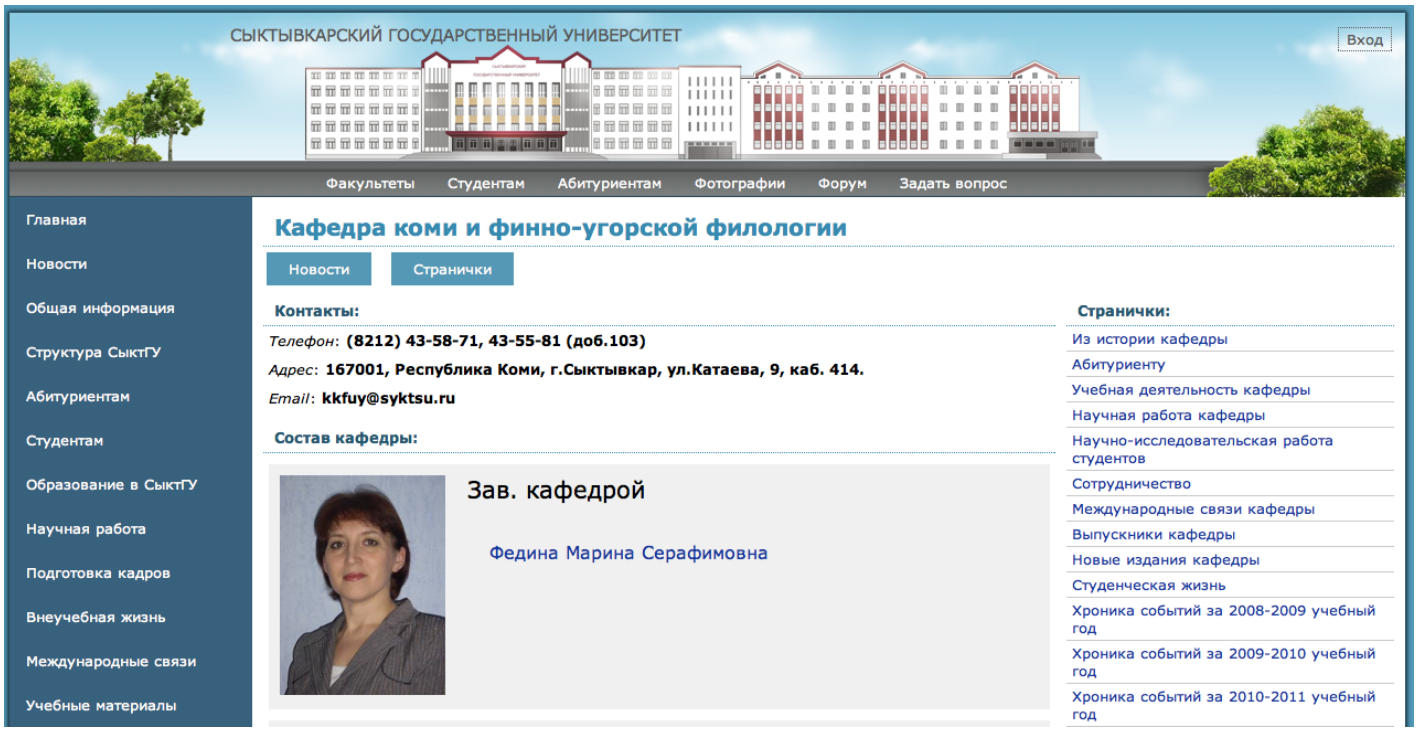
Роль языковой технологии в сохранении и ревитализации языка

Марина Федина и Трун Тростерўд

2 мая 2011 г.

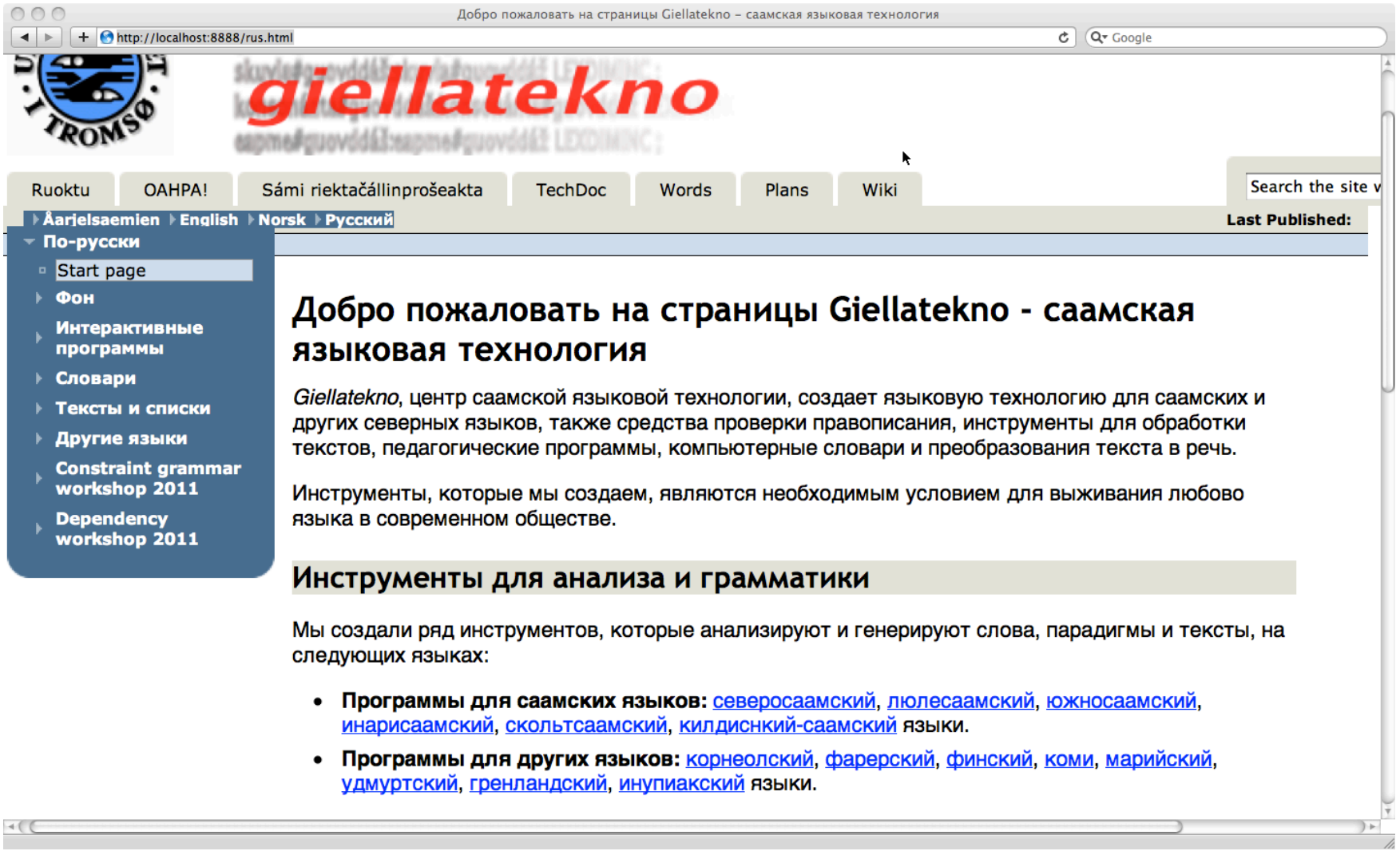
Сыктывкар - исследование коми языка

Исследование коми языка в Сыктывкаре.



