

Роль языковой технологии в сохранении и ревитализации языка

Марина Федина и Трун Тростерүд

2 мая 2011 г.

Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

Электронные словари

Машинный перевод

Техническая инфраструктура

Заключение

Сыктывкар – исследование коми языка

Исследование коми языка в Сыктывкаре.

The screenshot shows the website for the Department of Finnish-Uralic Philology at Syktyvkar State University. The header features a banner with the university's name and a building image, and a navigation menu with links for Faculties, Students, Applicants, Photos, Forum, and Ask a Question. A 'Вход' (Login) button is in the top right. The main content area is titled 'Кафедра коми и финно-угорской филологии' and includes tabs for 'Новости' (News) and 'Странички' (Pages). The 'Контакты' (Contacts) section provides phone, address, and email information. The 'Состав кафедры' (Department Staff) section features a portrait of Marina Serafimovna Fedina, the department head. The 'Странички' section lists various pages related to the department's history and activities.

СЫКТЫВКАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Вход

Факультеты Студентам Абитуриентам Фотографии Форум Задать вопрос


Главная
Новости
Общая информация
Структура СыктГУ
Абитуриентам
Студентам
Образование в СыктГУ
Научная работа
Подготовка кадров
Внеучебная жизнь
Международные связи
Учебные материалы

Кафедра коми и финно-угорской филологии

Новости Странички

Контакты:
Телефон: **(8212) 43-58-71, 43-55-81 (доб.103)**
Адрес: **167001, Республика Коми, г.Сыктывкар, ул.Катаева, 9, каб. 414.**
Email: **kkfuy@syktsu.ru**

Состав кафедры:

 **Зав. кафедрой**
Федина Марина Серафимовна

Странички:
Из истории кафедры
Абитуриенту
Учебная деятельность кафедры
Научная работа кафедры
Научно-исследовательская работа студентов
Сотрудничество
Международные связи кафедры
Выпускники кафедры
Новые издания кафедры
Студенческая жизнь
Хроника событий за 2008-2009 учебный год
Хроника событий за 2009-2010 учебный год
Хроника событий за 2010-2011 учебный год

Лингвистическая технология в Тромсё

Добро пожаловать на страницы Giellatekno - саамская языковая технология

Search the site v

Last Published:

Ruoktu OAHPA! Sámi riektáčáálliprošeakta TechDoc Words Plans Wiki

Aarfelsaemien English Norsk Русский

По-русски

- Start page
- Фон
- Интерактивные программы
- Словари
- Тексты и списки
- Другие языки
- Constraint grammar workshop 2011
- Dependency workshop 2011

Добро пожаловать на страницы Giellatekno - саамская языковая технология

Giellatekno, центр саамской языковой технологии, создает языковую технологию для саамских и других северных языков, также средства проверки правописания, инструменты для обработки текстов, педагогические программы, компьютерные словари и преобразования текста в речь.

Инструменты, которые мы создаем, являются необходимым условием для выживания любого языка в современном обществе.

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](#), [инуиакский](#) языки.

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](#), [килдиский-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.

Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

Электронные словари

Машинный перевод

Техническая инфраструктура

Заключение

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.	ö	3	6,92	ь	34	0,03
	і	22	1,32	щ	35	0,02
Коми яз.	ö	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
	ö	24	0,61	ь	34	0,02
	н	26	0,52	ф	33	0,09
Горно-марий. яз.	ӱ	2	6,84	щ	35	0,02
	ӓ	11	4,37	ю	34	0,07
	ӱ	31	0,29	ф	32	0,27
	ö	33	0,16	б	30	0,35

Клавиатура коми языка



Языки, в которых много национальных букв

Удмуртский язык:

25 й, 30 ө, 34 э, 36 ч, 37 ж

Кильдинский язык:

Долгие: ā, ē, ī, ū, ū̄, jā

Точки: ä, ë, ě

Растр.: й, љ, ѓ, џ, р

Другие: һ, ј, њ, њ̄,

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 Дизайне](#)

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](#)

Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

Электронные словари

Машинный перевод

Техническая инфраструктура

Заключение

Языковая технология в сохранении языка

└ Языковая технология в уральских языках

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>
```

Техс-файл для коми, созданный компьютером

