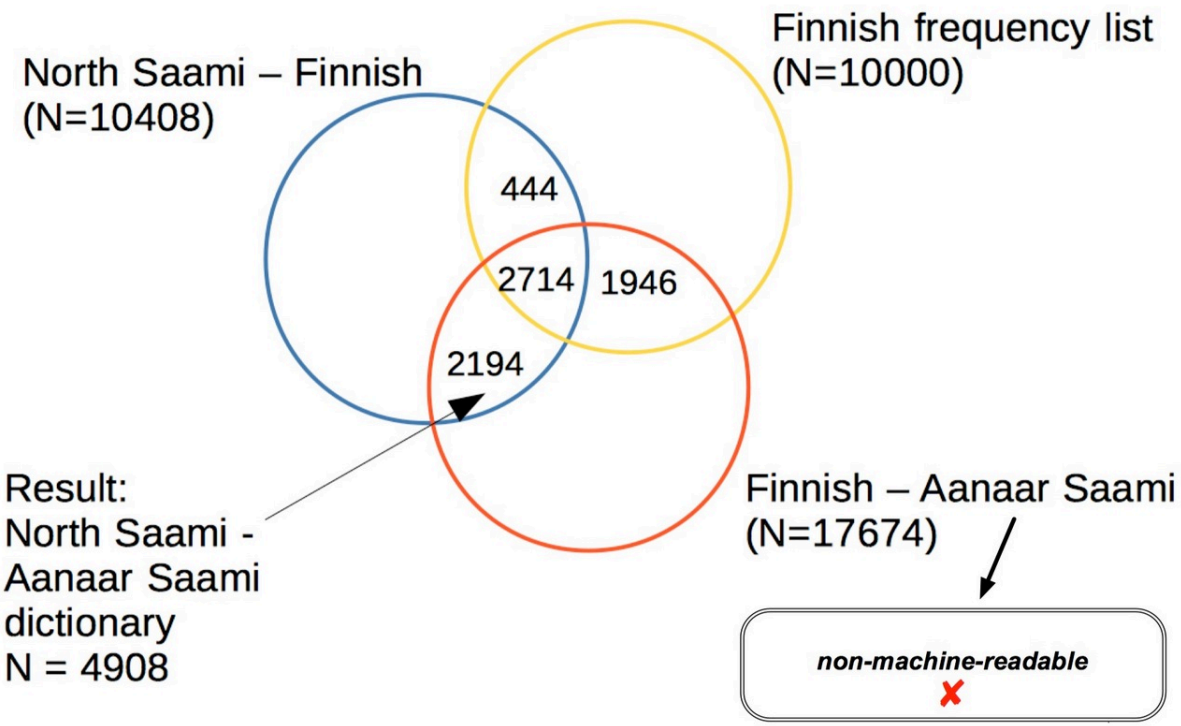
Resource building: monolingual corpus



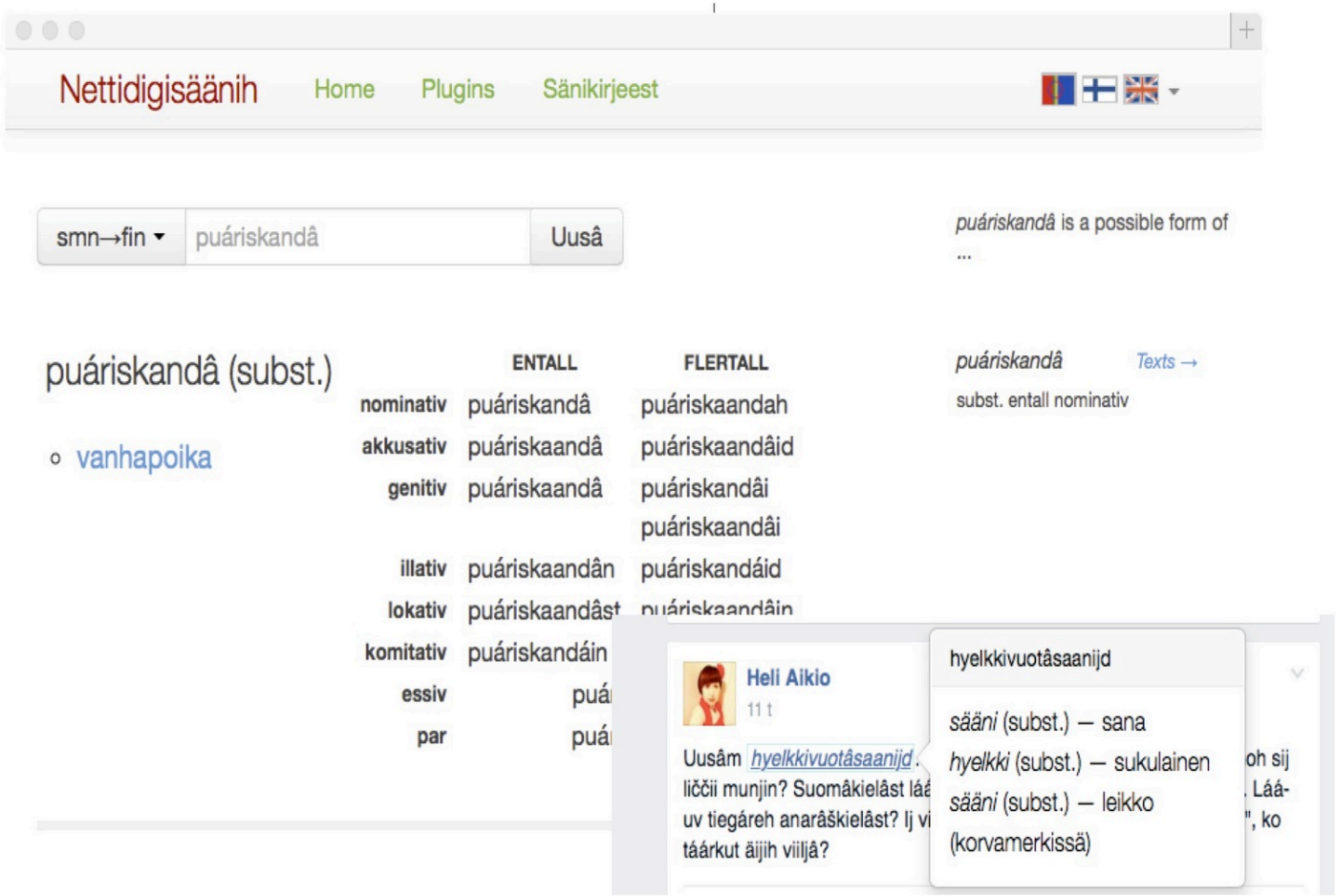
North Saami corpus



Resource building: bilingual dictionary



Online dictionary Nettidigisäänih (NDS)



Resource building: parallel corpus

North Saami ⟵ **Finnish** ⟶ **Aanaar Saami**

Size: 600 files

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

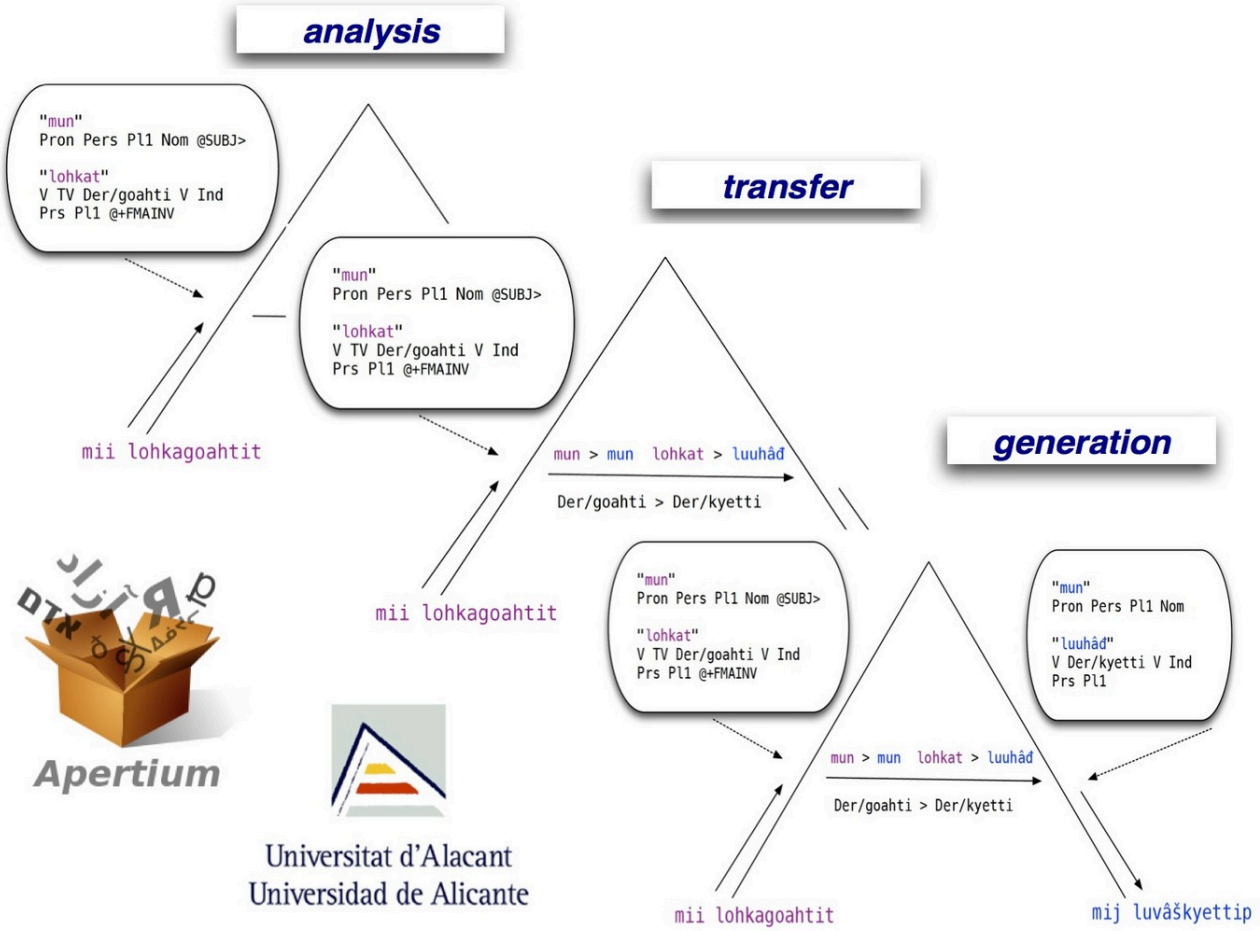
Format: Translation Memory eXchange (TMX)

