

Examples from community-oriented technological solutions

Course: NLP for Endangered Languages of the Amazon.
From a Uralic perspective. Lecture 6.

Jack Rueter, Niko Partanen,
Mika Hämmäläinen & Khalid Alnajjar

Scope of language technology

Language technology and Natural Language Processing focuses often to the use needed for researchers themselves:

- Linguistic analysis, digitizing existing materials into corpora etc.
- Linguistic materials available for *research use*

At the same time we have to recognize the needs of the language communities, and the fact that these materials also belong to them and originate from them.

The communities themselves may be very technically oriented, or can become so in a matter of years or decades.

We go through some examples how we have addressed this issue in our projects.

КОРПУС КОМИ ЯЗЫКА

КОРПУС КОМИ ЯЗЫКА — ЭТО ИНФОРМАЦИОННО-СПРАВОЧНАЯ СИСТЕМА, ОСНОВАННАЯ НА СОБРАНИИ КОМИ ТЕКСТОВ В ЭЛЕКТРОННОЙ ФОРМЕ. НА ЭТОМ САЙТЕ ПОМЕЩЕН КОРПУС СОВРЕМЕННОГО КОМИ ЯЗЫКА ОБЩИМ ОБЪЕМОМ БОЛЕЕ 60 МЛН СЛОВОУПОТРЕБЛЕНИЙ.

начать работу



Слово

Амазон

Тип поиска

Все варианты

Подкорпус

Поиск по всему корпусу

Автор

☐

Название

☐

Год

☐

Источник

☐

Сортировка результата

Год

Порядок сортировки

По убыванию

☐ Учитывать регистр

Поиск

Всего словоупотреблений: 67,271,275

Всего найдено результатов: 113

Время поиска: 0.984

Туй ёртыд век колö! / Ольга Тарачёва // Йöлöга (2021. №4)

... Гижысь, публицист Пётр Романовкöд да сылөн ёртъяскöд, индееч Бобкöд да терьер Джеррикöд, лыддысьысь мöдöдчас Тихой океанлөн Перуса вадорö, каяс Анды гöра вылö да видзöдлас Амазонкаса сельваö. ...

Эжва пöлөн рöдті вöлөн... / Наталья Кузнецова // Коми му (2020-01-23)

... Паджгаса амазонка ...

Медся аслыспöлöс пемöсьяс / Kodko // Йöлöга (2019-09-13)

... Тайö примат, öблезяна, олö Амазонка дорса зера вöръясын. ...

Пон ыджда черань / Kodko // Коми му (2019-07-04)

... Неважөн энтомолог Петр Наскрецки Амазонкаса тропической вöрын паныдасьöма ыджыдсьыс-ыджыд черанькöд. ...

Ватöгыд некыдз / Kodko // Йöлöга (2019-06-21)

... Сы бöрын кузьта серти лоö Амазонка (6.400 километр). ...

Вом тыр змей / Kodko // Коми му (2019-05-10)

... Бразилияысь экология дорйысь Артевал Дуарте важөн нин сулалö вöлöм Амазонкаын вөр пöрöдöмлы паныд. ...

Вом тыр змей / Kodko // Коми му (2019-05-10)

... Мöрччис эз тайö Амазонкаысь вөр пöрöдысьяслы, абу тöдса. ...

Бöръя туземеч / Kodko // Коми му (2018-11-08)

Тавогся июлыкын учёнйöс казвлöмась Амазонияса вöрын олгыкöс весиг мойвийöма снимайтыштны

Сохранить результат

Komicorpora

Created by FU-Lab (The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages) in Syktyvkar.

Led and run by native Komi speakers, and coordinates scanning of Komi books, proofreading, and negotiating the rights with the original authors.

They also translate legislation into Komi.

Several adjacent projects:

- Digital Komi dictionaries
- Online library
- Language learning resources



Коми-русский словарь

Русско-коми словарь

для русскоговорящих

Copyright © 2020. [FU-Lab](#).





krv->rus ▾

кань



Поиск

А Б В Г Д Е Ё Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Э Ю Я

кань



сущ:

кошка, кот

Словарь диалектов коми языка

[О сайте](#) [Инструкция](#)

А Б В Г Д Е Ё Ж З И Й К Л М Н О Ђ Ѓ П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю

Коми сёрнисикас кывчукёр. Словарь диалектов коми языка: в 2-х томах. Сыктывкар: ООО «Издательство «Кола», 2012.