```
янсодчыны:янсодчы Verb3 "part company" ;  
ярмыны:ярмы Verb3 "fly into a fury" ;  
явитны:явит Verb2 "spread" ;  
являйтчыны:являйтчы Verb3 "be" ;  
абутöмасьны:абутöмась Verb2 "feign poverty" ;  
адавны:адал Verb1 "gobble up" ;  
аддзавны:аддзал Verb1 "find" ;  
аддзöдчыны:аддзöдчы Verb3 "meet" ;  
аддзöдлыны:аддзöдлы 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

```

72:0 Vdz4i3k
73
74 LEXICON Verb1 ! яйсявны, в - л change
75 +V: Finiteforms ;
76 +V+Inf:%>ны K ;
77
78 LEXICON Verb2 ! яно́дны, 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:%>a PresPret1 ; ! 1,2 Future
129 +Ind+Prs:%>a PresPret1 ; ! 1,2 Present
130 +Ind+Prt1:%>i PresPret1 ; ! 1,2 Preterite1. i и variation?
131 +Ind+Fut:%>ac PresPret1-SgPl3 ; ! 3 PresPret1-SgPl3
132 +Ind+Prs:%>ö PresPret1-SgPl3 ; ! 3
133 +Ind+Prt1:%>i P3 ; ! 3
134 +Ind+Prt2:%>ом Pret2a ;
135
136 Non-finiteforms ;
137
138 LEXICON Non-finiteforms
139 / Page 346 in the Коми grammar

```

Результат: Анализатор коми языка



бета-версия Trondtr сёрнитӧм лист бокӧй лӧсьӧдӧмьяс видзӧдӧм лыддьӧгӧй чӧжӧсӧй сванс эштӧдӧм

гижӧд сёрнитӧм веськӧдны вылӧм ним вежны видзӧдны

Медшӧр лист бок

Материал из Wikipedia



Тайӧ субдоменсӧ видзӧны, медым вӧчны
Википедия
пытшкын **КОМИ КЫВ** кыв вылын гижӧдъяс.

1503 гижӧд

Сямыд кӧ эм сёрнитны комиӧн, сӧлӧм кӧ пӧсь да вирӧн тыр, гиж тӧ комиӧн, комиӧн, медым олим ми дыр! Тайӧ сайт вылын позьӧ гижны унатор. Сӧмын гиж. Кут ассыд здуктӧ. Вӧч-кер мый колӧ.

Юриндалысь  **Бур гижӧдъяс**

Йӧзкотыр · **Му вӧдитӧм** · **Важвылӧм** **Коми республика**
Серпасалун · **Югӧрпас** · **Мӧвпаланбур** Коми Республика

корсысьсӧм

Вуджны

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

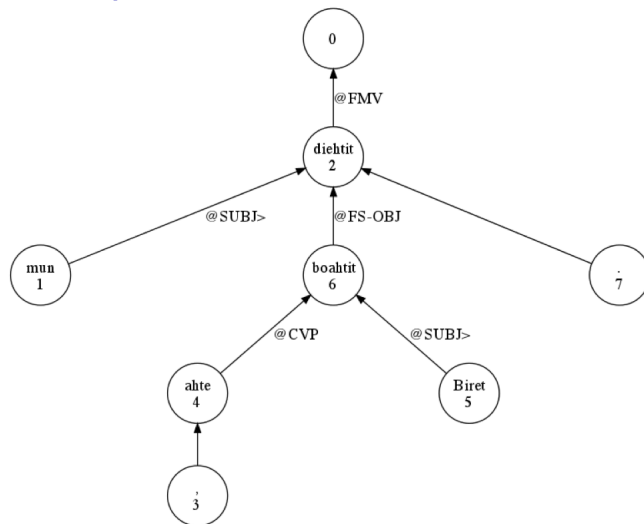
(Анализатор саамского языка)

```

"<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
"<boah tá>"
  "boah tít" <mv> V IV Ind Prs Sg3 @FS-OBJ #6->2
"<.>"
  ". " CLB #7->2

```


(Анализатор саамского языка)