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

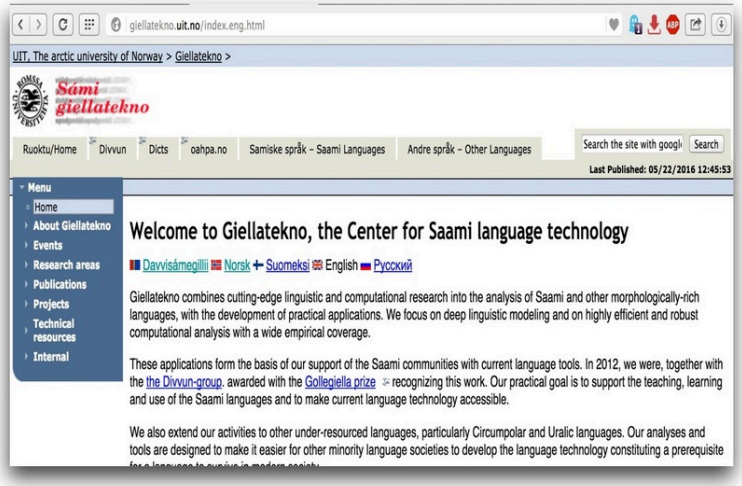
Usage: Evaluation of the Machine Translation output

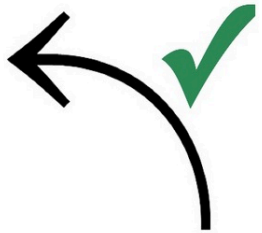
Senaatti-kiinteistöt vástida Sajosa hukseheamis ja hálddaša giddodaga .
Seenaat-kiddoduvah västid Saijoos huksiimist sehe haaldâš kiddoduv ton valmâštum maŋa .
*Senaatti-*kiinteistöt västid Sajos #huksiđ<vblex><der_tt><der_nomact><n><sg><loc> já haaldâš kiddoduv .

Machine Translation process



After 2016?


giellatekno.uit.no/index.eng.html
UiT, The arctic university of Norway > Giellatekno >
Sámi giellatekno
Ruoktu/Home
Divvun
Dicts
oahpa.no
Samiske språk – Saami Languages
Andre språk – Other Languages
Search the site with google
Search
Last Published: 05/22/2016 12:45:53
Menu
Home
About Giellatekno
Events
Research areas
Publications
Projects
Technical resources
Internal
Welcome to Giellatekno, the Center for Saami language technology
Davvisámegillii Norsk Suomeksi English Русский
Giellatekno combines cutting-edge linguistic and computational research into the analysis of Saami and other morphologically-rich languages, with the development of practical applications. We focus on deep linguistic modeling and on highly efficient and robust computational analysis with a wide empirical coverage.
These applications form the basis of our support of the Saami communities with current language tools. In 2012, we were, together with the the Divvun-group, awarded with the Gollegiella prize recognizing this work. Our practical goal is to support the teaching, learning and use of the Saami languages and to make current language technology accessible.
We also extend our activities to other under-resourced languages, particularly Circumpolar and Uralic languages. Our analyses and tools are designed to make it easier for other minority language societies to develop the language technology constituting a prerequisite




?


"Mon kielânsun tuu kolgâččij tiervâttid?"
"In what language should i greet you?"

Overview

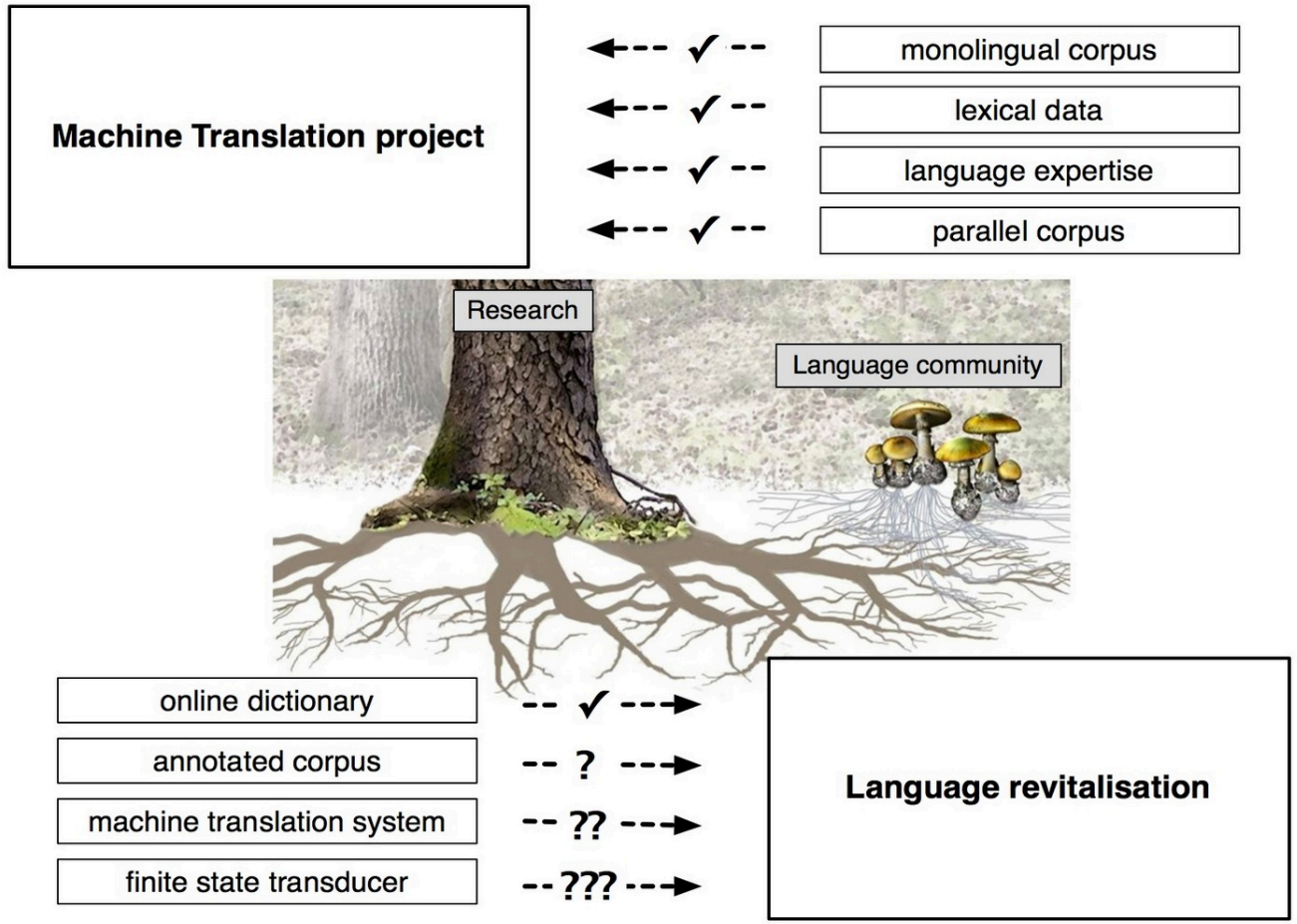
Aanaar Saami language community
the revitalisation of a small indigenous language

Saami Language Technology
the Machine Translation project between North Saami and Aanaar Saami