- Т. I.: А–О. – 1096 с. ISBN 978-5-7934-0541-6
- Т. II.: Ђ–Я. – 888 с., илл. ISBN 978-5-9906269-0-4

Настоящий Словарь является результатом совместной работы четырёх авторов. Это научные сотрудники **Института языка, литературы и истории Коми научного центра УрО РАН:**

- **Люция Михайловна Безносикова** (буквы А–И),
- **Надежда Кимовна Забоева** (буквы Й–М),
- **Евгения Авенировна Айбабина** (буквы О–С),
- **Раиса Ивановна Коснырева** (буквы Н, Т–Я).

Работа подготовлена под руководством и редакцией Л. М. Безносиковой; ею проверена каждая словарная статья рукописи и в необходимых случаях внесены коррективы и дополнения; ею же написаны предисловие и сведения для пользующихся словарём, а также составлен список наименований населённых пунктов Республики Коми, отличающихся от официальных названий.

Составители Словаря приносят глубокую благодарность Н. И. Лоскутовой, Э. К. Павловой за участие в пополнении картотеки на начальном этапе работы, а также за постоянную помощь и добрые советы в уточнении значений вновь выявленных диалектных слов; В. К. Хабаровой за квалифицированное осуществление компьютерного набора значительной части рукописи; С. А. Сажинной за предоставление экспедиционных записей печорского диалекта. Авторы Словаря выражают признательность зав. сектором фольклора Института языка, литературы и истории КНЦ УрО РАН Ю. А. Крашенинниковой, научному сотруднику названного сектора А. Н. Рассыхаеву за предоставление ряда слов, зафиксированных ими в лузско-летском диалекте коми языка; любителям народной речи Н. С. Калинин, Е. Е. Афанасьевой, уроженцам Удорского района, и С. И. Елфимову, жителю г. Сыктывкара, за предоставление довольно большого количества диалектных слов, бытующих в их говорах.

Оригинальный текст словаря предоставил директор издательства «Кола» Николай Вахнин.

Над разработкой электронной формы словаря работали Инна Андрианова, Елена Кожевина и Дмитрий Левченко.



кась

[kaś]

вс. (Гр. уж.) лл. сс. (Кур.) уд.; см. [кань](#) /

лл. [касьыс парччасьӧ](#) кошка царапается

уд. (Лат.) [мысьтӧм кась тэ \(аб кӧ мысьсьӧма\)](#) ты грязная кошка (если не умыт)

лл. (Пр.); примета [кась парччасьӧ — лым вайӧ](#) кошка царапается — к снегу

сс. (Кур.); примета [пон да кась кӧ турун сёвӧны — зэр водзӧ](#) если собака и кошка едят траву — будет дождь

уд. [Ӧ синмас эд касьӧн он ка](#) насильно мил не будешь (букв. в глаза кошкой не кинешься)

вс. (Уж.) [касьӧн синад он чеччыш](#) насильно мил не будешь (букв. в глаза кошкой не кинешься)

уд. (Крив. Пучк.) [Ӧ касьӧн каас тэ вылад](#) так и лезет на тебя

уд. [Ӧ касьӧн каны](#) всё время придирается к кому-л.

лл. (Лет.) [Ӧ как кась олӧ](#) кошачья жизнь; живёт как кошка (о лёгкой жизни)

вс. (Уж.) лл. (Зан. Об.) [Ӧ кась козинӧн козынасьны](#) отобрать подарок или просить вернуть дарёное

Videocorpora / Komi Mediateka

In 2014–2016 a large language documentation project made new recordings from one Komi dialect.

The materials are archived to the Max Planck Institute in Nijmegen and the Language Bank of Finland for research use.

Selected materials were also put online into open multimedia platform, in order to provide easy access to the communities themselves.

The scientific archiving solutions are not necessarily accessible, so we ended up to this duplicate system. Not ideal, but works!



КОМИ МЕДИАТЕКА

[О ПРОЕКТЕ](#) [ПОИСК](#) [ВСЯ МЕДИАТЕКА](#) [НАША КОМАНДА](#) [КАРТА](#) [КАК ПОЛЬЗОВАТЬСЯ](#) [КОНТАКТЫ](#)

© 2016 Все права защищены.

Рохир Блокланд, Василий Чупров, Дмитрий Левченко, Мария Федина, Марина Федина, Нико Партанен, Михаэль Русселер.