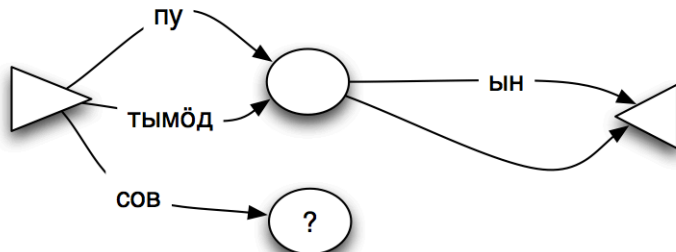


Результат: Словарь коми языка

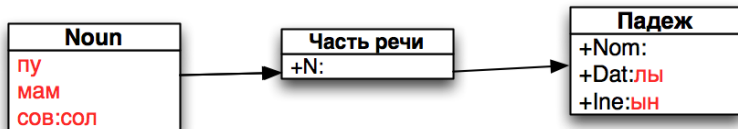
The screenshot shows a web browser window titled "Kaikki viitelähteet". The search bar contains the word "аддзӧдчыны". Below the search bar, there are navigation tabs: "Kaikki", "Sma:Nob-Swe", "Vuosttaš Digisánit", "Oxf", and "FKV:NOB". The main content area displays the word "аддзӧдчыны" in bold, followed by the part of speech "verb". Underneath, there are two sections: "English:" with the translation "1. meet" and "Finnish:" with the translation "1. tavata". A mouse cursor is visible over the text.

Фонология и морфонология

Склонение как автомат



Автомат в программе lex

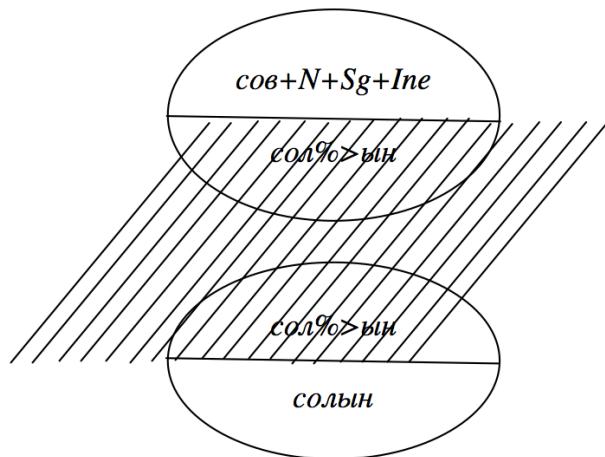


Автомат в программе xfst

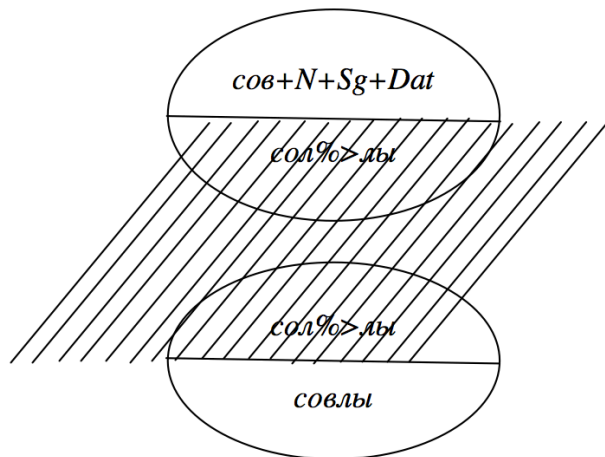
```
[ л -> в | | _ [ .# . | %> Cns ] ] ;
```

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



Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

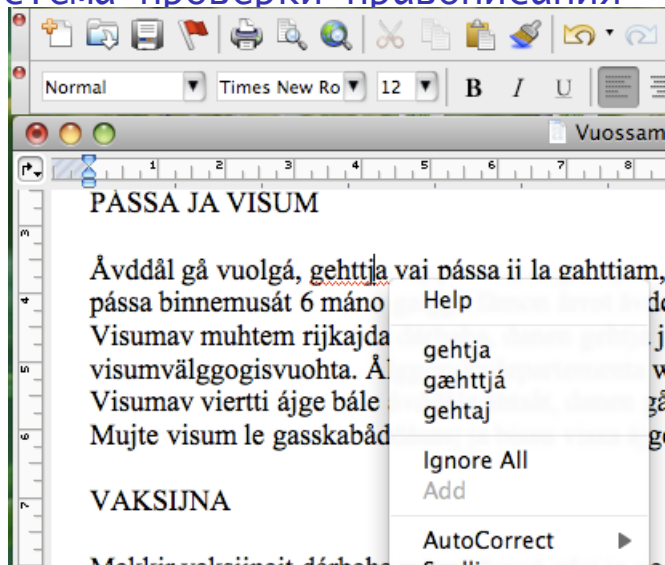
Электронные словари

Машинный перевод

Техническая инфраструктура

Заключение

Система проверки правописания



Электронные словари

Váldosiidu-Hovedsiden | Ođđasat | Mánát mátkkošte čuođi jagi maŋos

Mánát mátkkošte čuođi jagi maŋos

johanante

jahki (subst.)