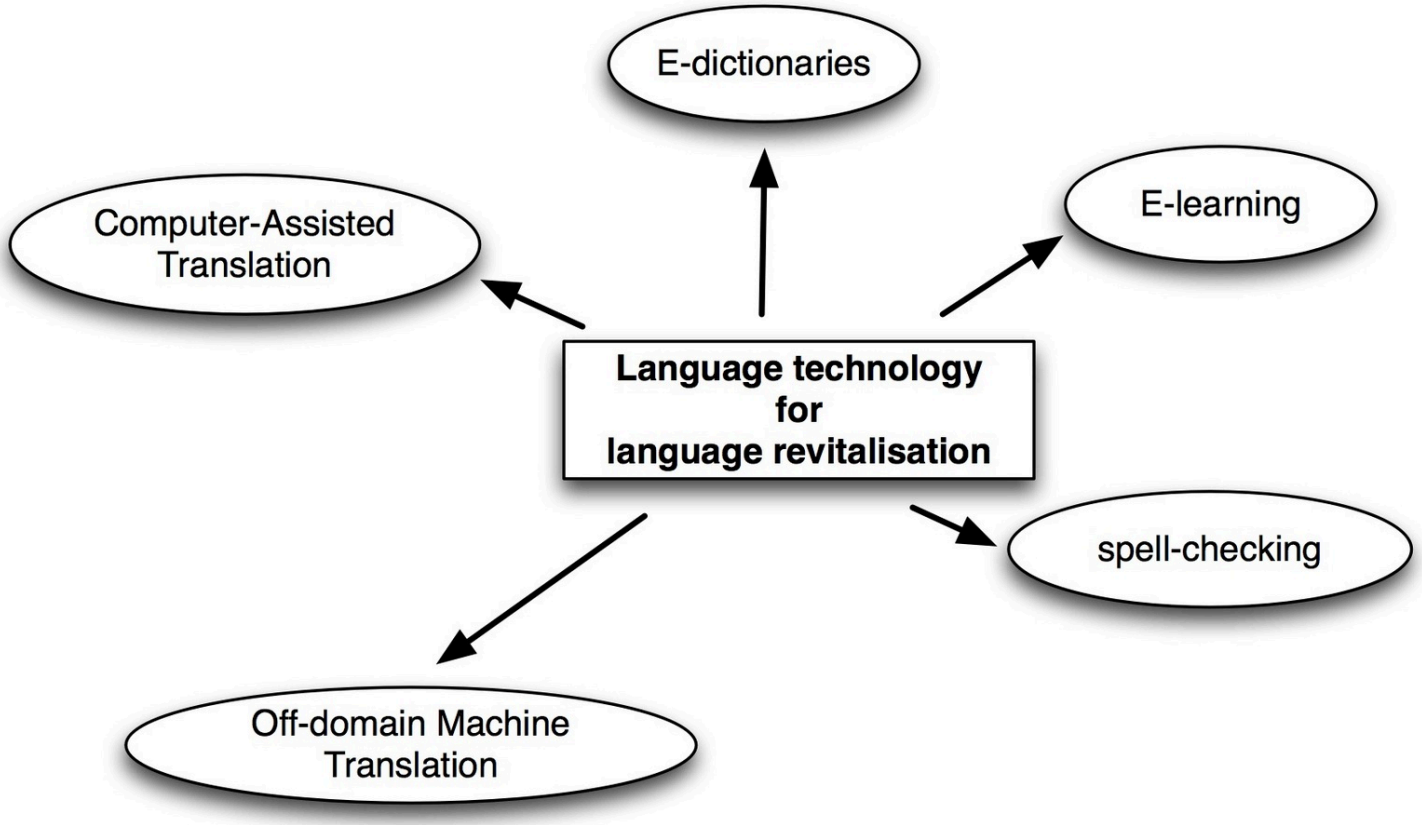
Language technology for language revitalisation
beyond Machine Translation as proof-of-concept