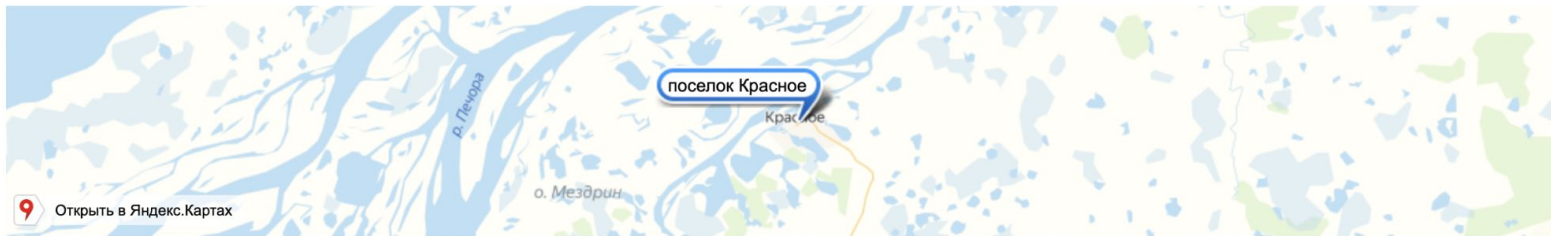
2016. Коми медиатека. Сыктывкар: FU-Lab. URL: <http://videocorpora.ru>



Login

[Home](#)[Whole media collection](#)[Map](#)[Our team](#)[About the project](#)[Instructions](#)[Contacts](#)

Анна Степановна Власова



▶ || song(00:00:00 - 00:00:05)

KPV: Коми му кузя ми мунам,

RUS: Мы идем по Коми земле,

ENG: We go along the Komi land,

▶ || song(00:00:05 - 00:00:12)

KPV: Гӱгер сулалэ съӱд вӱр.

RUS: Кругом стоит темный лес.

ENG: The black forest stands around us.

▶ || song(00:00:12 - 00:00:17)

KPV: Вӱрыс вувті-вувті уна,

RUS: Леса очень-очень много,

Videocorpora / Komi Mediateka

Technical questions:

- Maintaining the same materials in several locations is difficult
- Komi Mediateka is a local custom solution (code not easy to reuse)
- The transcriptions in Komi Mediateka have not been updated for a while, but there have been many corrections to the ELAN files
- Maintaining this costs work and money, and both are always a challenge
- Still, the materials are out there, and the users have been satisfied

It's always possible to imagine the perfect solution for everything!

If that is not available, some solution must be good enough!

Komi dictionaries

The work is currently ongoing to combine dictionary work in Komi and Finland

We are planning the further work, and discuss with collaborators

- The goal would be to combine the available dictionaries and sound examples

There is a strong tradition of individual groups or researchers making their own dictionaries, but we think a more collaborative approach is needed

- This is the main motivator behind our work on Ve'rdd

Especially the dialectal lexicon is important when we analyze ELAN transcriptions

Work principles arising from technical needs

- Ideally we store each information only once
- We must be able to update whatever system we use
- The materials should be licensed somehow, so we know what reuse is ok
- The same information is useful for different purposes!

Examples:

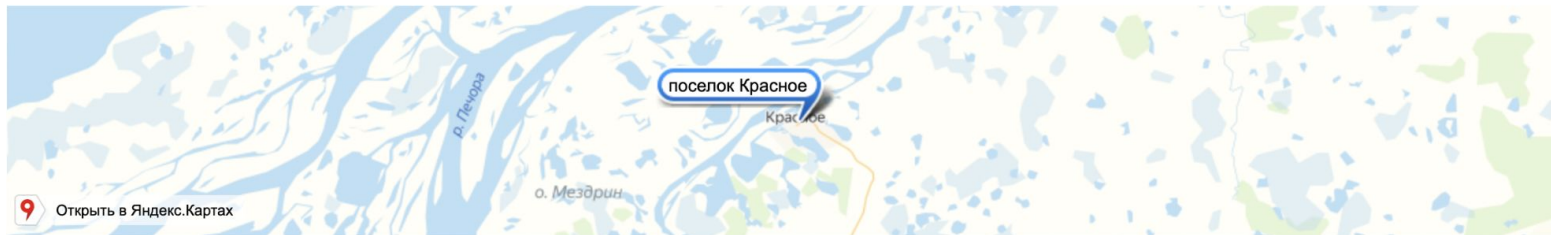
- Transcriptions and audio can be used to train ASR systems
- Dictionary word and sentence data is also useful for this
- Dictionary is an important component of different morphological analysers

