

# How to help languages to survive during modern time by means of language technologies?

---

Trond Trosterud

University of Tromsø



Jack Rueter

University of Helsinki





# Sámi giellatekno

Ruoktu

ОАҢРА!

Divvun

TechDoc

Dicts

Barents

Wiki

Search the site with googl

Search

► [Åarjelsaemien](#) ► [English](#) ► [Norsk](#) ► [Русский](#)

Last Published: 10/24/2012 04:29:29

## По-русски

▫ [Start page](#)

► [Фон](#)

► [Интерактивные программы](#)

► [Словари](#)

► [Тексты и списки](#)

► [Другие языки](#)

► [UnivOahpa-  
Vorkshop 2012](#)

## Добро пожаловать на страницы Giellatekno - саамские языковые технологии

[Davvi](#) [Åarjel](#) [Norsk](#) [English](#) [Suomeksi](#) [Русский](#)

## Ресурсы, которые существуют для разные языки

- **Саамские языки:** [северосаамский](#), [луле-саамский](#), [южносаамский](#) // [пите-саамский](#), [инари-саамский](#), [колтта-саамский](#), [кильдин-саамский](#).
- **Остальные уральские языки:** [вепсский](#), [горномарийский](#), [ижорский](#), [квенский](#), [КОМИ](#), [ливский](#), [ливвиковсий](#), [лугово-восточный марийский](#), [мокшанский](#), [нганасанский](#), [удмуртский](#), [финский](#), [хантыйский](#), [эрзянский](#).
- **Остальные языки:** [бурятский](#), [гренландский](#), [инупиак](#), [корнский](#), [оджибве](#), [русский](#), [фарерский](#),