Challenges
communication between researchers and language community

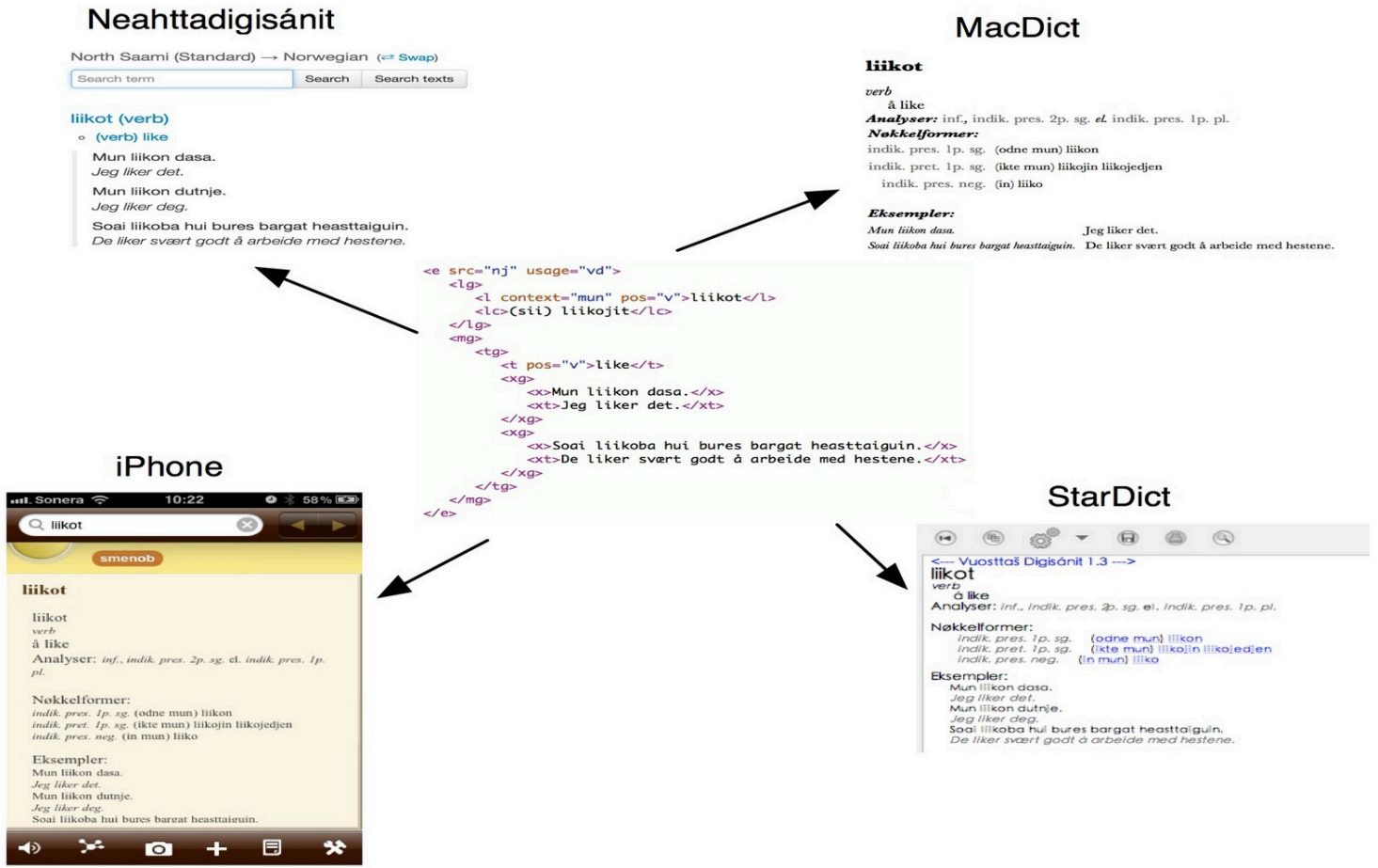
Research-Revitalisation symbiosis



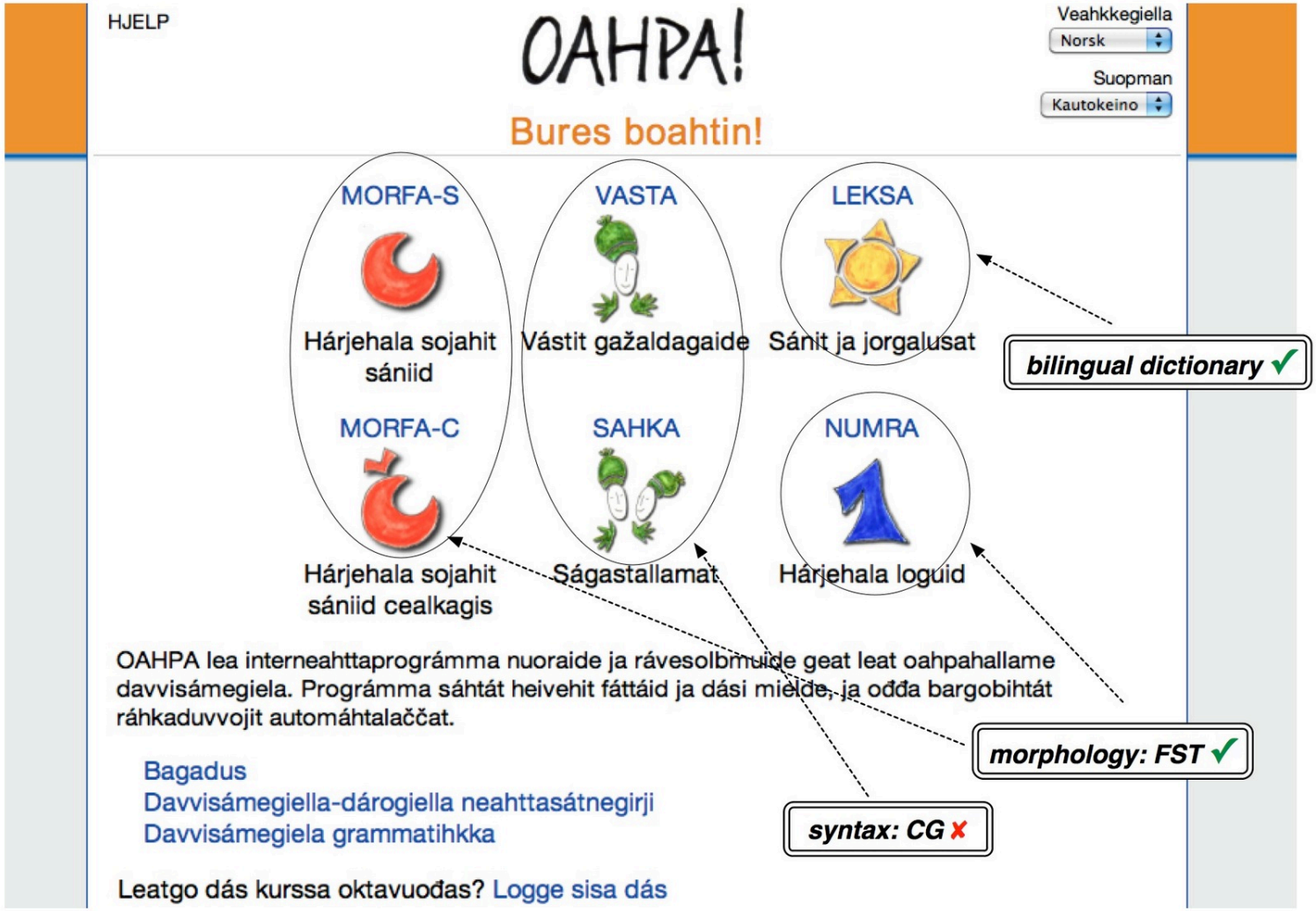
Applications for language revitalisation



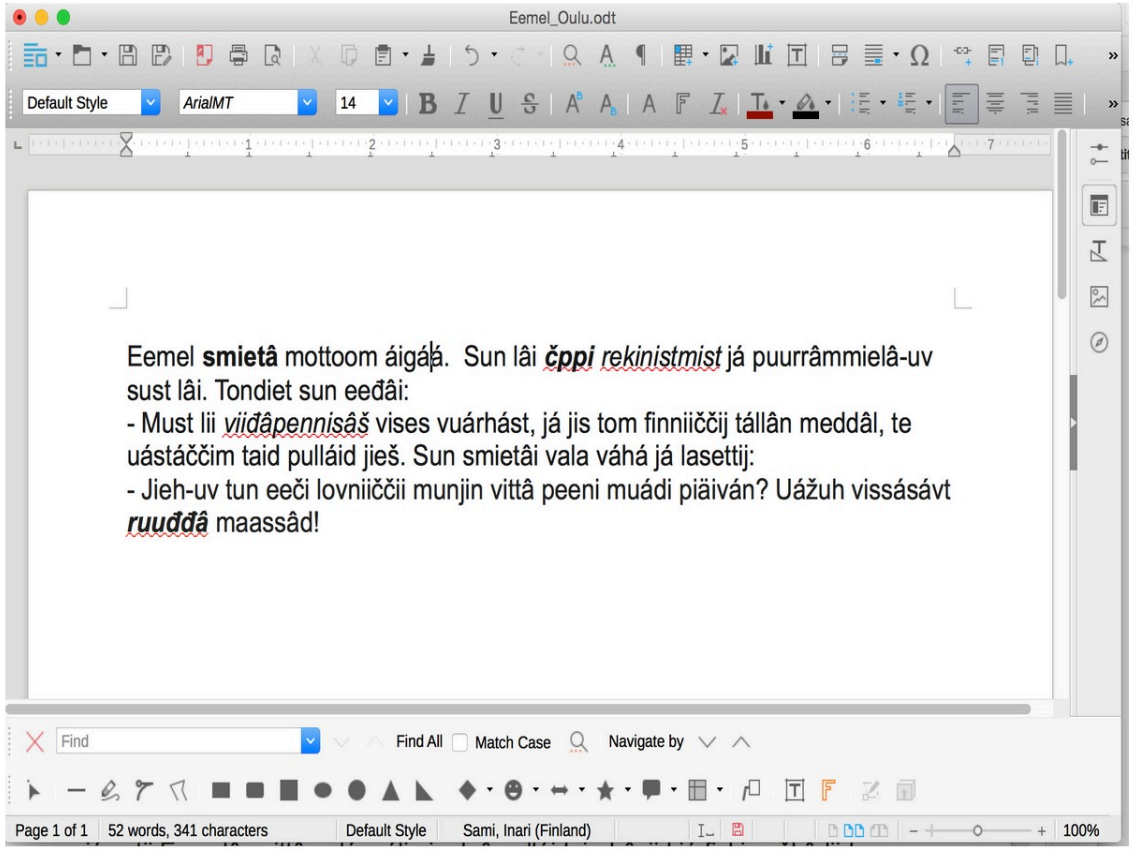
Cross-platform online/offline intelligent dictionaries



ICALL application: Oahpa



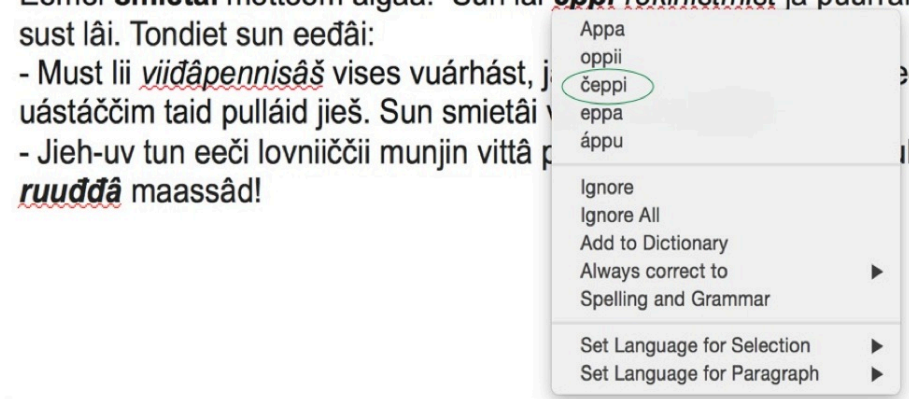
Spellchecker: non-detectable word



→ need for a grammar checker

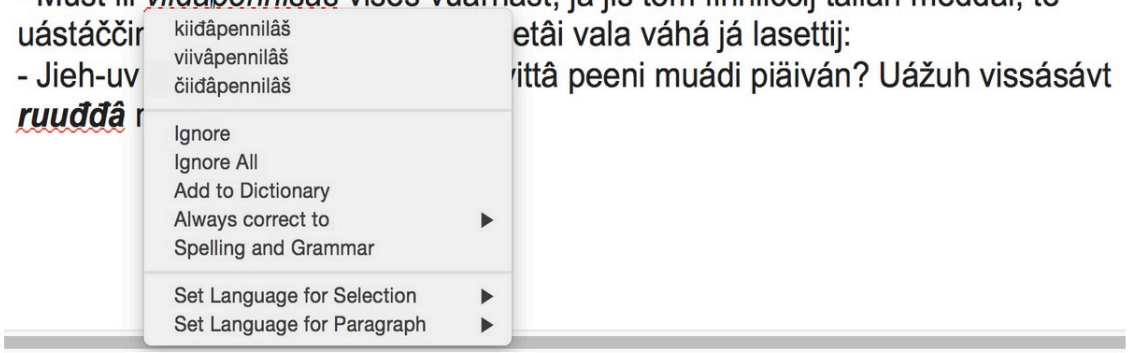
Spellchecker: correcting a typo

Eemel **smietâi** mottoom áigáá. Sun lâi ***čppi*** *rekinistmist* já puurrâmmielâ-uv sust lâi. Tondiet sun eeđâi:
- Must lii *viiđâpennisâš* vises vuárhást, já jis tom finniiččij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniiččii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ***ruuđđâ*** maassâd!