Example: Coordinates

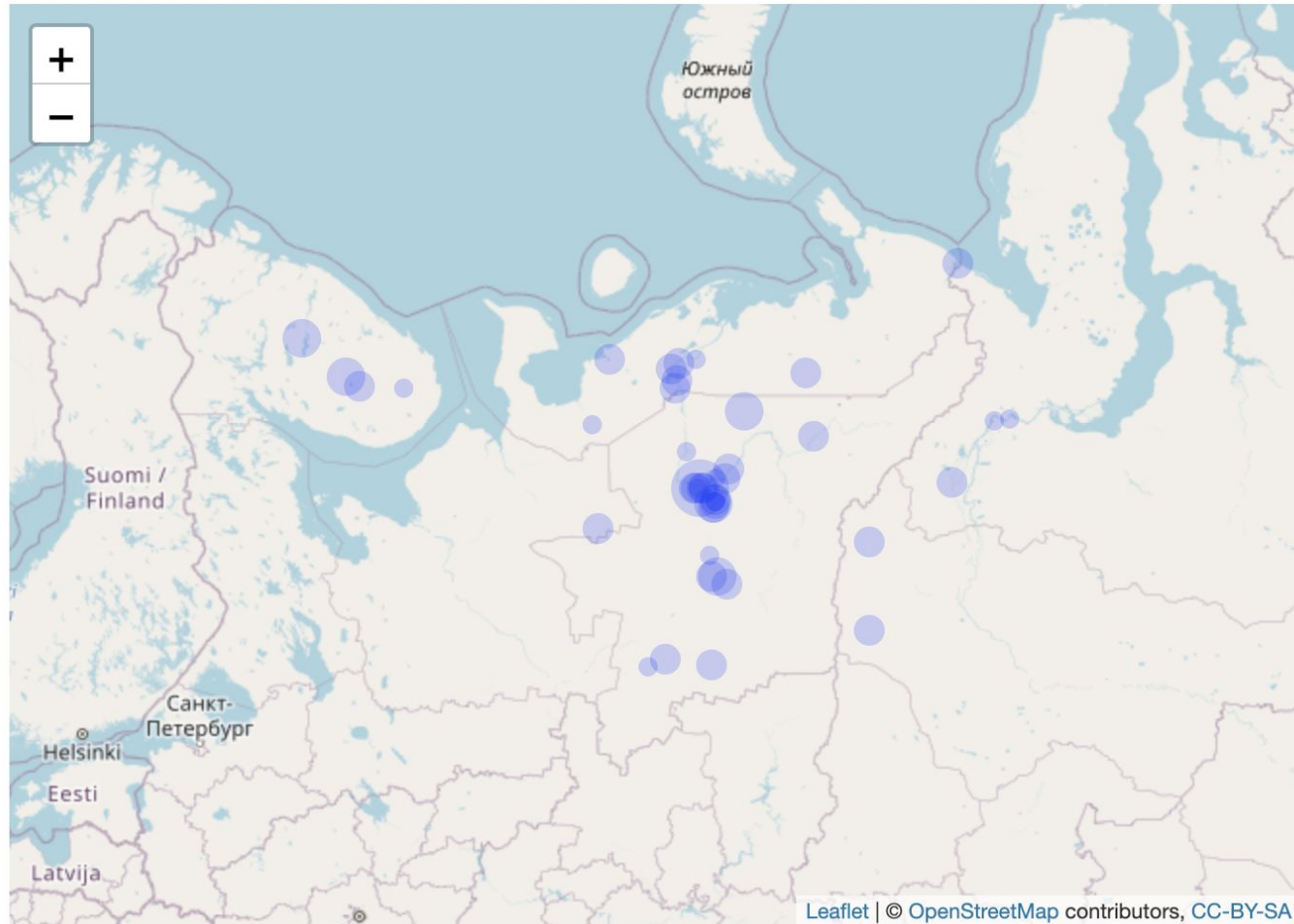
Recordings usually have at least three types of coordinates:

- Recording location
- Birthplace of the speaker
- Place of residence of the speaker

We can use them to visualize our collections in different environments, and they are very valuable for our research. We are also wondering what to do!