../../img/logoWeb070sh

Лингвистическая технология в Тромсё



../../img/logoWeb070sh

Инструменты для анализа и грамматики

Мы создали ряд инструментов, которые анализируют и генерируют слова, парадигмы и тексты, на следующих языках:

- **Программы для саамских языков:** северосаамский, люлесаамский, южносаамский, инарисаамский, скольтсаамский, килдиснкий-саамский языки.
- **Программы для других языков:** корнеолский, фарерский, финский, коми, марийский, удмуртский, гренландский, инупиакский языки.

Что такое языковая технология и почему она важна?

Технология - это ремесло создания инструментов.

Языковая технология - это ремесло создания инструментов, которые необходимы для использования языка.

- Некоторые языковые технологии независимы от самого языка,
- другие наоборот, тесно связаны с конкретно определенными языками.

Список языковых технологий

- Основные инструменты
  - клавиатура;
  - морфологический анализатор и синтаксический анализатор.
- С их помощью можно создать множество важных ресурсов:
  - системы проверки правописания (проверка орфографии);
  - электронные словари;
  - машинный перевод на русский язык и с русского на финно-угорские языки, и между ф-у яз.

Клавиатура

Философия для клавиатур

Для профессионального использования:
«Национальная клавиатура»: i и ö как отдельные буквы (ц, щ = AltGr + i, ö), или в цифровом ряде

Для других пользователей: «Русская клавиатура»: i = AltGr + и, ö = AltGr + о

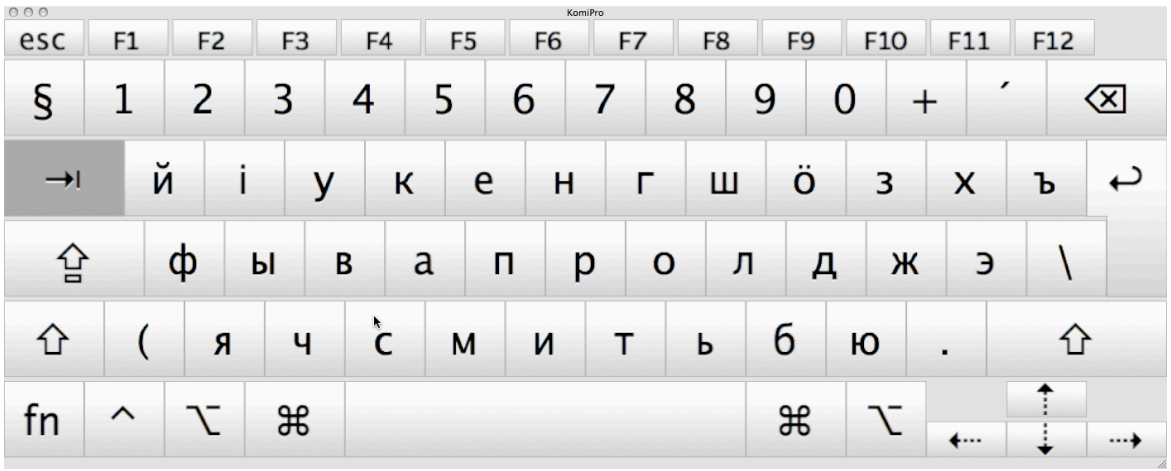
Для иностранцев: Клавиатура яшерты и уральские буквы под AltGr

Без отдельной клавиатуры: Insert Symbol в Microsoft Word :-(

Статистика букв в уральских языках

	бкв	пор	%	бкв	пор	%
Коми-перм. яз.	ӧ	3	6,92	ъ	34	0,03
	і	22	1,32	щ	35	0,02
Коми яз.	ӧ	4	6,38	щ	34	0,02
	і	27	0,81	ц	33	0,16
Лугово-марий. яз.	ӱ	22	1,12	щ	35	0,02
	ӧ	24	0,61	ъ	34	0,02
	ҥ	26	0,52	ф	33	0,09
Горно-марий. яз.	ӹ	2	6,84	щ	35	0,02
	ӓ	11	4,37	ю	34	0,07
	ӱ	31	0,29	ф	32	0,27
	ӧ	33	0,16	б	30	0,35

Клавиатура коми языка



Языки, в которых много национальных букв

Удмуртский язык:
25 й, 30 ӧ, 34 ӟ, 36 ч, 37 ӝ
Кильдинский язык:

Долгие: ā, ē, ӣ, ӯ, ю̄, я̄
Точки: ӓ, ӟ, ё̄
Растр.: ӈ́, ԯ, ӎ, ӊ, ҏ
Другие: ԧ, ј, ӈ, ҭ,

Проект клавиатура

The Uralic keyboard project

The goal of this project is to contribute to the creation of good and efficient keyboards for minority languages, especially in Russia.

Background documents

- The application text (somewhere)
- Layout discussion~ О Дезигн

Keyboard layouts

tbw.

Letter frequencies

Letter frequencies in running text for:

Meadow Mari ~ Hill Mari ~ Komi Permyak ~ Komi Zyrian ~ Udmurt ~ Kildin Sami

For reference, consider also the Russian letter frequency.

These data are based upon the respective Wikipedias, except the Kildin ones, who are based upon the lemma list of Kuruch's dictionary.

Layout images

- Pictures of layout drafts may be found here

Языковая технология в уральских языках

Структура коми словаря

```
<entry>
   <lemma>аддзӧдчыны</lemma>
   <stem>аддзӧдчы</stem>
   <contlex>Verb3</contlex>
   <pos>V</pos>
   <article>
      <sem>RECIP</sem>
      <syn>SVP/-кӧд</syn>
      <eng>
         <choice>
            <variant>meet</variant>
         </choice>
      </eng>
      <fin>
         <choice>
            <variant>tavata</variant>
         </choice>
      </fin>
   </article>
</entry>
```

Структура коми словаря

```
<entry>
   <lemma>аддзӧдчыны</lemma>
   <stem>аддзӧдчы</stem>
   <contlex>Verb3</contlex>
   <pos>V</pos>
   <article>
      <sem>RECIP</sem>
      <syn>SVP/-кӧд</syn>
      <eng>
         <choice>
            <variant>meet</variant>
         </choice>
      </eng>
      <fin>
         <choice>
            <variant>tavata</variant>
         </choice>
      </fin>
   </article>
</entry>
```

lexc-файл для коми, созданный компьютером

```
ярмыны:ярмы Verb3  "fly into a fury" ;
явитны:явит Verb2  "spread" ;
являйтчыны:являйтчы Verb3  "be" ;
абутӧмасьны:абутӧмась Verb2  "feign poverty" ;
адавны:адал Verb1  "gobble up" ;
аддзавны:аддзал Verb1  "find" ;
аддзӧдчыны:аддзӧдчы Verb3  "meet" ;
аддзӧдлыны:аддзӧдлы Verb3  "see" ;
аддзыны:аддзы Verb3  "see" ;
аддзывны:аддзыл Verb3  "experience" ;
аддзысьны:аддзысь Verb2  "be found" ;
аддзысьлыны:аддзысьлы Verb3  "see one another" ;
```

Другой файл: Спряжение коми

gtsvn/kt/kom/src/verb-kom-morph.txt

72:0 Vdz4i3k

```
73
74  LEXICON Verb1 ! яйсявны,  в - л change
75  +V:  Finiteforms ;
76  +V+Inf:%>ны K ;
77
78  LEXICON Verb2 ! янӧдны, контролируйтны. Ending in -дны, -тны.
79  Verb1 ;
80  +V+Inf:%>ны K ;
81
82  LEXICON Verb3 ! ярмыны. Ending in -ыны.
83  Verb1 ;
84  +V+Inf:%>ыны K ;
```