# Uralic languages

URALISCHE SPRACHEN

## F Finnisch-Ugrisch

FO Ostseefinnisch

FO1 Finnisch

FO2 Karelisch

FO3 Wepsisch

FO4 Ischorisch

FO5 Estnisch

FO6 Wotisch

FO7 Liwisch

FS Samische Sprachen

FS1 Westsamisch

FS2 Zentralsamisch

FS3 Ostsamisch

FU Ugrisch

FU1 Ungarisch

FU2 Mansisch / Wogulisch

FU3 Chantisch / Ostjakisch

FP Finnisch-Permisch

FP1 Komi-Syrjänisch

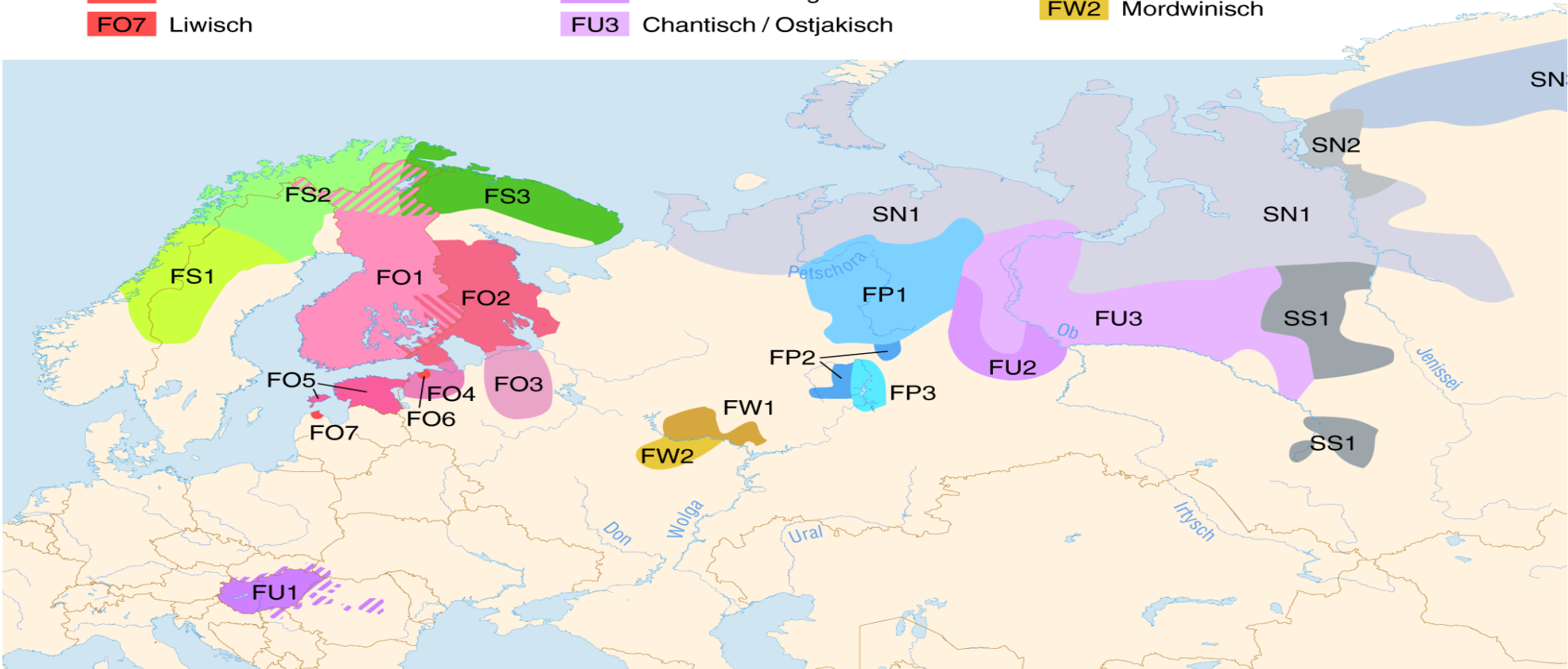
FP2 Komi-Permjakisch

FP3 Udmurtisch / Wodjakisch

FW Finnisch-Wolgaisch

FW1 Tscheremissisch

FW2 Mordwinisch



# Language policy in 1955

- Northern Norway
  - Forbidden to speak Saami in the boarding schools
- South Africa
  - The children got education in their mother tongue
  - ... but got no access to English
- Mari El
  - The children got 9 years education in their mother tongue
  - ... and they also learned Russian

# Norway since 1980

- A new language policy for
  - North Saami (20000)
  - Lule Saami (500)
  - South Saami (300)
- Right to use the mother tongue, and get official information in Saami
- Right to education in Saami



Radiologisk avd. venterom  
Radiologijja osd. vuordinlatnja

BRANNDØR  
holder lukket

← Røntgen

Foto: Jan Fredrik Frantzen



# RÁÐÐEHUS.NO

Dieđut ráđđehusas ja departemeanttain



Oza olles ráđđehusa.no:as

Oza

Čálánhápmi a a a

## Norgga ekonomii ja ovdána bures

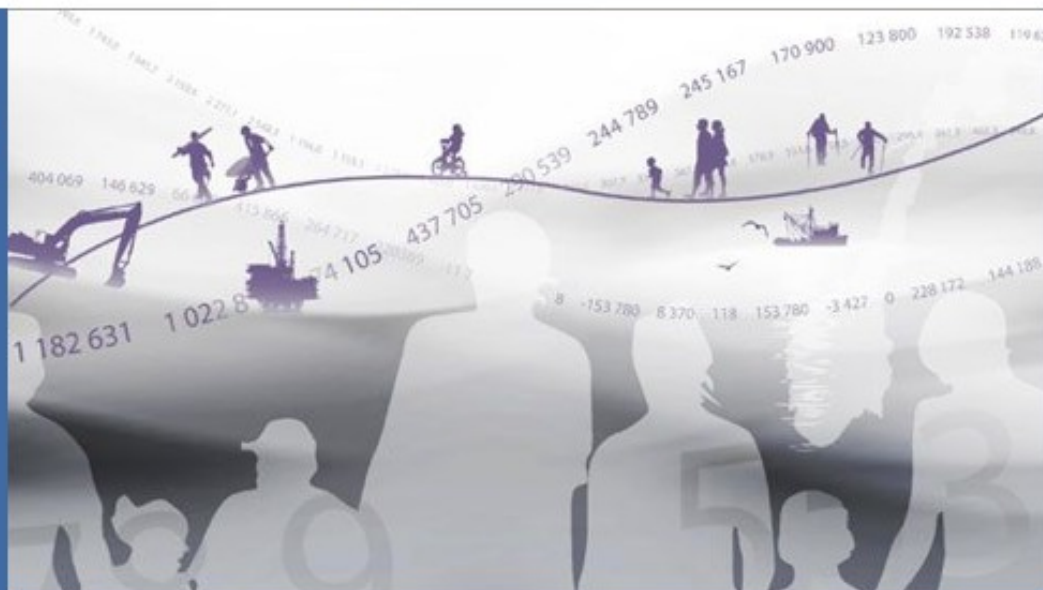
- Norgga ekonomii ja lea ceavzán bures ruhtadanroasu váttisvuodaid maŋŋá. Árvoráhkadeapmi lea dál lassánan ovcci jahkenjealjádasa maŋŋálagaid. Barggolašvuotta lea buoret go 2008 buoremus áiggi, ja bargguhisvuotta bissu vuollegaš dásis. Ráđđehusa reviderejuvvon bušehta árvalusain mii láchit dilálašvuodaid dasa, ahte norgga ekonomii ja ovdána bures ain viidáseappotge, ruhtadanministtar Sigbjørn Johnsen dadjala. **(Ruhtadandepartemeanta, 15.5.2012)**