1. ár

Analyser: sg. gen. el. sg. akk.

Vuosttaš Digisánit

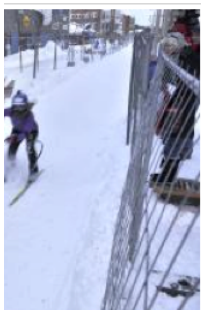
Mer...



skuvlamánáide kulturbealvi gos
ohppe dábálaš dološ doaimmaid
Straumen gillis.

Gaskavahkku ledje Stener, Kvaløya ja
Sandøya skuvla viđat ja guđat

Электронные словари



imunenis šattai garra gilvu
agodii Kuhmunen fiškalit
wven 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 Áltá guovddášgáhttii gilvalit
201 mehtera sprinttas, 16 hearggi
senioroasis, ja guokte hearggi
junioroasis. Son gii **beasa** illudit
eanemusat manne **beassat (verb)**

Risten Sara.

Guvttiin herggiin

ggit, namalassii beakkán Dolgi ja su ođđa hear
nit loahppavurrui ja ii mannan ge guhká ovdal ge žilggai ahte

beasa illudit
beassat (verb)

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

Analyse: indik. pret. 3p. sg.
Vuosttaš Digisánit Mer...

The screenshot shows a web browser window titled "Vuosttaš Digisánit". The search bar contains "beassat". The browser's address bar shows "Alt Wikipedia Lule Saami-X Vuosttaš Digisánit VD_2011". The main content area displays the word "beassat" in a large, bold font. Below it, the word is identified as a verb ("verb"). A numbered list provides three meanings: 1. fritas, 2. slippe (inn), komme (til), and 3. få (anledning til). The entry includes sections for "Analyser" (inf. el. indik. pres. 1p. pl.) and "Nøkkelformer" (key forms) with their respective grammatical forms: indik. pres. 1p. sg. (odne mun) beasan, indik. pret. 1p. sg. (ikte mun) bessen, and indik. pres. neg. (in) beasa. An "Eksempler" section provides three example sentences in Sami with their English translations: "Son beasai militearabálvalusas." (He was in military service.), "In beassan lahka ge wssa." (I was able to get away.), and "Ja dál beassá bárdni sihkastit bivastaga ja bosihastit." (And now the father-in-law can finally get away and get his own house.). To the right of the examples, there are three English sentences: "Han ble fritatt fra militærtjenesten." (He was exempted from military service.), "Jeg kom meg ikke i nærheten av døra engang." (I didn't even get close to the door.), and "Og nå får gutten tørke svetten og puste ut." (And now the boy can dry his sweat and breathe out.). The browser's status bar at the bottom right shows the URL "://.../img/LogoWeb070sh" and standard navigation icons.

Vuosttaš Digisánit

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.

In beassan lahka ge wssa.

Ja dál beassá bárdni sihkastit bivastaga ja bosihastit.

Han ble fritatt fra militærtjenesten.
Jeg kom meg ikke i nærheten av døra engang.
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



navigation

- [Main Page](#)
- [Community portal](#)
- [Current events](#)
- [Recent changes](#)
- [Random page](#)
- [Help](#)

search

toolbox

- [What links here](#)
- [Related changes](#)
- [Upload file](#)
- [Special pages](#)

[page](#) [discussion](#) [edit](#) [history](#) [move](#) [watch](#)

We're in the 2011 [Google Summer of Code!](#) Check out the [ideas page!](#) Trans

Руководство по созданию новой языковой пары

В этом руководстве описывается порядок создания новой языковой пары для системы машинного перевода Apertium. От вас не требуются какие-либо лингвистические знания или знания по машинному переводу, кроме как способности различать части речи (отличать существительные от глаголов, например).