134:23 Finiteforms

```
126
127 LEXICON Finiteforms !Gives linking vowels for 3 tenses
128 +Ind+Fut:%>а PresPret1 ; ! 1,2  Future
129 +Ind+Prs:%>а PresPret1 ; ! 1,2  Present
130 +Ind+Prt1:%>і PresPret1 ; ! 1,2  Preterite1. і и variation?
131 +Ind+Fut:%>ас PresPret1-SgPl3 ; ! 3 PresPret1-SgPl3
132 +Ind+Prs:%>ӧ PresPret1-SgPl3 ; ! 3
133 +Ind+Prt1:%>і P3 ; ! 3
134 +Ind+Prt2:%>ӧм Pret2a ;
135
136 Non-finiteforms ;
137
138 LEXICON Non-finiteforms
```

Результат: Анализатор коми языка

../../img/logoWeb070sh

Результат: Анализатор коми языка

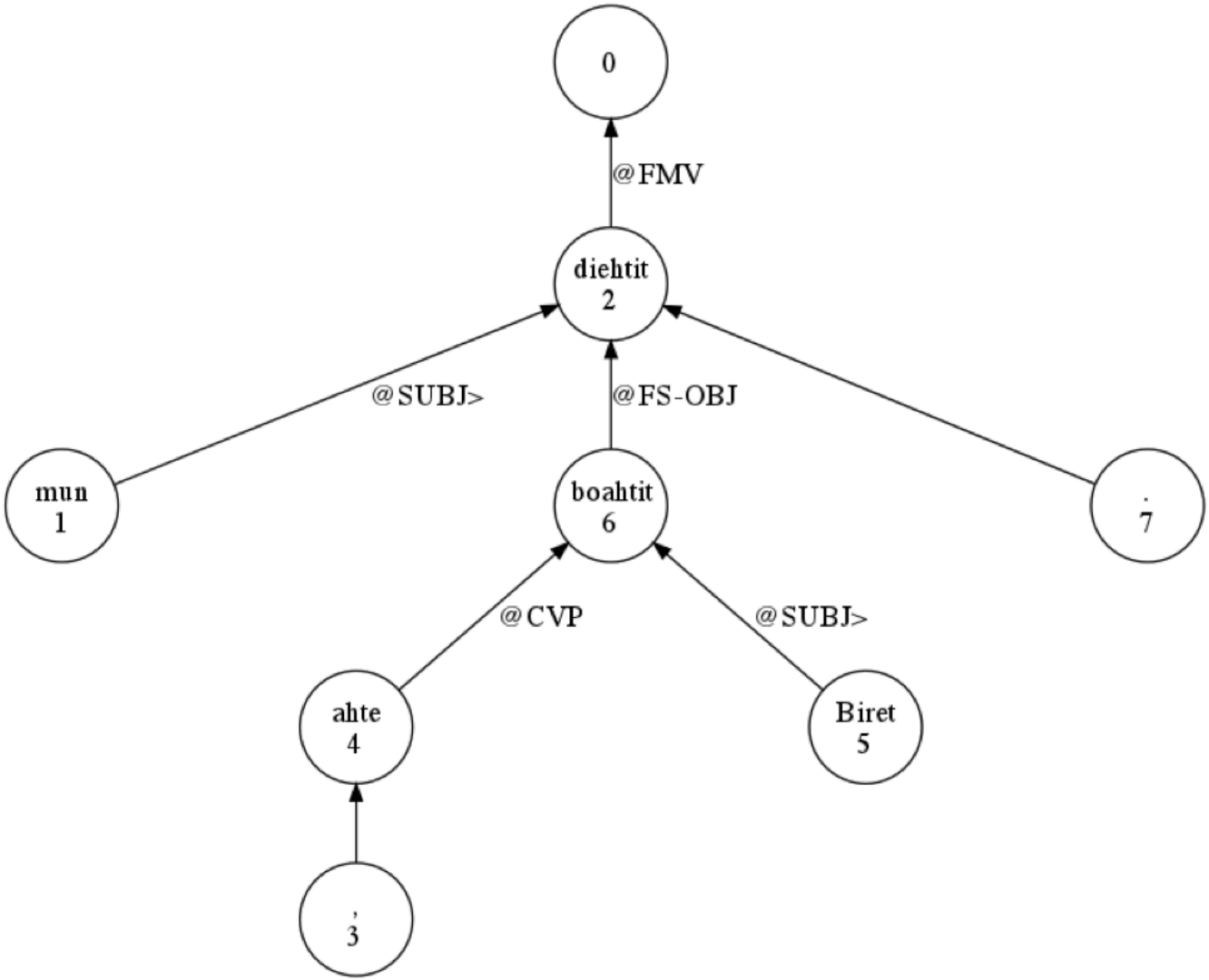
```
"<Тайӧ>"
    "тайӧ" Pron Dem
"<субдоменсӧ>"
    "субдоменсӧ" ?
"<видзӧны>"
    "видзны" V Ind Prs Pl3
"<,>"
    "," CLB
"<медым>"
    "медым" CC
"<вӧчны>"
    "вӧчны" V Inf
"<Википедия>"
    "Википедия" ?
```

```
"<пытшкын>"
    "пытшкын" ?
"<Коми>"
    "коми" N Sg Nom
    "коми" N Sg Acc
"<кыв>"
    "кыв" N Sg Nom
    "кыв" N Sg Acc
"<вылын>"
    "вылын" Adv
    "выв" N Sg Ine
"<гижӧдъяс>"
    "гижӧд" N Pl Nom
    "гижӧд" N Pl Acc
```

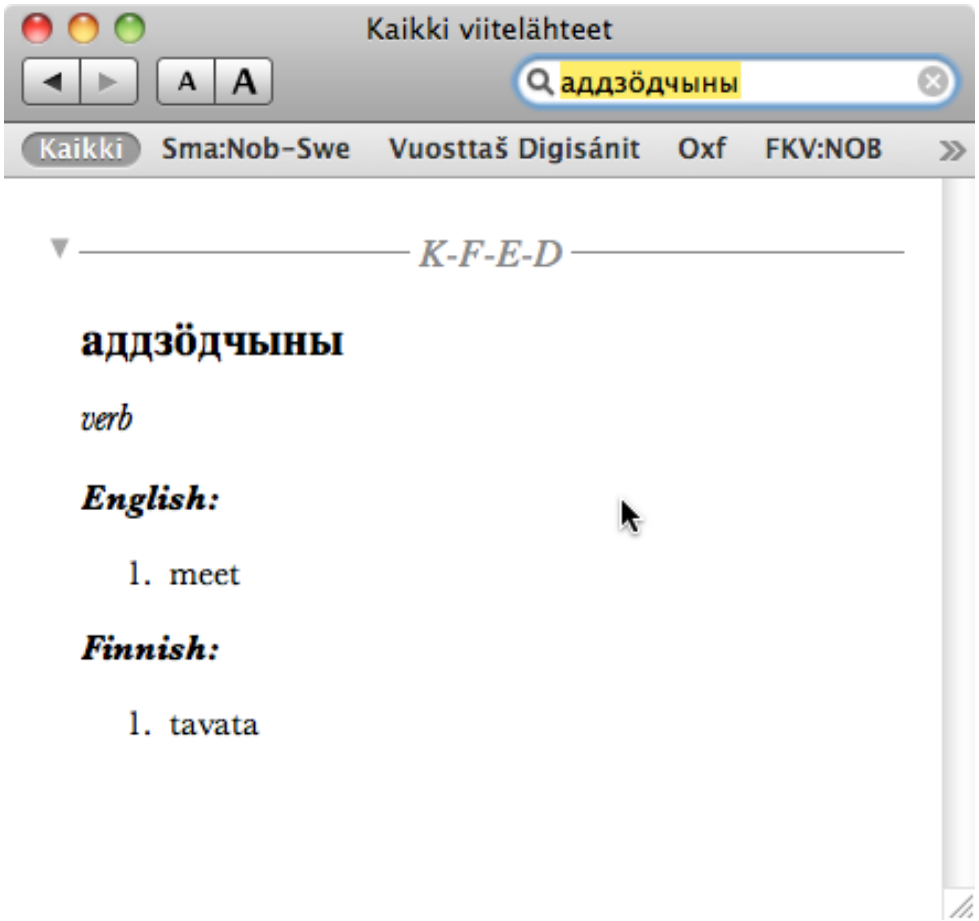
(Анализатор саамского языка)

```
"<Mun>"
    "mun" Pron Pers Sg1 Nom @SUBJ> #1->2
"<dieđán>"
    "diehtit" <mv> V TV Ind Prs Sg1 @FMV #2->0
"<,>"
    "," CLB #3->4
"<ahte>"
    "ahte" CS @CVP #4->6
"<Biret>"
    "Biret" N Prop Fem Sg Nom @SUBJ> #5->6
"<boahtá>"
    "boahtit" <mv> V IV Ind Prs Sg3 @FS-OBJ #6->2
"<.>"
    "." CLB #7->2
```