Vearro- ja divatrivevdadusat >

Elektrovnnalaš vearrokoarttat >

EDAG álkidahtta ealáhusaid barggu >

Norga áigu dahkat loatnasoahpamuša IMF:in >

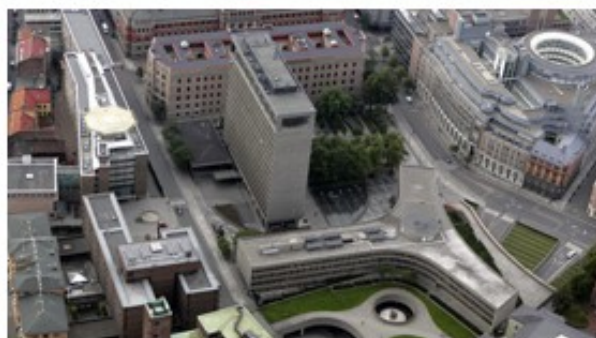


2012 dárkklistuvvon nátionálabušehta



### Galledii Sámedikki

Stáhtačálli Eli Blakstad galledii otna Sámediggeráđi Vibeke Larsena. Čoahkkima fáddán lei plánejuvvon elfápmollinnjá gaskal Ofuohta, Báhcavuona ja Hámmerfæstta. Čoahkkima galledii departemeanta



### Ráđđehuskvartála boahhteáigi

Ráđđehus lea mearridan čohkket departemeanttaid lahka ja birra dálá ráđđehuskvartála. Mo dálina áššiin galgá bargat viidáseappot, čielgá maŋŋil go ráđđehus lea árdáhan viiddis. Rabas proosaasas gos loahpuđet



### Ráđđehusa bargu

Ráđđehus-sátni geavahuvvo bealvválaččat guovtti áddejumis; olles ráđđehus čoahkis Gonagasa jodíheami vuolde stáhtaráđđechoahkkimis, ja olles ráđđehus čoahkis stáhtaministara jodíheami vuolde

Russia



# The language planning years

- New literary languages
  - with new letters!
- Mother tongue education
  - Books
  - National press
- New administrative borders





составитель карты Никола  
Карта не учитывает административные границы Российской империи на территориях до войны, поскольку территориальные границы между ними должны быть договорными.

ТОБОЛЬСК

Нижнетагильский

ПЕРМЬ

Екатеринбург

Верхнеуфалейск

Златоуст

Катав-Ивановский

УФА

ОРЕНБУРГ

УЛЕАБОРГ  
ВЕЛИКОЕ  
КНЯЖЕСТВО  
ФИНЛЯНДСКОЕ

НИКОЛАЙСТАД  
ВАЗАСКАЯ  
ГУБЕРНИЯ

КУОПИО

АРХАНГЕЛЬСК

ТАВАСТУС

ПЕТРОЗАВОДСК

ОЛОНЕЦКАЯ  
ГУБЕРНИЯ

НЮЛАНДСКАЯ  
ГУБ.

ВЫБОРГ

ТЕЛЬСИНГФОРС

Котлас

РЕВЕЛЬ

ПЕТРОГРАД

Нарва

Колпино

ВОЛОГДА

НОВГОРОД

ВЯТКА

ЭСТАЛАНДСКАЯ ГУБ.