Geographic distribution



Interactive reports

For our project I have set up a web page that shows the corpus status:

- <https://langdoc.github.io/kpv/>

In principle it should update automatically (this gets bit complicated).

The idea is very simple:

- A script reads the corpus and the metadata
- Different features are counted, and visualized in various ways
- This helps us to understand when things go wrong

Next few other examples.

Paasonen's Mordvin materials

Map of Paasonen's fieldwork (1891–1912) of the Erzya and Moksha

Map of Paasonen's Erzya fieldwork (1891–1912)

Map of Paasonen's Moksha fieldwork (1891–1912)

Feoktistov & Saarinen, 2005 Moksha Dialects

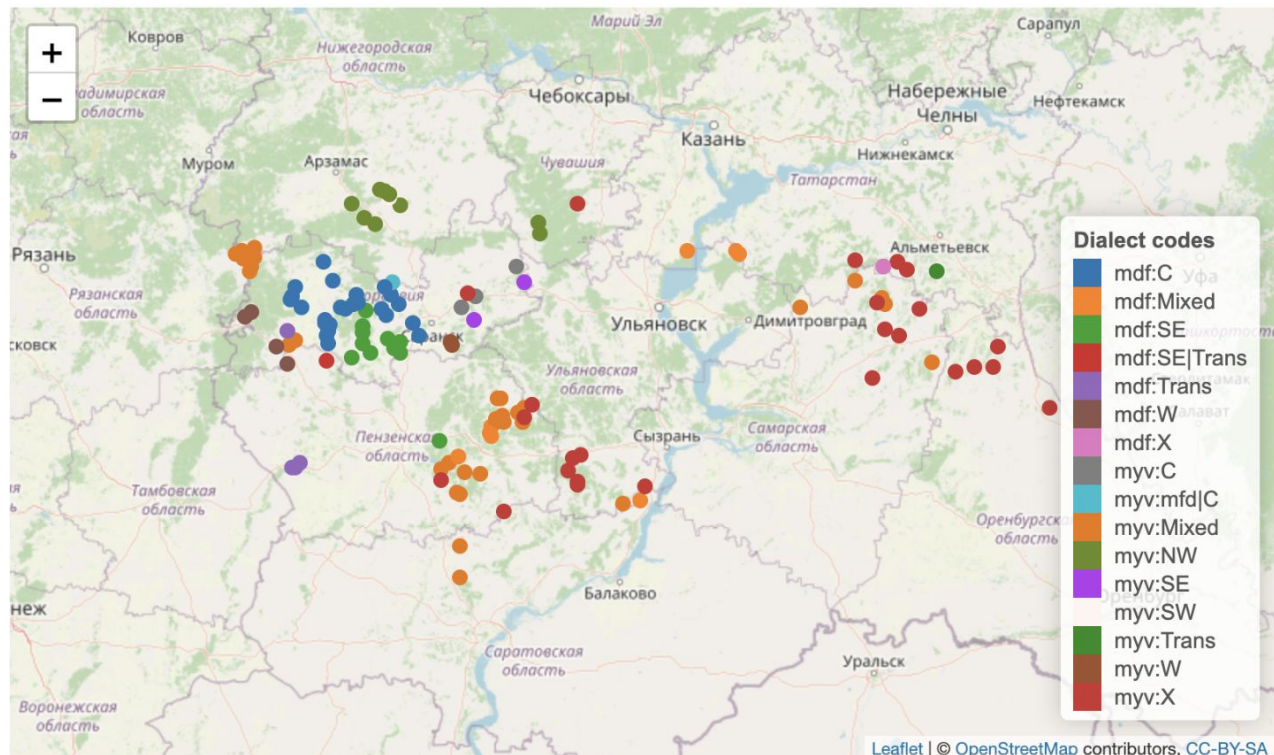
Mordvin settlements in Bashkortostan (2014)

Mordvin populus attested in from census registry (2002)

Paasonen's Mordvin materials

Map of Paasonen's fieldwork (1891–1912) of the Erzya and Moksha

These are all locales, Erzya and Moksha, where fieldwork materials were collected for and Heikki Paasonen 1891–1912. The dialects of each language within the Republic of Mordovia have been classified as five groups by Aleksandr Pavlovich Feoktistov. The some locales have not been classified, especially outside of the Republic, and other locales have more than one classification. By clicking the dot, you will reveal the location. (<https://www.sgr.fi/fi/items/show/413>)