(Анализатор саамского языка)

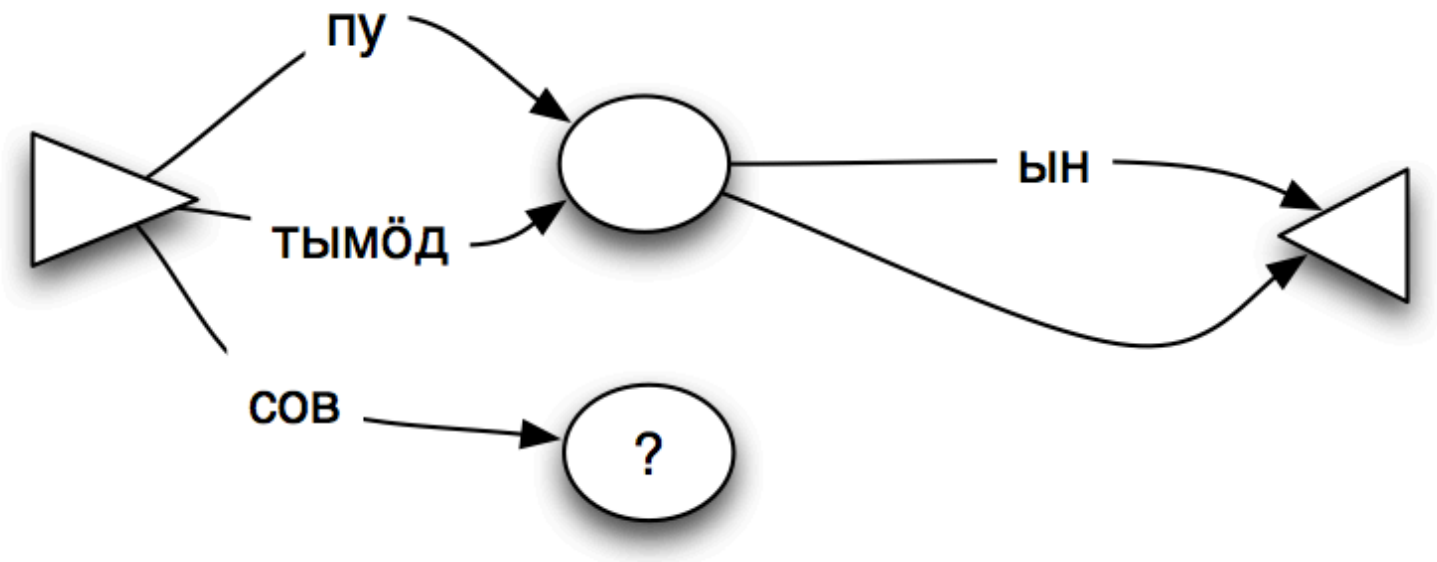


Результат: Словарь коми языка

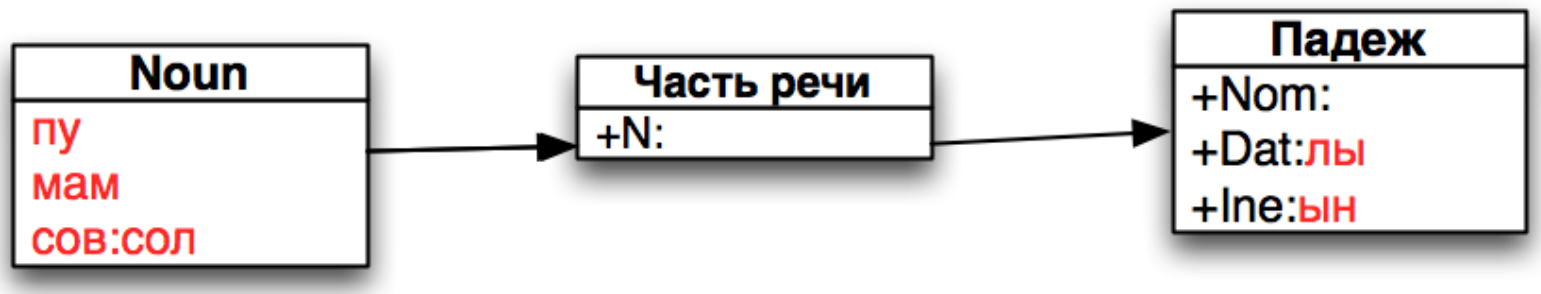


Фонология и морфонология

Склонение как автомат