Пернов

РИГА

ПСКОВ

Рыбинск

КОСТРОМА

Воткинский

Екатеринбург

Лифляндская  
Губерния

Люцин

ТВЕРЬ

ЯРОСЛАВЛЬ

Кинешма

Ижевский

Двинск

Полоцк

МОСКВА

Ковров

НИЖНИЙ  
НОВГОРОД

КАЗАНЬ

Катав-Ивановский

Вильно

ВИТЕБСК

СМОЛЕНСК

ВЛАДИМИР

Кинешма

Алатырь

Борисов

МИНСК

МОГИЛЕВ

КАЛУГА

РЯЗАНЬ

Симбирск

Барановичи

Пинск

Бобруйск

Гомель

Клинцы

Брянск

ТУЛА

Рязск

Сызрань

Пенза

Кузнецк

ОРЕЛ

ТАМBOB

ПЕНЗА

Сызрань

САМАРА

ОРЕЛ

ТАМBOB

ПЕНЗА

Кузнецк

САМАРА

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

ОРЕНБУРГ

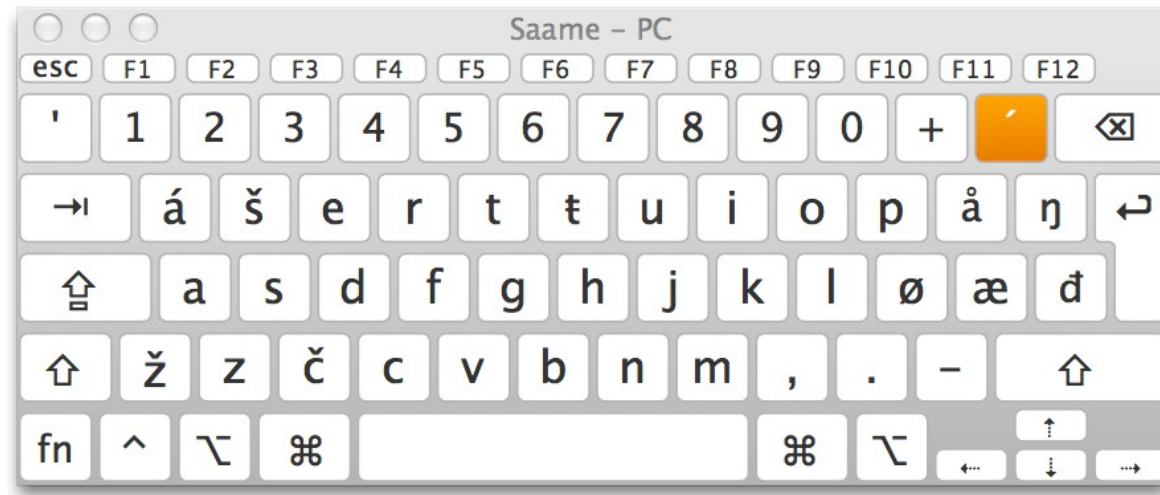
# Today: New challenges

- We move to the digital age of text processing
  - And we want our languages with us

# Programs needed

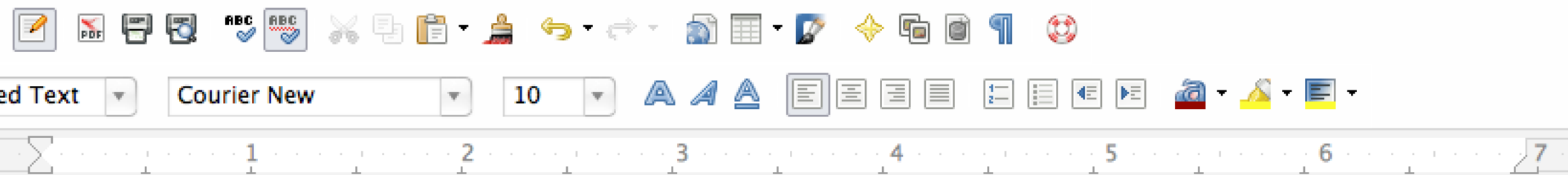
- Keyboard containing national letters
- Spell checkers
- Dictionaries who understand inflected forms
- Machine translation
- Information retrieval
- Sound: Synthetic speech and speech recognition

# Keyboards



- In the Komi Wikipedia, 66% of i, ö are Latin letters, only 33% are Cyrillic
- (2/3 write without a Komi keyboard

# Spell checkers



өшйыліс Латкинлөн сизимөд номера вугыръяс йылө. А нидзувсө яндысьтөма джыбөдісны.

Но мөдлаө Латкинлы эз мунсы. Шыр кыйөдысь кань моз куткырасис дебаркадер бөжын да кыйкнитавліс кыр йывлань. И быдөн пөсьмунлі, кор аддзис кык нылөс, горув посниа да тэрыба тувчалысьөс: өтиыс вөлі почта новлөдлысь, а мөдыс – Тамара.

Печора карысь кывтысь теплоходыс мыйлакө сөрмөдчис, эз на кыв ни тыдав, да нывъяслы, кылө, дышөдіс дебаркадер вылын өти- даың варгынысө, найө кутісны ветлөдлыны гөгөрыс. Водзджыксө кытшовтісны нырөдыс, а сэсся тотшсьөдісны и бөжланьыс, Латкин дінө.