Comparative Mordvinic database

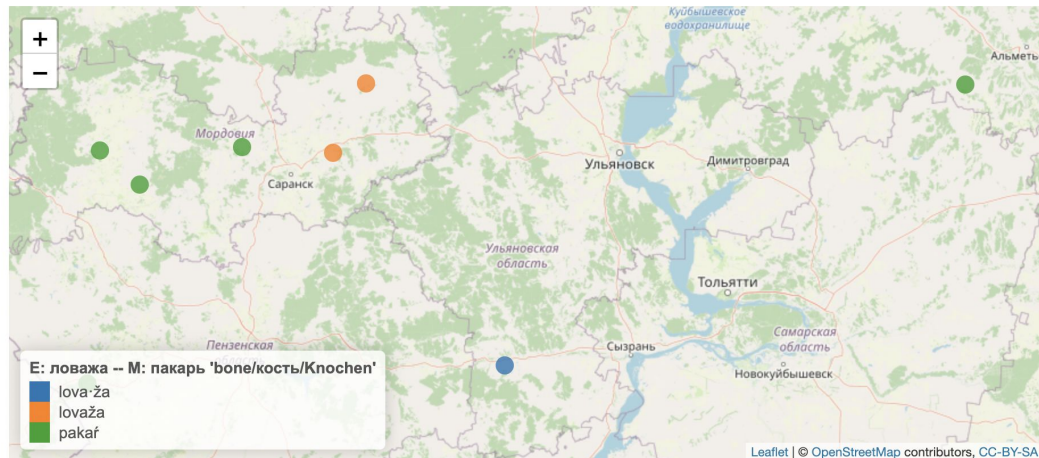
Essentially contains simple tables with the feature variants and locations

These are rendered to HTML

The system is extremely simple, but thereby easy to maintain

Similar map-views can also be made directly from the corpus data, if we have the coordinates.

```
Erzya-Lit,lovaža,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Moksha-Lit,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Erzya-Mar,lovaža,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Erzya-Ba,lovaža,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Erzya-Af,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Moksha-Čemb,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Moksha-Saz,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Moksha-Lemd,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,  
Moksha-Kars,pakař,E: ловажа -- M: пакарь 'bone/кость/Knochen',lexical,
```



Prerequisites for database printouts

language, value, feature, category, comment

- Collection (language variant)
 - Geographical coordinates for place of collection
 - Temporal coordinates
- Item (phonology, morphological forms, lexicon)
 - In transcription
 - Normalized forms
- Definitions
 - Morphological describers
 - Various languages, perhaps even own language
- Categories
 - Phonology, morphology, lexicon
- Others

Example

We search from the corpus all verbs that end to *-in~~i~~*, *-isn~~i~~*, *-in~~i~~s* or *-isn~~i~~s*