Автомат в программе lexc

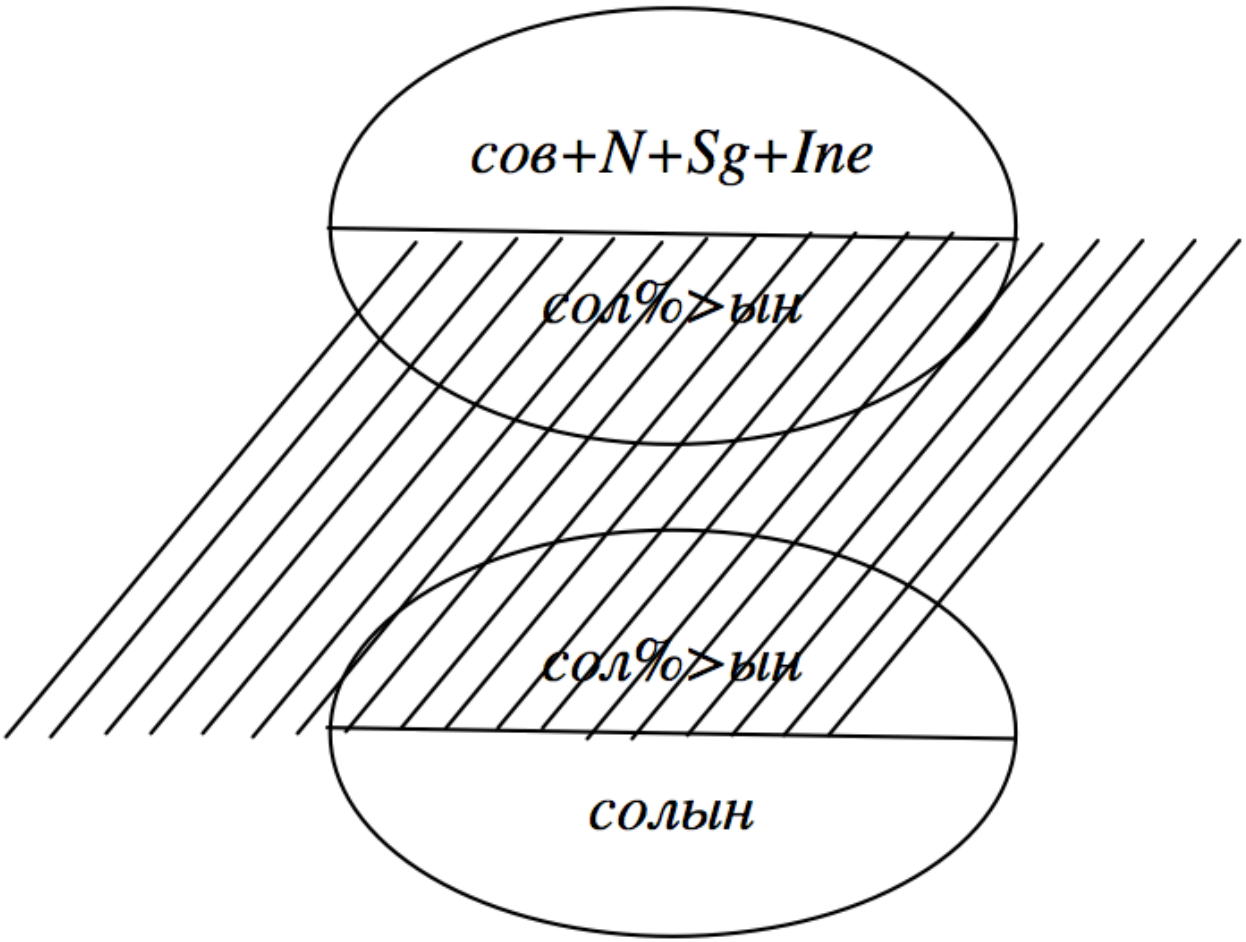


Автомат в программе xfst

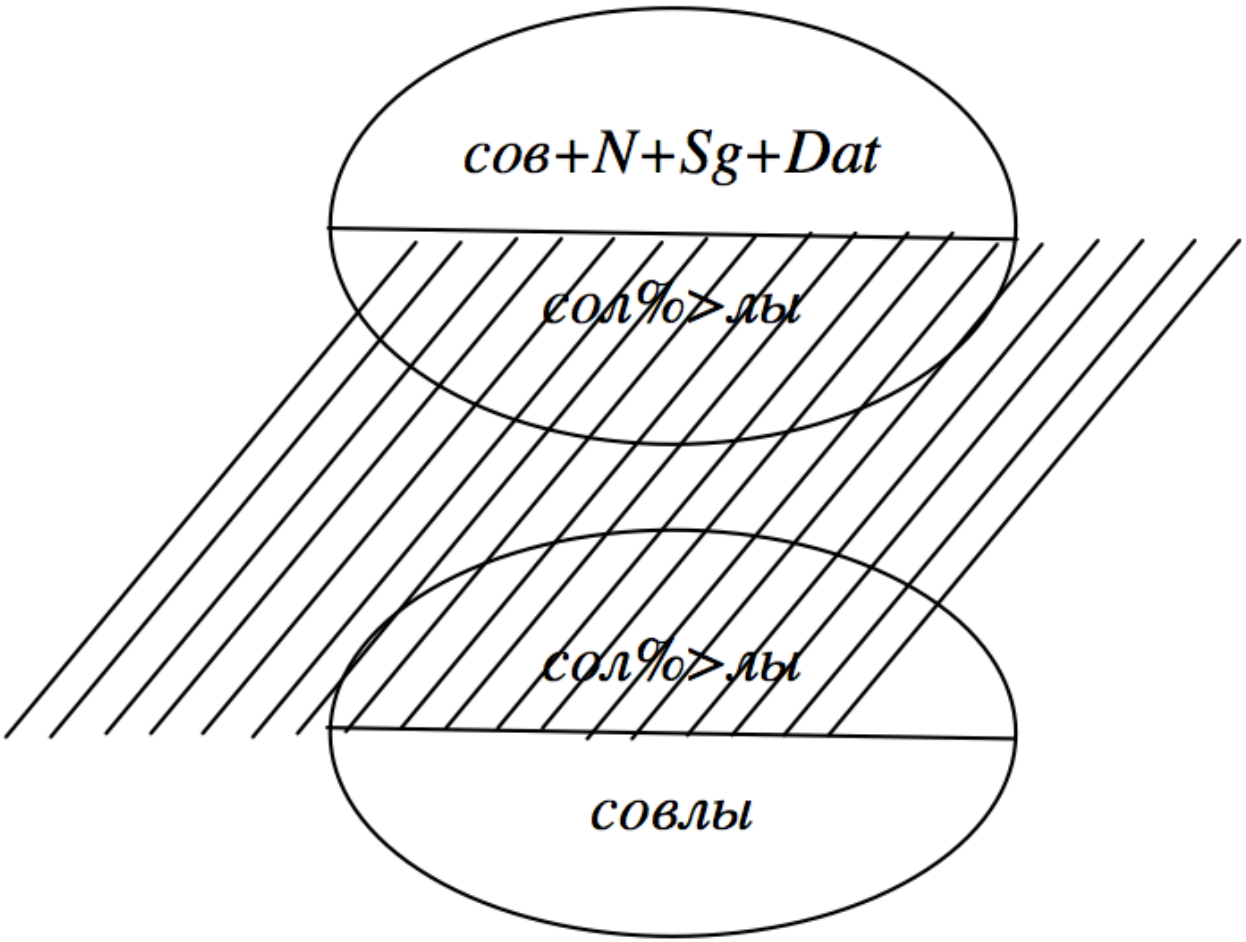
```
[ л -> в || _ [ .#. | %> Cns ] ] ;
```