Eemel **smietâi** mottoom áigáá. Sun lâi ***čeppi*** *rekinistmist* já puurrâmmielâ-uv sust lâi. Tondiet sun eeđâi:
- Must lii *viiđâpennisâš* vises vuárhást, já jis tom finniiččij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniiččii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ***ruuđđâ*** 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" ;    !
viiđâpennisâš:viiđâ#pennis 4LAS_NOUN "viisipenninen" ;  !
sajasâš:sajas 4LAS_NOUN "viransijainen" ;    !
čiähásiemmânsâš:čiähá#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

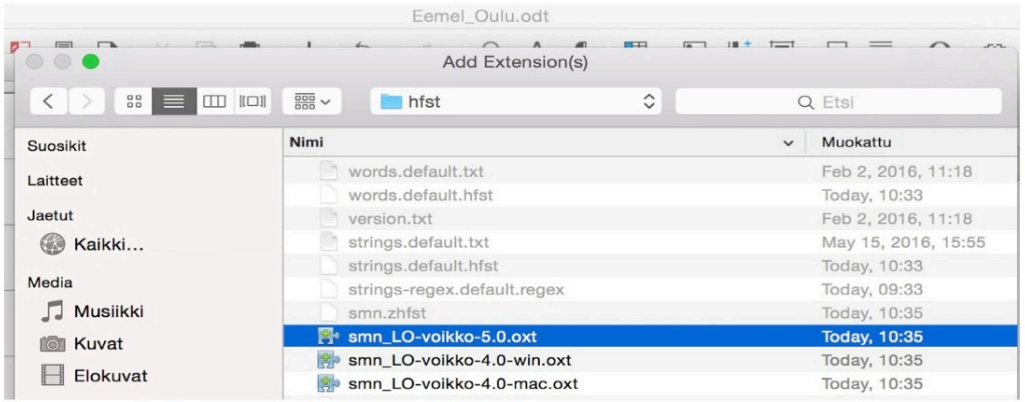
```
tf4-hsl-m0019:smn marjaliisa$ make

*** 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