- Абу тай дыш, – эльтана шыасис Тамара.
- А мый нө таянлы, горт олөм оз төдны-а, – вочавидзис почта новлөдлысь.
- Вола пестө поткөдлыны, – бергөдчис да шуис Ааткин, коді таөдз мышка на вөлі.
- Миян важөн нин чипасын.

# Dictionaries understand inflection

Báhčaveaijohka | Industriija | Okta johka – golbma stáhta | Geavvnis | Čálalaš meattábeassanlohporzi | Maŋemus mátkki | Dálá elfápmu



## BÁHČAVEAIJOHKA

Báhčaveaijohka lea odne rádjin Norgga ja Ruošša gaskkas. Johka vuolgá mearihis unnit jogážat mat leat vuodđun etnui mii golga Báhčaveaivutnii Girkonjárgga bol **jogas** ja Norggas leat guokte.

Vuosttaš Digisánit

*subst.* → **johka**

bekk, elv

**Analyse:** sg. lok.



# Machine translation 1

- What do they write about me in *Komi Mu*?



# Machine translation 1

- What do they write about me in *Komi Mu*?

# Machine translation 2

## Коми Республика официальной портал

[Русская версия](#) | [English version](#)

Каналан Сöвет


Веськöдлан котыр

Олөмö пöртысь власьт органьяс

### Выльторьяс

17.10.2012 17:39



«Коми Республика да Индия костын эмöсь ыджыд позянлуныяс экономика да культура юкөнныясын сöвмöдны öтвылысь удж öта-мöдлы пöльза вайöмөн» [ фото]

Юöр сетан технологиясын да промышленностьын öтувья балаяс збыльмöдан позянлуныяс йылысь, а сідзжö культура да велöдан да наука йитöдьяс йылысь сёрнитисны Сыктывкарын удж серти паныдасьлігөн Коми Республикаса Юралысьöс вежысь Александр Буров да Санкт-Петербургын Индияса генеральной консул господин Вишвас Сапкал.

Поиск

Гражданалөн шыöдчöмъяс

"Индöд-тшöктöмъяс..." журнал

Фотосерпаскуд

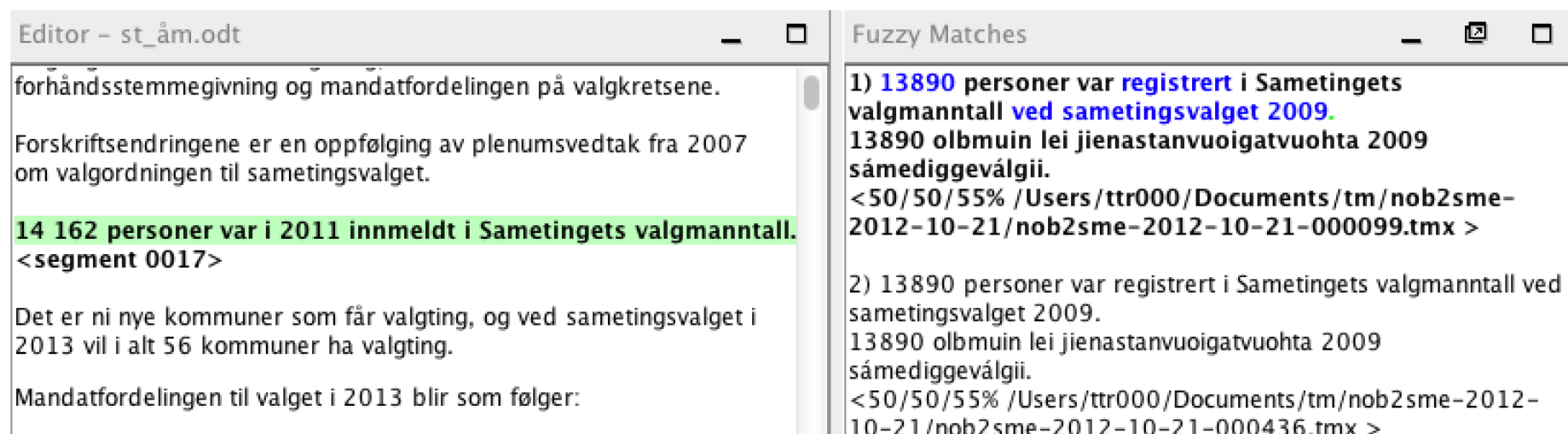
Воча йитöд

Сайт тэчас



ПРИЕМНАЯ ПРЕЗИДЕНТА

# Computer-assisted translation



- The program has access to a translation memory
- Sentences that have been translated earlier are offered as suggestions