Л переходит в В в абсолютном конце слова или перед согласным

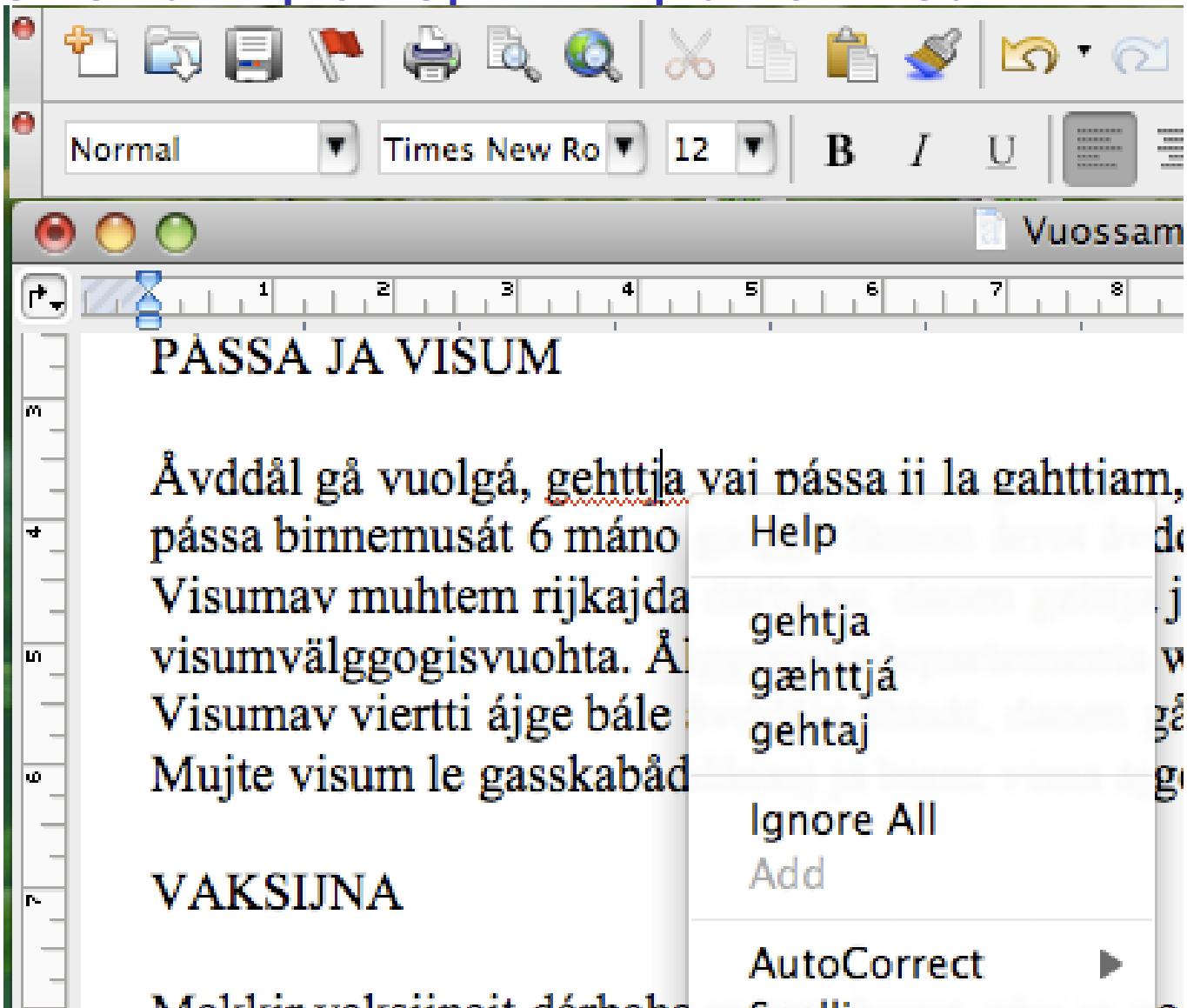
Местный падеж в двух автоматах: морфология и фонология



Дательный падеж в двух автоматах: морфология и фонология

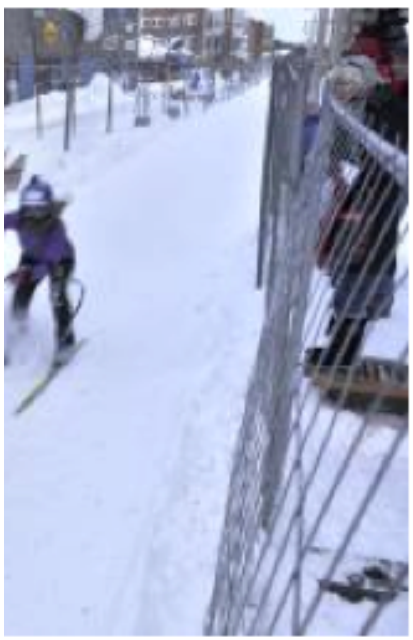


Система проверки правописания



Электронные словари

Электронные словари



ımunenis šattai garra gilvu
agodii Kuhmunen fiškalit
ɔvven buot govaid: Máret Eli

Son lea guhká niegadan ahte galggai oažžut guokte hearggi loahppavurrui, ja mannan vahkkoloahpa ollašuvai ge su niehku.

Mannan vahkkoloahpa ledje 18 hearggi boahtán Álttá guovddášgáhttii gilvalit 201 mehtera sprinttas, 16 hearggi senioroasis, ja guokte hearggi junioroasis. Son gii beasai illudit eanemusat maŋŋe
Risten Sara.

beassat (verb)

1. fritas;
2. slippe (inn), komme (til);
3. få (anledning til)

Analyse: indik. pret. 3p. sg.

Vuosttaš Digisánit | Mer...

Guvttiin herggiin

ggit, namalassii beakkán Dolgi ja su ođđa hear


Vuosttaš Digisánit
beassat
Alt
Wikipedia
Lule Saami-X
Vuosttaš Digisánit
VD_2011
beassat
verb
1. fritas
2. slippe (inn), komme (til)
3. få (anledning til)
Analyser: inf. el. indik. pres. 1p. pl.
Nøkkelformer:
indik. pres. 1p. sg. (odne mun) beasan
indik. pret. 1p. sg. (ikte mun) bessen
indik. pres. neg. (in) beasa
Eksempler:
Son beasai militearabálvalusas.
Han ble fritatt fra militærtjenesten.
In beassan lahka ge uvssa.
Jeg kom meg ikke i nærheten av døra engang.
Ja dál beassá bárdni sihkastit bivastaga ja bosihastit.
Og nå får gutten tørke svetten og puste ut.