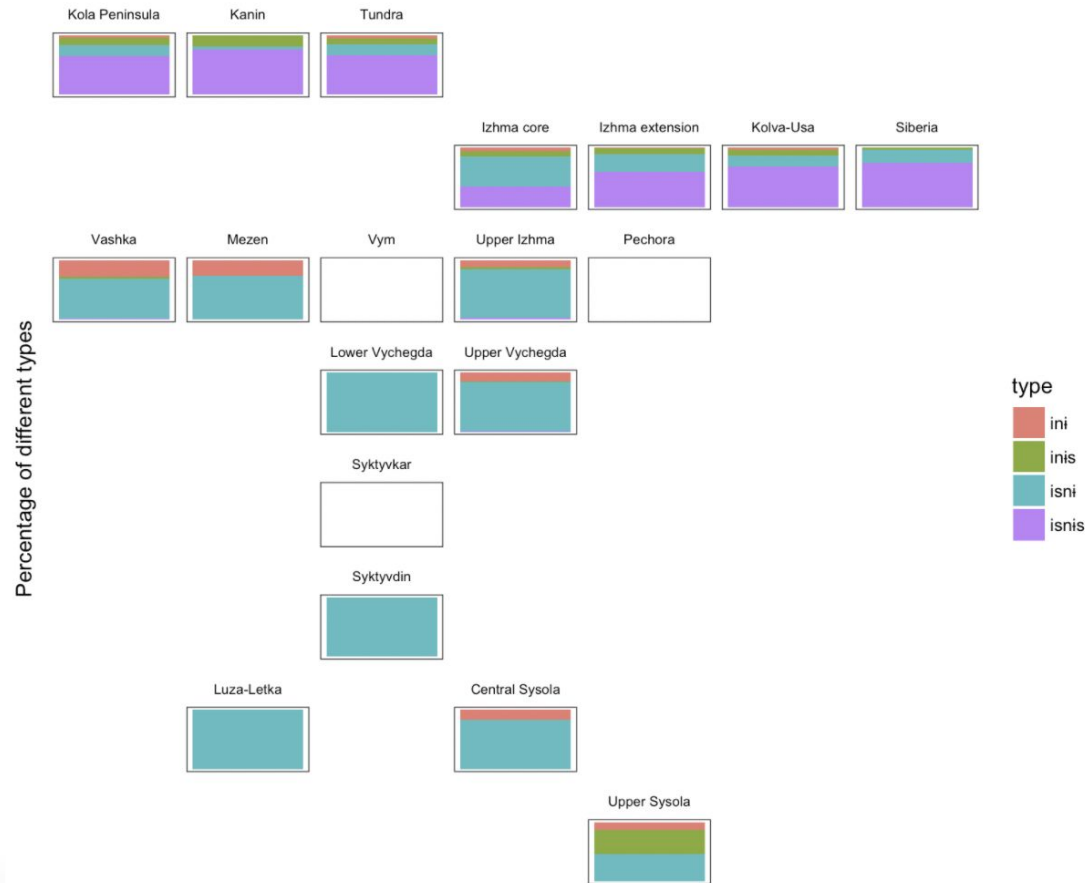
Each speaker is associated to a dialect area (unclear cases are ignored)

The verbs are classified to **four types**, corresponding to the allomorphs

The frequencies of the allomorphs are counted for each **dialect area**.

First preterite plural verb allomorphs in Komi-Zyrian dialects

Map approximates the location and contact relations of the dialects. For blanks no data available.

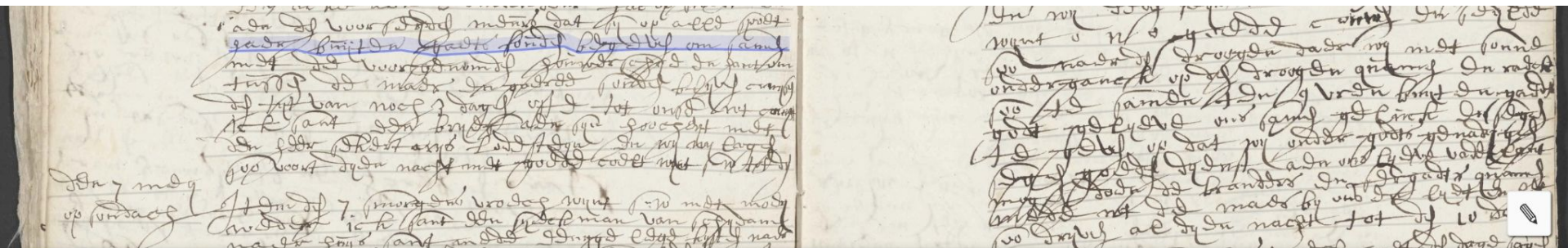


Next week

Jack will be on the Livonian coast in Latvia

Niko will be hosting a text recognition workshop with some colleagues next week

Suggestion: We unite the forces, and do the workshop at this time?



x²

x₂

B

/

U

abe

Special Characters

Annotate ...

?! Unclear

16	van nouwerschye en isack nantom in see nijet	#
17	vernomen dye selven naer myddach gaf ick	#
18	een acke aen comandeur Jacop swart en	#
19	aen den voorseyden meurs dat sy op alle spoet	#
20	haer buijten gaets souden begeven om samen	#
21	met de voorgenomden houwerschve en hantom	#

Links:

Online corpus: <http://komicorpora.ru/>

Online Komi–Russian–Komi dictionary: <http://dict.komikyv.ru/>

Other digital dictionaries: <https://dict.fu-lab.ru/>

Online library: <http://komikyv.org/>

Multimedia collection: <https://videocorpora.ru/>

Language learning platform: <http://komikyv.ru/>

Databases:

- <https://rueter.github.io/Mordvin-Varieties/database.html>
- <https://langdoc.github.io/comparative-permic-database/>