```
smn — bash — 112×28
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$
```

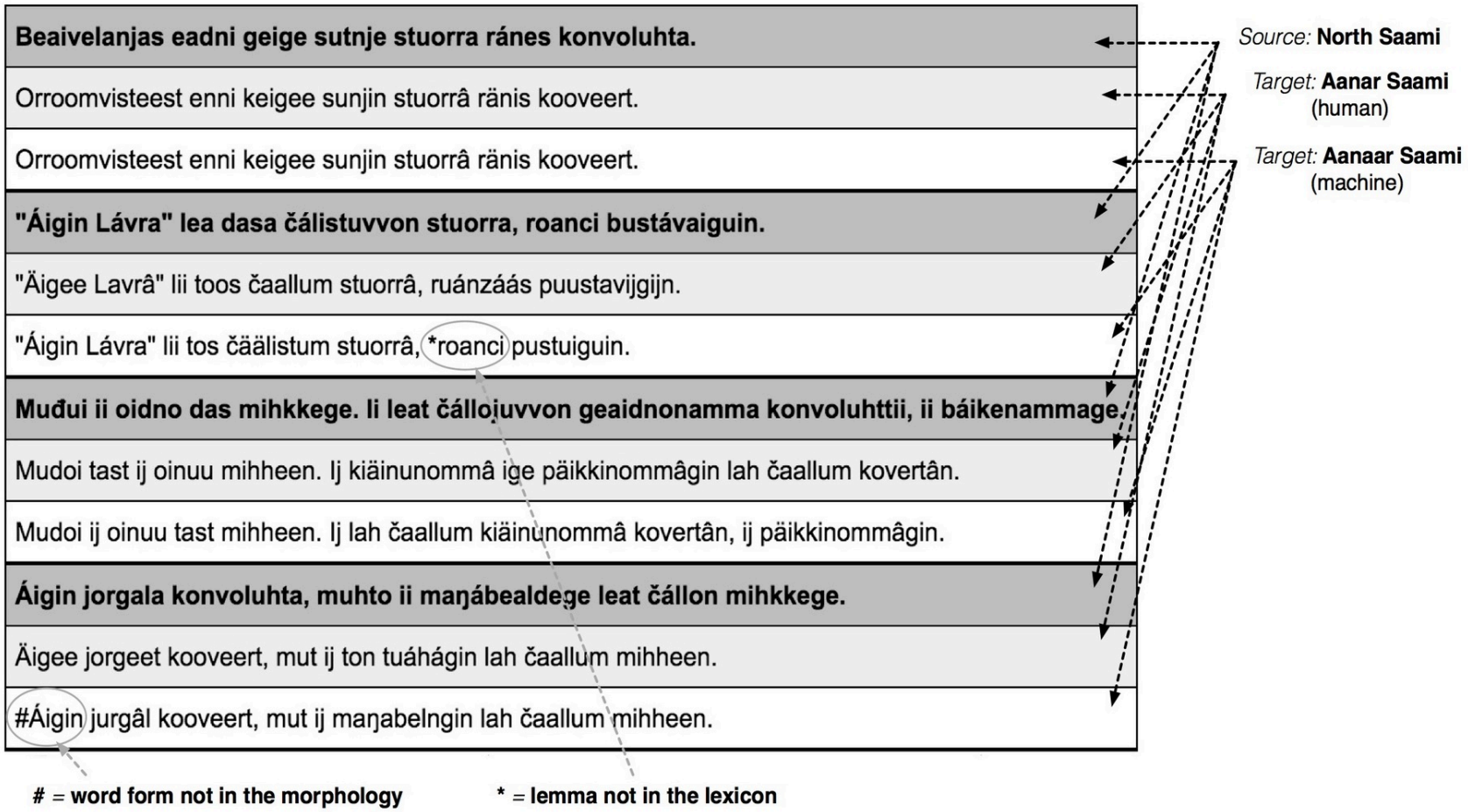
==> add the plugin to the editing tool



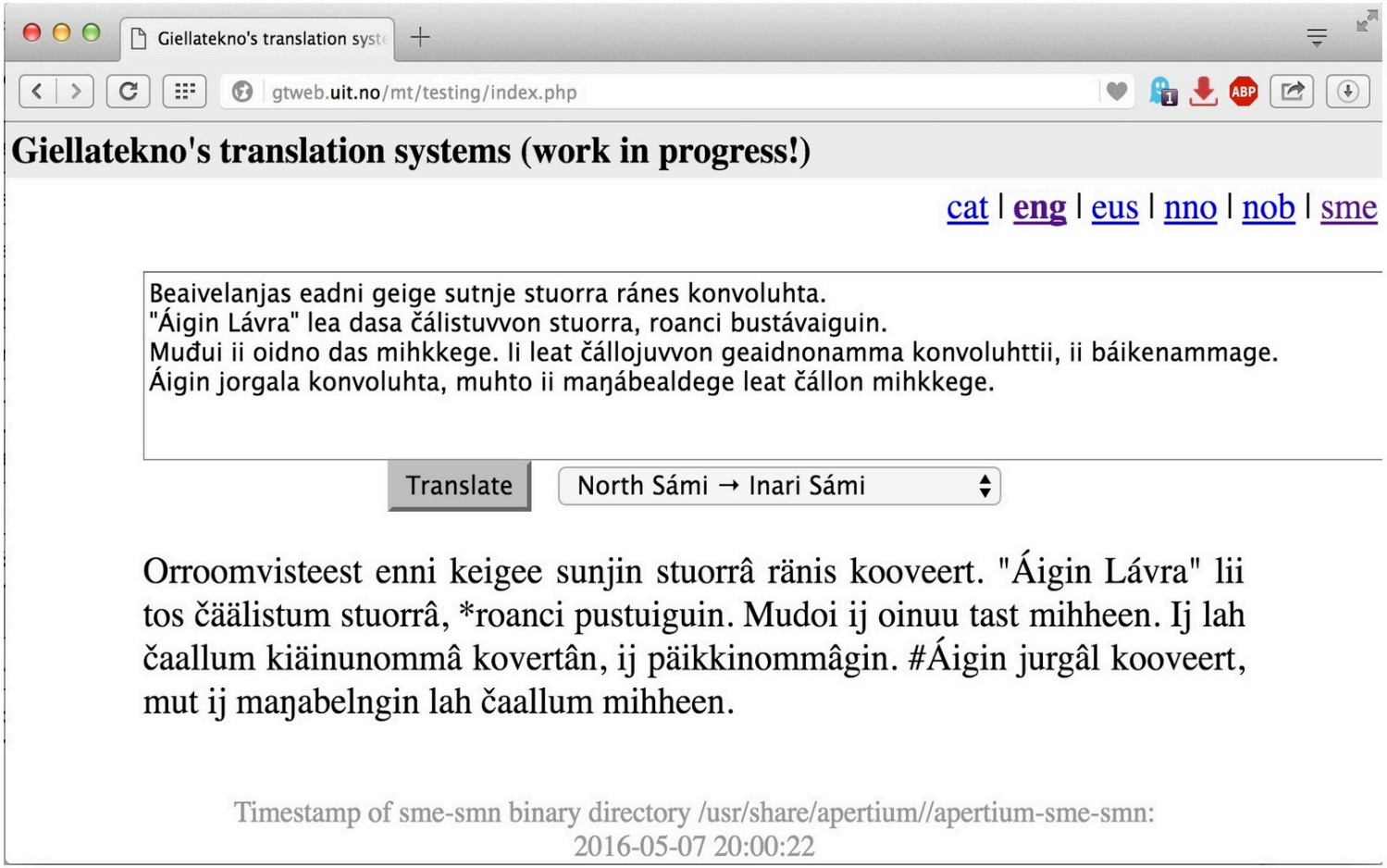
==> reopen the text file

Eemel **smietâi** mottoom áigáá. Sun lâi ***čeppi*** *rekinistmist* já puurrâmmielâ-uv sust lâi. Tondiet sun eeđâi:
- Must lii *viiđâpennisâš* vises vuárhást, já jis tom finniiččij tállân meddâl, te uástáččim taid pulláid jieš. Sun smietâi vala váhá já lasettij:
- Jieh-uv tun eeči lovniiččii munjin vittâ peeni muádi piäiván? Uážuh vissásávt ***ruuđđâ*** maassâd!

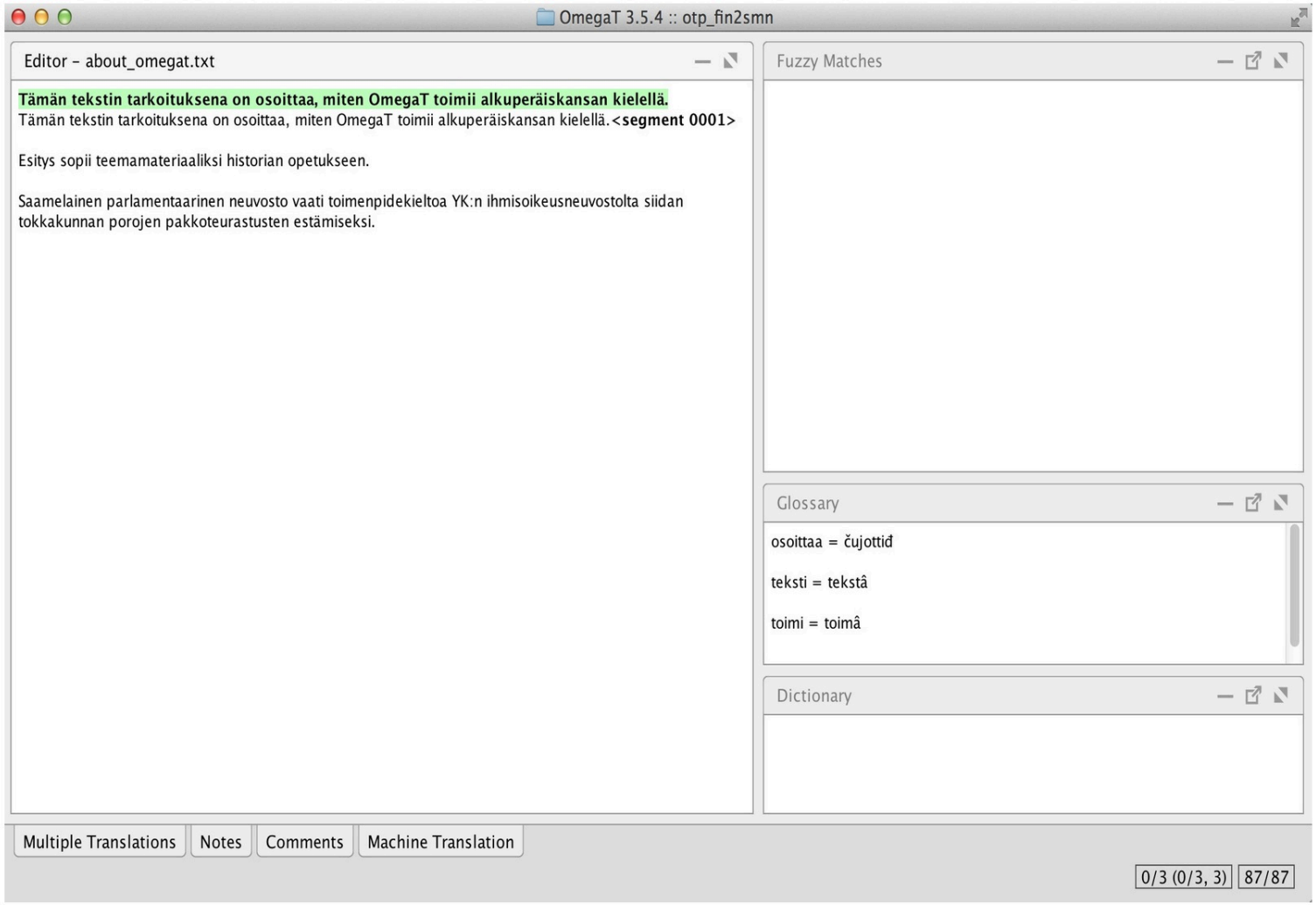
Off-domain Machine Translation test: children's books



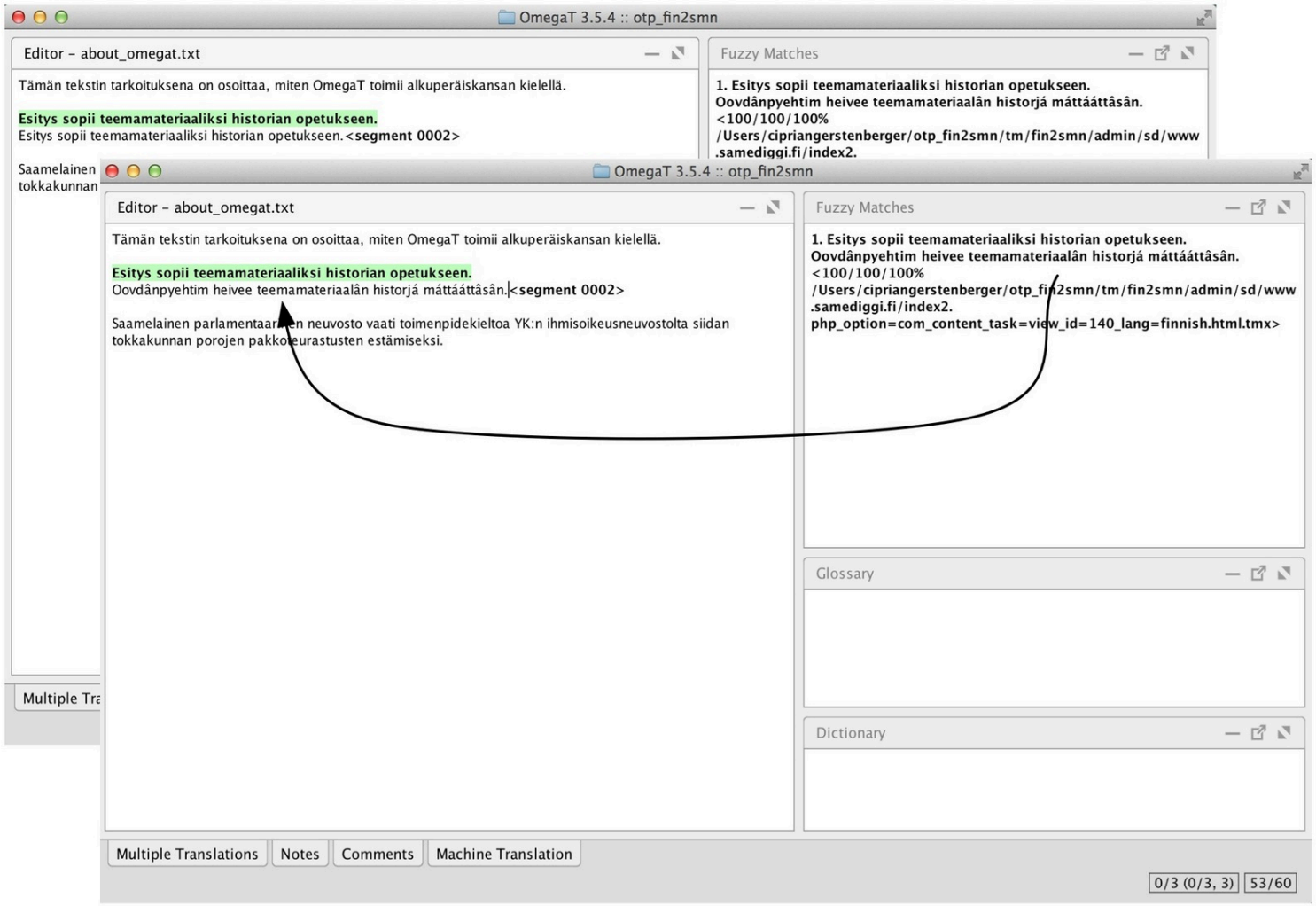
Online use of the Machine Translation system



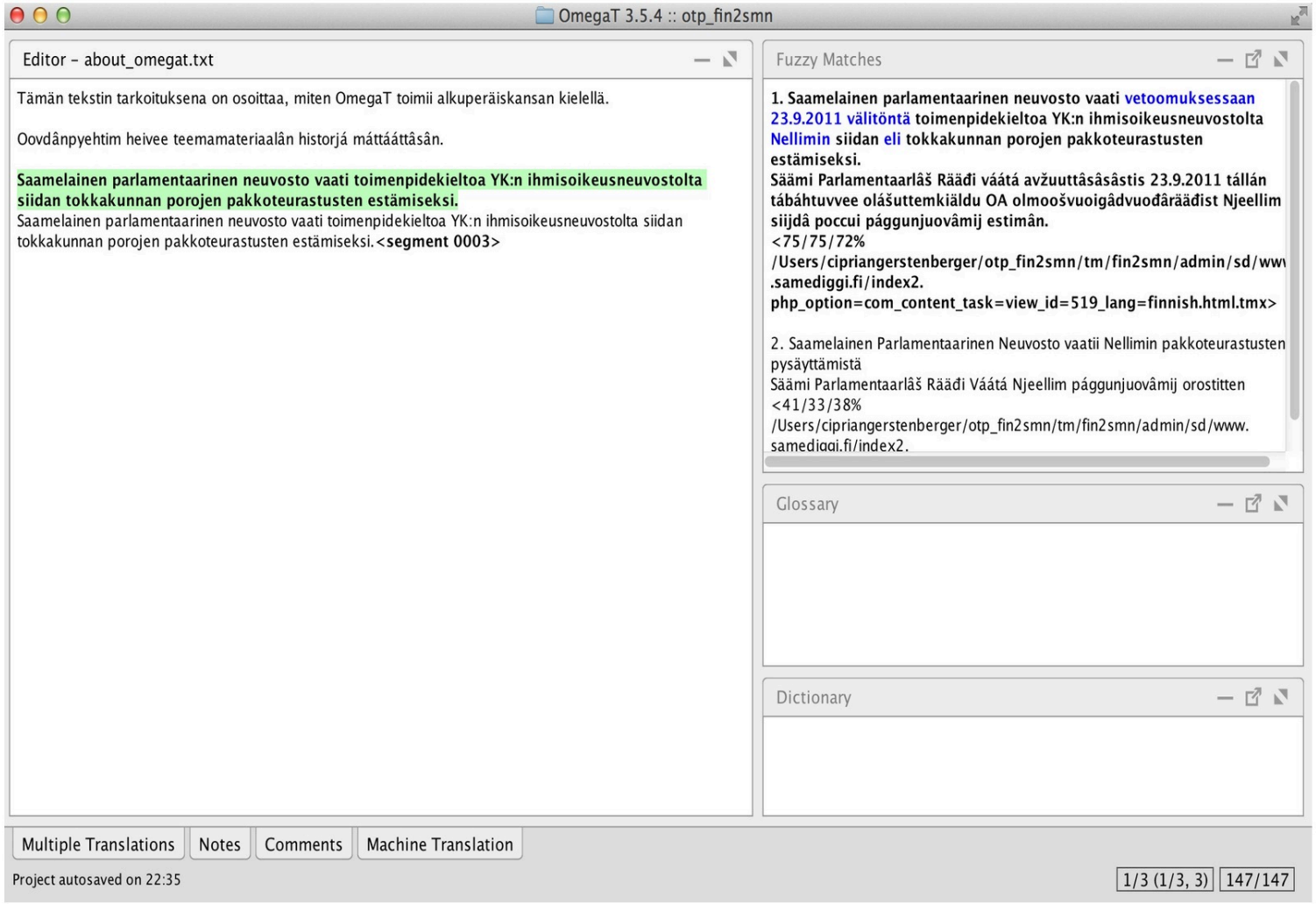
Computer-Assisted Translation (CAT)


OmegaT 3.5.4 :: otp_fin2smn
Editor – about_omegat.txt
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.
Fuzzy Matches
Glossary
osoittaa = čujottiđ
teksti = tekstâ
toimi = toimâ
Dictionary
Multiple Translations
Notes
Comments
Machine Translation
0/3 (0/3, 3)
87/87

CAT: total string match


OmegaT 3.5.4 :: otp_fin2smn
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>
Saamelainen
tokkakunnan
Fuzzy Matches
1. Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdânpyehtim heivee teemamateriaalân historjá máttááttâsân.
<100/100/100%
/Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www
.samediggi.fi/index2.
OmegaT 3.5.4 :: otp_fin2smn
Editor – about_omegat.txt
Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.
Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdânpyehtim heivee teemamateriaalân historjá máttááttâsân.<segment 0002>
Saamelainen parlamentaarikkojen neuvosto vaati toimenpidekieltoa YK:n ihmisoikeusneuvostolta siidan
tokkakunnan porojen pakkoteurastusten estämiseksi.
Fuzzy Matches
1. Esitys sopii teemamateriaaliksi historian opetukseen.
Oovdânpyehtim heivee teemamateriaalân historjá máttááttâsân.
<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>
Glossary
Dictionary
Multiple Translations
Notes
Comments
Machine Translation
0/3 (0/3, 3)
53/60

CAT: partial string match


OmegaT 3.5.4 :: otp_fin2smn
Editor – about_omegat.txt
Tämän tekstin tarkoituksena on osoittaa, miten OmegaT toimii alkuperäiskansan kielellä.
Oovdânpyehtim heivee teemamateriaalân historjá máttáättâsân.
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.
Sämi Parlamentaarlâš Rääđi váátá avžuuttâsâsâstis 23.9.2011 tállán tábáhtuvvee olášuttemkiäldu OA olmoošvuoigâdvuođâräädist Njeellim siijdâ poccui pággunjuovâmij estimân.
<75/75/72%
/Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/ww
.samediggi.fi/index2.
php_option=com_content_task=view_id=519_lang=finnish.html.tmx>
2. Saamelainen Parlamentaarinen Neuvosto vaatii Nellimin pakkoteurastusten pysäyttämistä