Машинный перевод

- перевод, чтобы понимать текст
- перевод, чтобы произвести текст

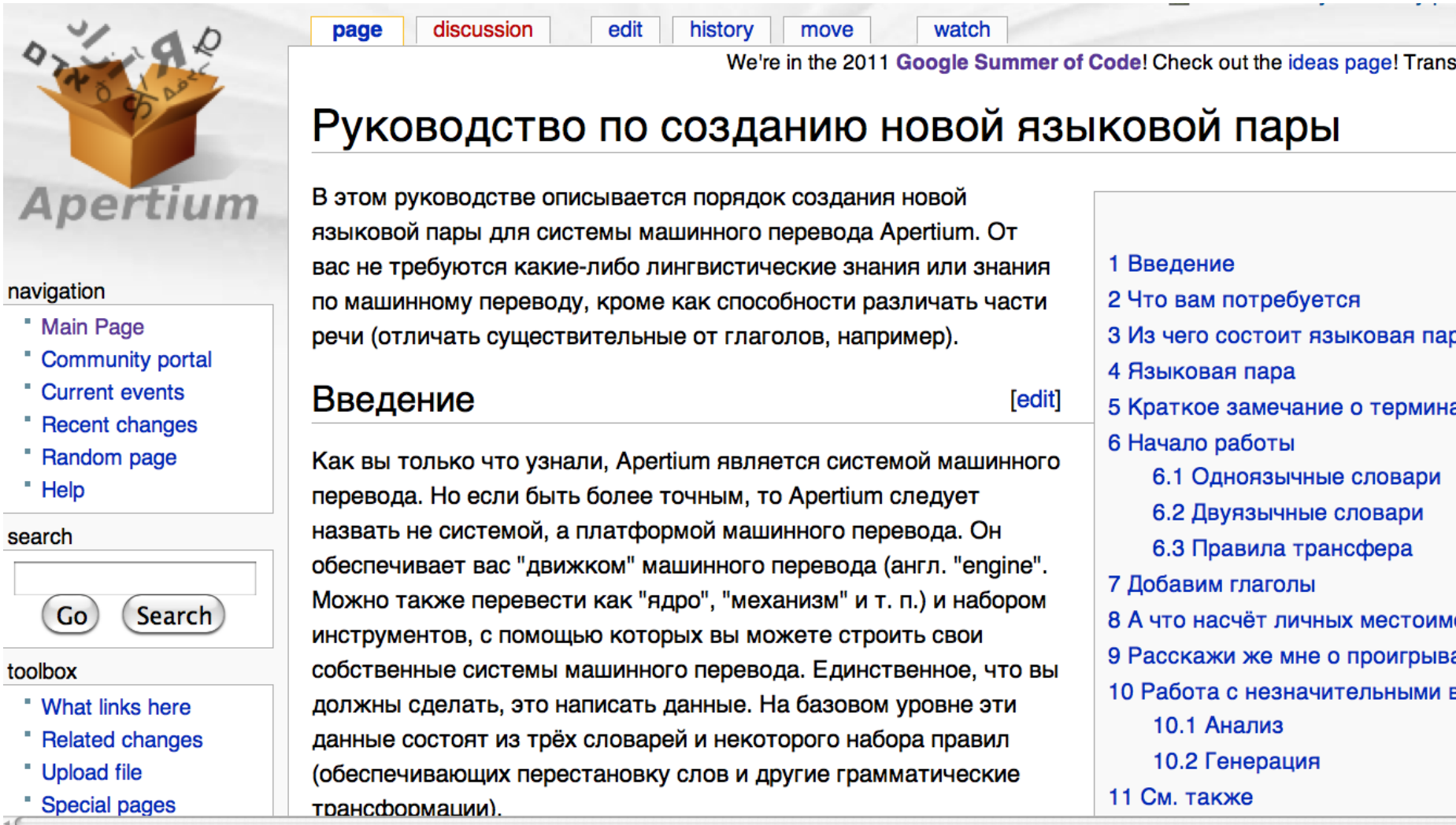
Перевод, чтобы понимать текст

Перевод с языка национальных меньшинств на русский язык

Перевод, чтобы произвести текст

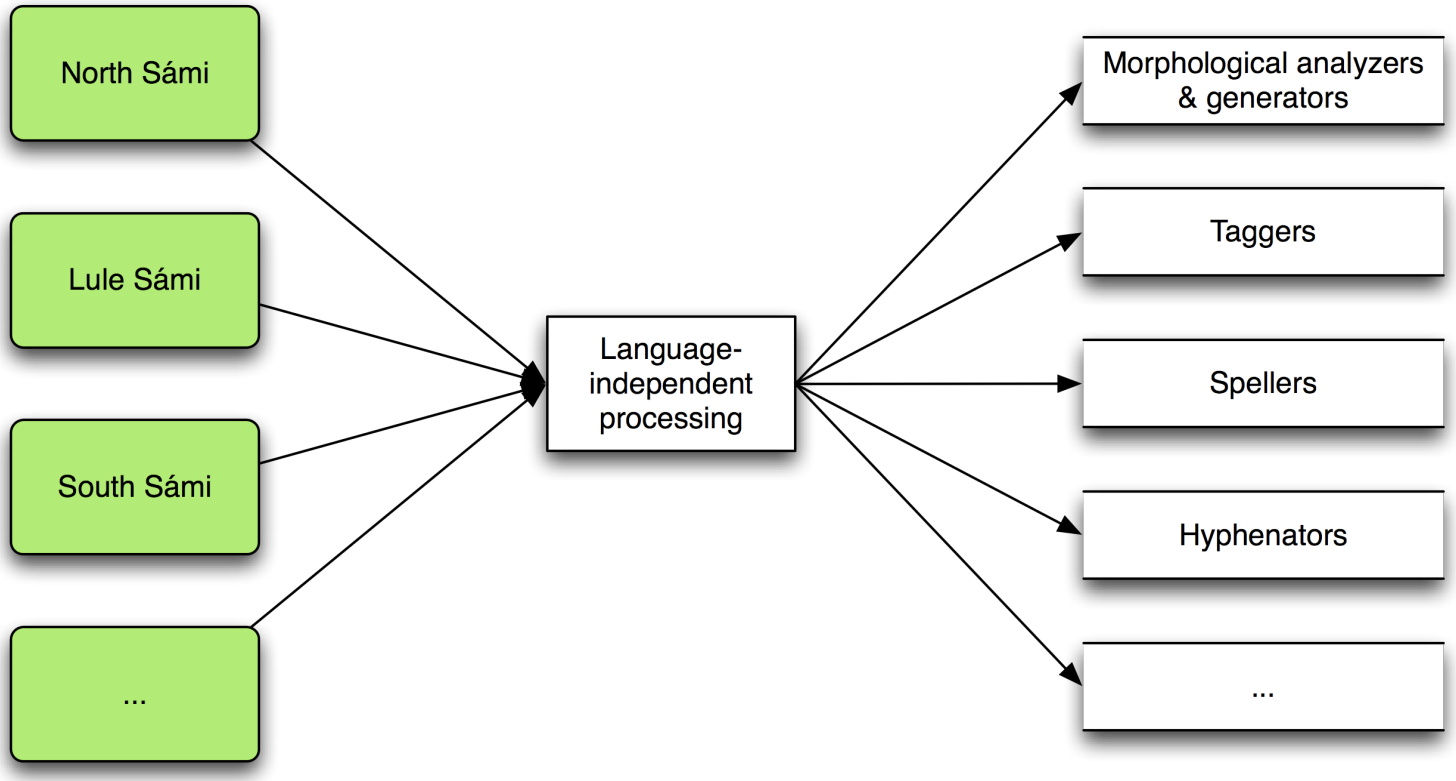
Например: Перевод с финского языка
Тоже: Перевод между языками меньшинств в России