# Scientific cooperation

- Partners
  - FU-lab in Syktyvkar
  - Giellatekno at the University of Tromsø
  - The Kone Foundation Language Programme 2012-2016
- Two-fold goal:
  - Make programs for efficient processing of Komi text
  - Make the scientific foundation for better language research on Komi and related languages

# FU-lab

- Focus on Komi, also Komi-Permyak
-

# Kone Foundation Language Program

- Documentation
  - Text collections, dictionaries
- Building analysers
  - вепсский, горномарийский, ижорский, коми, ливский, ливвиковский, лугово-восточный марийский, мокшанский, нганасанский, удмуртский, хантыйский, эрзянский.

# Giellatekno

- Focus on Saami languages
  - Also work on other northern languages
- Fields of interest:
  - Grammatical analysis
  - Interactive pedagogical programs

# Our common foundation

- Models of the grammar of Komi and other languages
- Without these models, none of the practical programs can be built



```
echo << defining the rules >>
```

```
! The famous L/V  
! -----
```

```
define LtoV [ л -> в || _ Flags* [ .#. | %> Cns ] ] ;
```

```
! These words are underlyingly л, and changed to в.
```

```
!€ ньл>ыс:ньлыс
```

```
!€ ньл:ньв
```

```
!€ ньл>сö:ньвсö
```

```
! Paragogic consonants  
! -----
```

```
! These are consonants that are visible in front of vowel-initial suffixes.
```

```
define mDeletion м -> 0 || н _ [ .#. | %> Cns ] ;
```

```
!€ синм>ыс:синмыс
```

```
!€ синм:син
```

```
!€ синм>сö:синсö
```

## LEXICON CASEPOSSLEX

DERIVEDADJ ; ! Deriving prenominal adjectives, республикаса  
CASEPOSSLEX-afterDerivation ;

## LEXICON CASEPOSSLEX-afterDerivation

+Nom: K ;  
!+ZeroAcc: K ;  
+Gen:%>лөн K ;  
+Abl:%>лысь K ;  
+Dat:%>лы K ;  
+Acc:%>öс K ;  
+Instr:%>өн K ;  
+Com:%>көд K ;  
+Car:%>төг K ;  
+Cns:%>ла K ;  
+Comp:%>ся K ;  
+Ine:%>ын SPAT-COMPARATIVE ;  
+Ela:%>ысь SPAT-COMPARATIVE ;  
+Ill:%>ö IllCompar ;  
+Apr:%>лань SPAT-COMPARATIVE ;  
+Egr:%>сянь SPAT-COMPARATIVE ;  
+Prl:%>öd K ;  
+Tra:%>ti K ;  
+Ter:%>ödз K ;  
+Apr+Ine:%>ланьын SPAT-COMPARATIVE ; ! New cases in last grammar  
+Apr+Ela:%>ланьысь SPAT-COMPARATIVE ; !  
+Apr+Ill:%>ланьö IllCompar ; !  
+Apr+Egr:%>ланьсянь SPAT-COMPARATIVE ; !  
+Apr+Prl:%>ланьöd K ; !  
+Apr+Tra:%>ланьti K ; !  
+Apr+Ter:%>ланьödз K ; ! Tähän asti

Коми Республика да Индия костын эмось ыджыд позянлуньяс экономика да культура юкөнъясын сөвмөдны өтвылысь удж өта-мөдлы польза вайомон.

Дать все формы слов

Устранить неоднозначность

Отправить форму

Сбросить форму

utf-8  latin

"<Коми>"

"КОМИ" N Sg Nom

"<Республика>"

"республика" N Sg Nom

"<да>"

"да" CC

"<Индия>"

"Индия" N Prop Plc Src/F Sg Nom

"<КОСТЫН>"

"КОСТЫН" Po Sg Ine

"КОСТ" N Sg Ine

"<ЭМӨСЬ>"

"ЭМ" A Pred Pl

"<ЫДЖЫД>"

"ыджыд" CmpTest

"<ПОЗЯНЛУНЬЯС>"

"позянлун" N Pl Nom

# Conclusion

- Without the tools presented here, a language cannot be used in digital contexts
- The administrative structures of Russia are based upon its ethnic composition, and language play a central role
- Scientific and applied work go hand in hand
- Cooperation and resource sharing are key factors