Sämi Parlamentaarlâš Rääđi Váátá Njeellim pággunjuovâmij orostitten
<41/33/38%
/Users/cipriangerstenberger/otp_fin2smn/tm/fin2smn/admin/sd/www.
samediggi.fi/index2.
Glossary
Dictionary
Multiple Translations
Notes
Comments
Machine Translation
Project autosaved on 22:35
1/3 (1/3, 3)
147/147

Overview

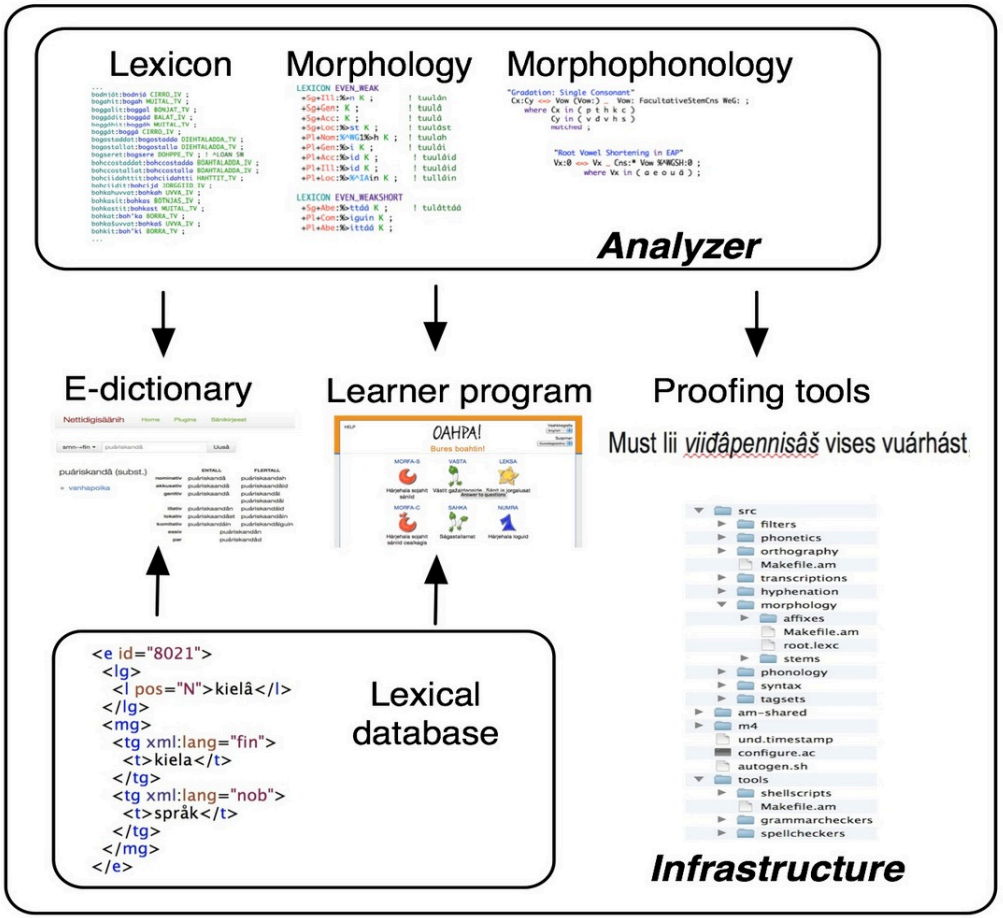
Aanaar Saami language community
the revitalisation of a small indigenous language

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

Data & software & experts



Linguist

Computational linguist

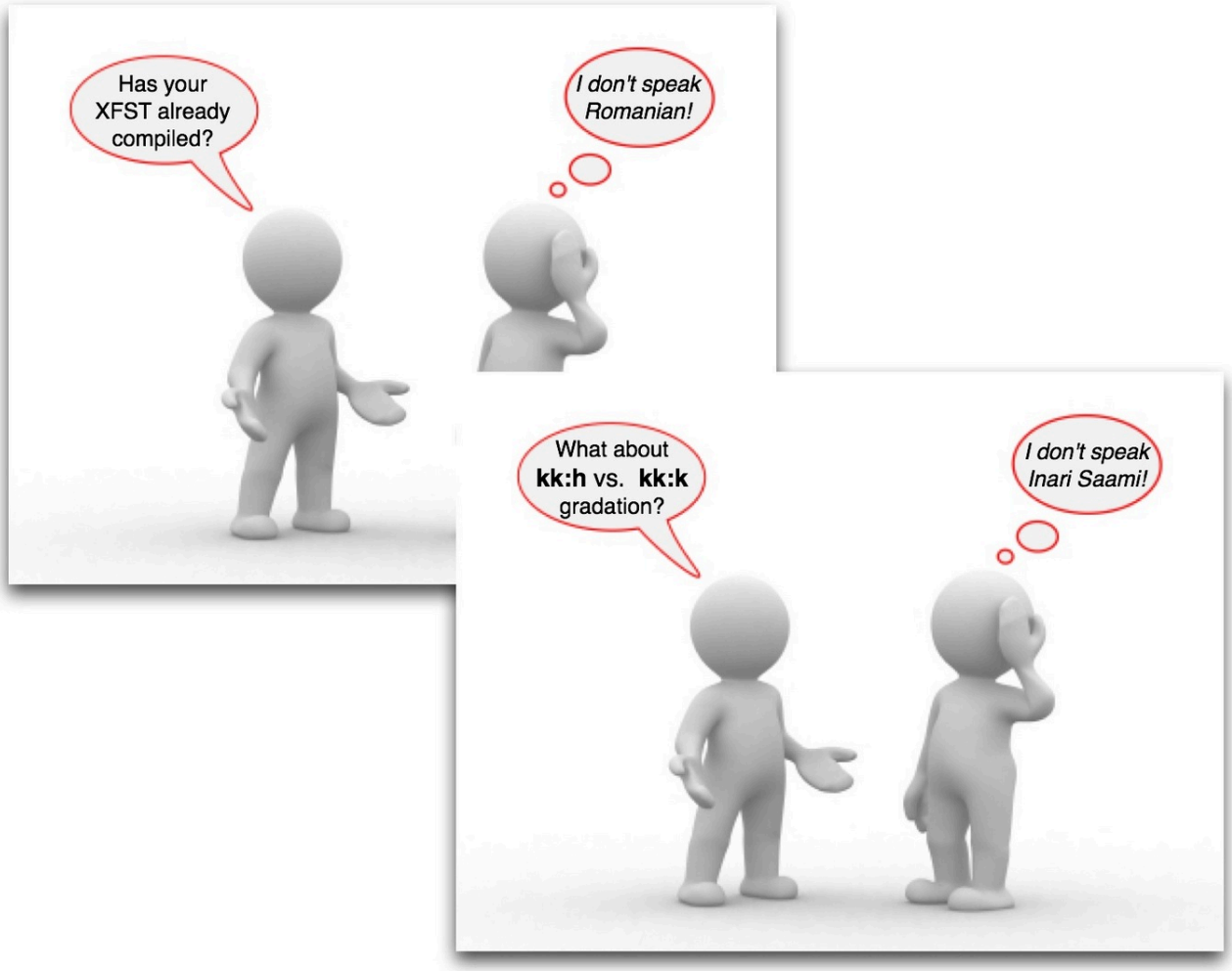
Programmer

Lexicographer

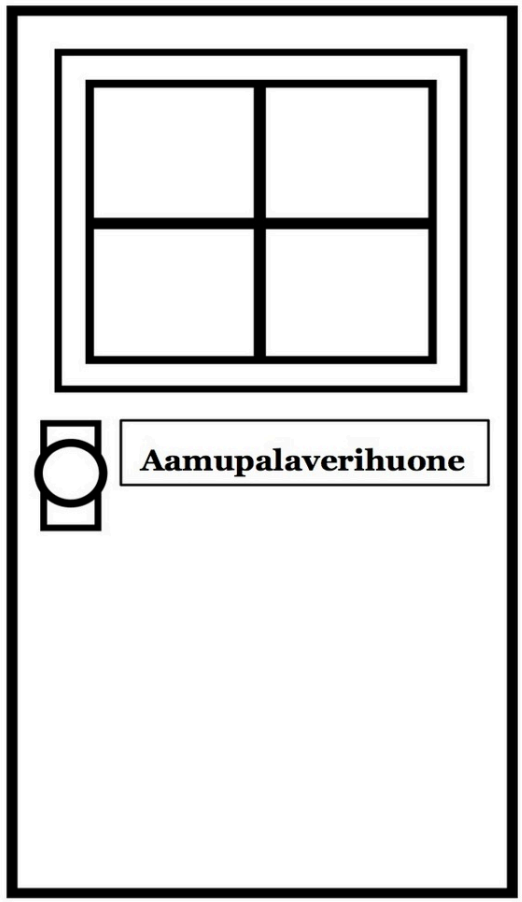
Teacher

.....

Do we understand each other?



I understand Finnish perfectly!



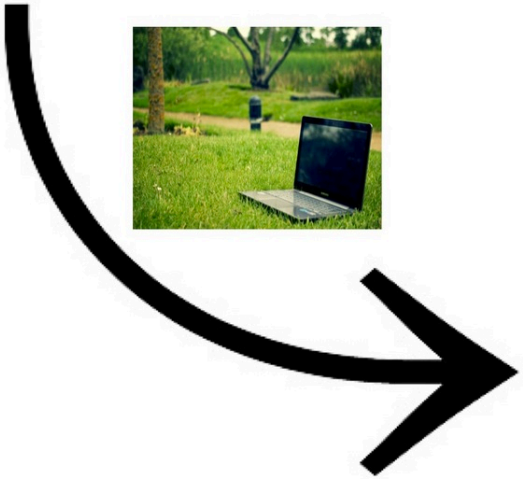
Machine-readability

aakkostaa *ornið puustavij mild* *[eng. "to alphabetise"; lit. "to organise alphabets accordingly"]*
ahdistaa *atâštið atâštâm aatâšt* *[eng. "to haunt"; lit. "to haunt", "I haunt", "(s)he haunts"]*

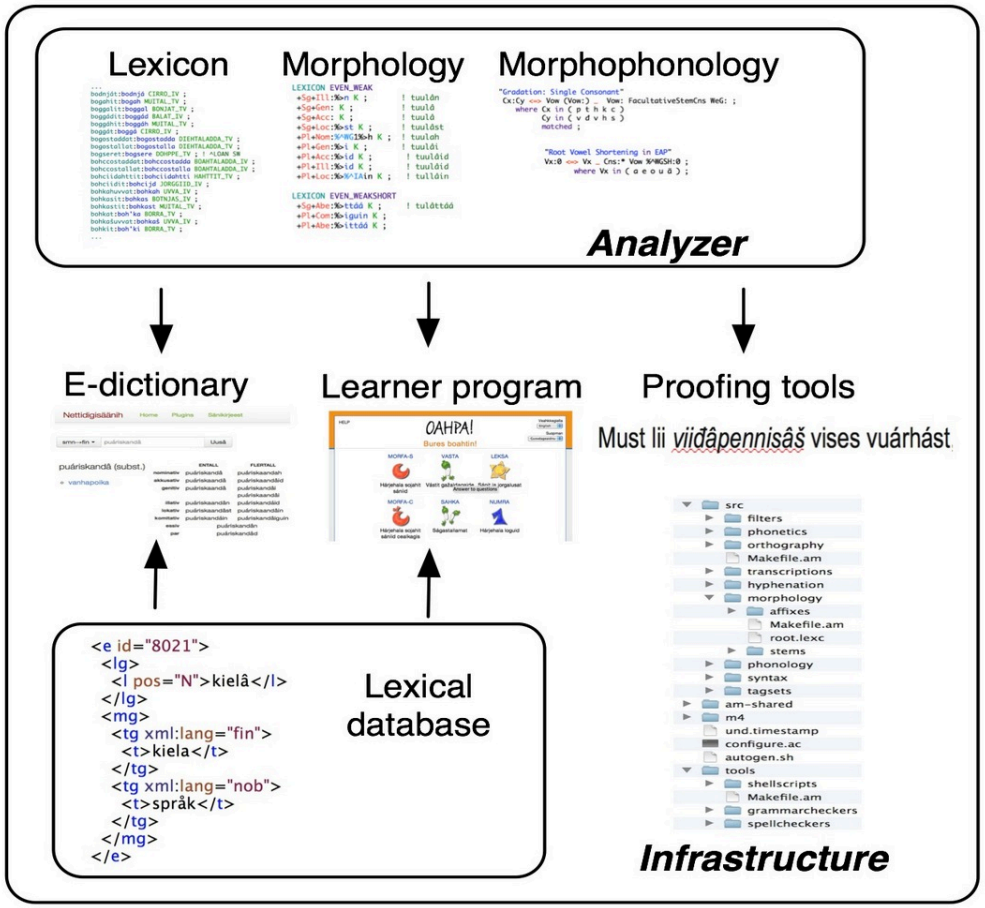
whitespace ambiguity → *disambiguate*
→ *label items appropriately*

```
<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">atâštið</t>
    </tg>
  </mg>
</e>
```

Teaching how language tools work



Human resources



Linguist

Computational linguist

Programmer

Lexicographer

Teacher

.....

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!