Открытая платформа по машинному переводу

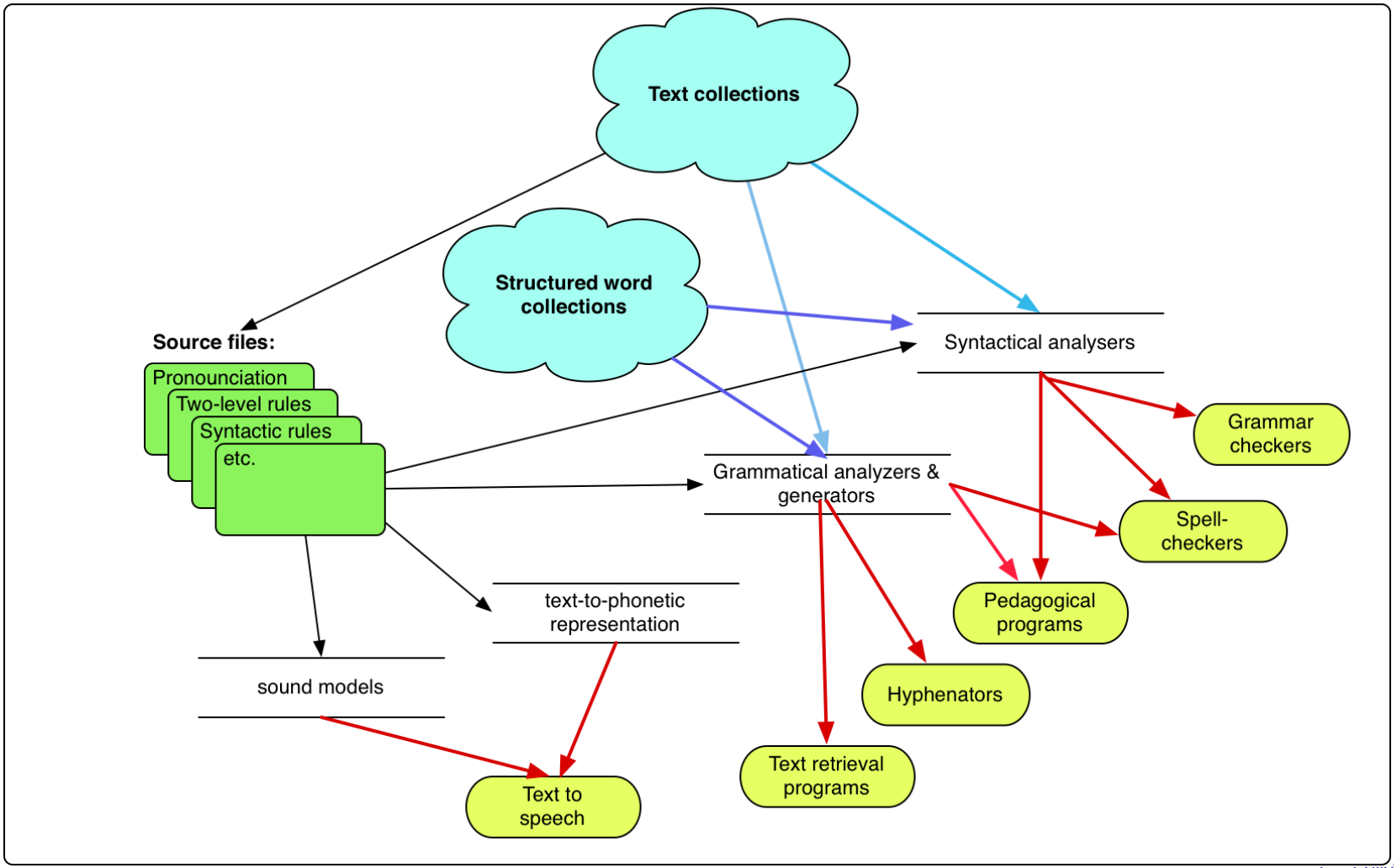


Техническая инфраструктура

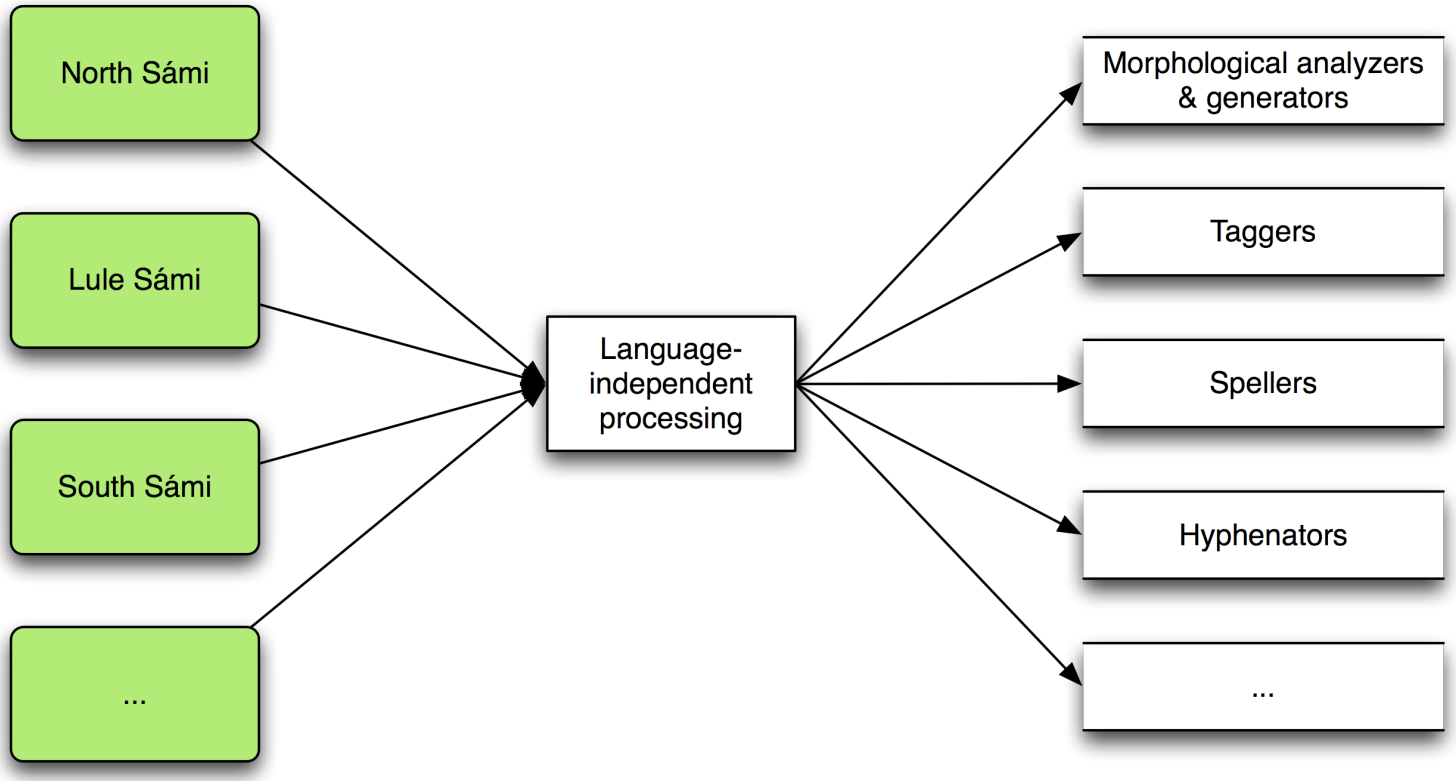
Панорама


North Sámi
Lule Sámi
South Sámi
...
Language-independent processing
Morphological analyzers & generators
Taggers
Spellers
Hyphenators
...

Панорама


Text collections
Structured word collections
Source files:
Pronounciation
Two-level rules
Syntactic rules
etc.
Syntactical analysers
Grammar checkers
Grammatical analyzers & generators
Spell-checkers
text-to-phonetic representation
Pedagogical programs
sound models
Hyphenators
Text retrieval programs
Text to speech

Панорама


North Sámi
Lule Sámi
South Sámi
...
Language-independent processing
Morphological analyzers & generators
Taggers
Spellers
Hyphenators
...

Инфраструктура для всех

- Политика открытого кода
  - Храните свои файлы на наших компьютерах согласно свободным лицензиям
  - Добавьте lexica, и, вместе с нами, напишите грамматические законы
  - Результат будет преобразован в различные практические программы

Предложение всем Вам

- Используйте свободную инфраструктуру (например, нашу)
- Храните свои файлы на наших компьютерах согласно свободным лицензиям
- Добавьте лексику и вместе с нами напишите грамматические законы
- Результат будет преобразован в различные практические программы

Некоторые из языков Apertium

- Азербайджанский, Чувашский, Баскский, Казахский, Турецкий, Татарский,
- Белорусский, Болгарский, Церковнославянский, Литовский, Латышский, Польский, Русский, Словацкий, Словенский, Украинский,
- Чеченский, Баскский, Ингушский, Курдский, Армянский, Непальский, Осетинский,
- Эстонский, Финский, Венгерский, Коми-зырянский, Кильдинский,

Заключение

- Разделение наших ресурсов.
- Доступность ресурсов, сохранение их во взаимозаменяемом формате.
- Открытость для совместной работы, чтобы могли участвовать люди из различных мест.

Тау!