Введение

[\[edit\]](#)

Как вы только что узнали, Apertium является системой машинного перевода. Но если быть более точным, то Apertium следует назвать не системой, а платформой машинного перевода. Он обеспечивает вас "движком" машинного перевода (англ. "engine"). Можно также перевести как "ядро", "механизм" и т. п.) и набором инструментов, с помощью которых вы можете строить свои собственные системы машинного перевода. Единственное, что вы должны сделать, это написать данные. На базовом уровне эти данные состоят из трёх словарей и некоторого набора правил (обеспечивающих перестановку слов и другие грамматические трансформации).

- 1 Введение
- 2 Что вам потребуется
- 3 Из чего состоит языковая пара
- 4 Языковая пара
- 5 Краткое замечание о терминах
- 6 Начало работы
 - 6.1 Одноязычные словари
 - 6.2 Двухязычные словари
 - 6.3 Правила трансфера
- 7 Добавим глаголы
- 8 А что насчёт личных местоимений
- 9 Расскажи же мне о проигрывающих
- 10 Работа с незначительными вкладами
 - 10.1 Анализ
 - 10.2 Генерация
- 11 См. также

Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

Электронные словари

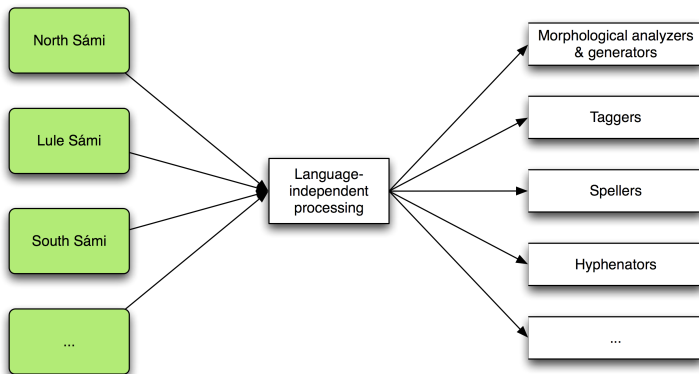
Машинный перевод

Техническая инфраструктура

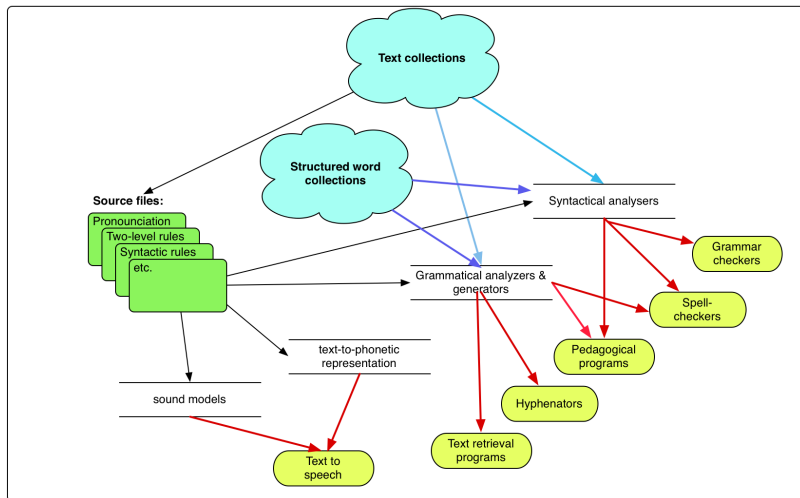
Заключение

Техническая инфраструктура

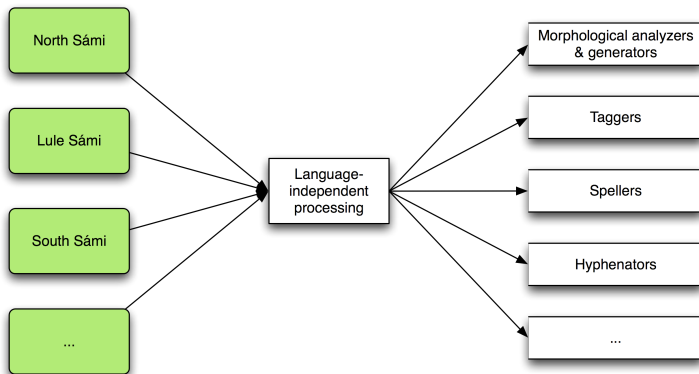
Панорама



Панорама



Панорама



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,

Фон

Клавиатура

Языковая технология в уральских языках

Программы

Система проверки правописания

Электронные словари

Машинный перевод

Техническая инфраструктура

Заключение

Заключение

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

Tay!