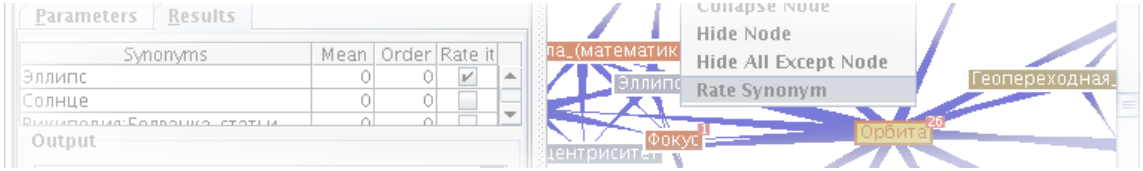


Корпус вепсского языка: проблемы и перспективы

Veps corpus

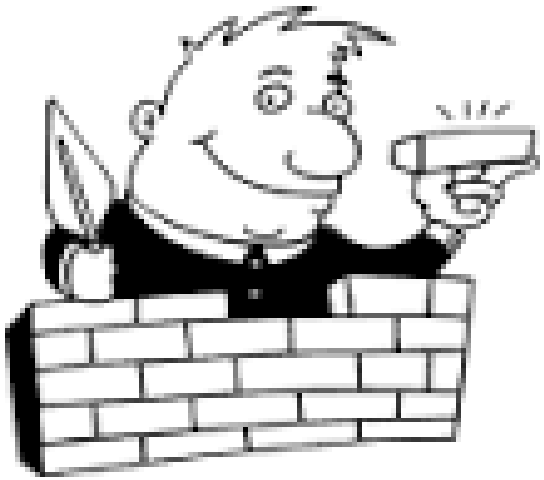
[1]Institute of Language, Literature and History
[2]Institute of Applied Mathematical Research of the Karelian Research Centre of the RAS

[1]Nina Zaitseva
[2]Andrew.Krizhanovsky gmail.com

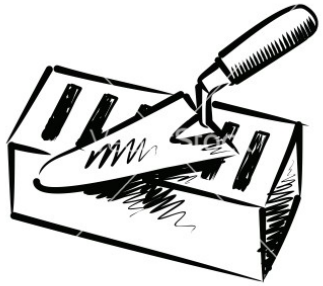
Goal

To create spoken

and written corpus



First task



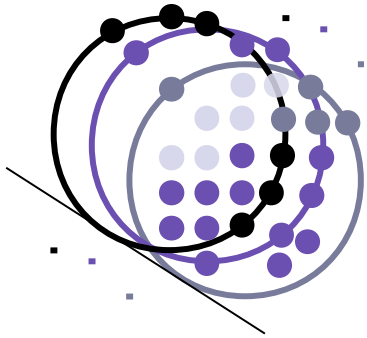
Develop software
(corpus as a web-site).

Second task

Fill data:

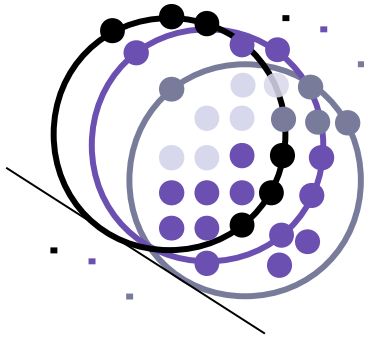
add texts, add words.

What is Corpus Linguistics?



Corpus Linguistics is the study of language/linguistic phenomena through the analysis of data obtained from a corpus.

Why to use a *corpus*?

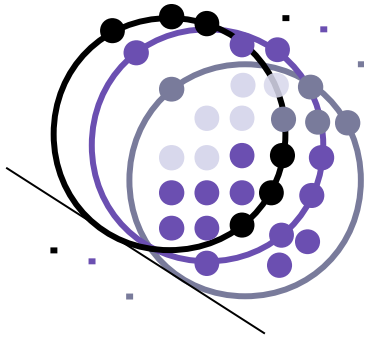


- Intuition alone is not enough
  - Is "*starting*" always replaceable by "*beginning*"?
  - Is it only "*time*" that is "*immemorial*"?
  - "*think of*" vs. "*think about*"

- Native speaker intuition is unreliable
  - provides no information on frequency of occurrence
  - "***head***" => body part - Is this the most used sense?

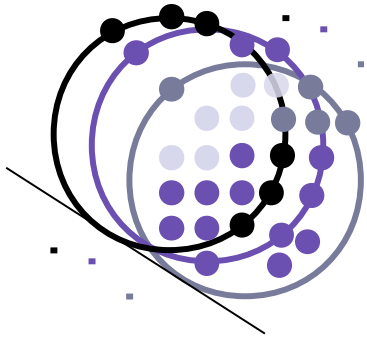
- Create corpus once, use for different linguistics tasks

Тексты, входящие в состав корпусов

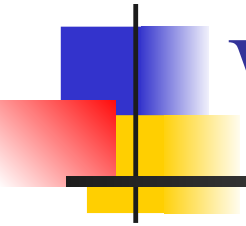


- отобраны исходя из определенных принципов,
- специально подготовлены и размечены,
  - машиночитаемый формат + разметка
- с помощью специальных программ в них можно искать необходимые фрагменты текста по заданным параметрам.

Types of corpora



- spoken vs. written
- monolingual vs. bi/multilingual
- parallel vs. comparable corpora (translation corpora)
- general language purpose vs. specialised language purpose
- diachronic vs. synchronic
- plain text vs. annotated (tagged) text



Web corpora

Корпусы(а) в Интернете

Russian corpora

Национальный корпус русского языка **http://ruscorpora.ru**	>500 млн слов
Открытый корпус русского языка **http://opencorpora.org**	> 1 млн слов

Finno-Ugric corpora

Языковой банк Финляндии **http://www.csc.fi/tutkimus/alat/kielitiede**	Broken link
Справочный корпус эстонского языка **http://www.keeletehnoloogia.ee/projects-1/the-reference-corpus-of-the-estonian-language/comprehensive-corpus-of-estonian?set_language=en**	Broken link
Фонетический корпус спонтанной эстонской речи **http://www.murre.ut.ee/phonetic-corpus/**	20 hours, 20 speakers
Венгерский национальный корпус **http://mnsz.nytud.hu/index_eng.html**	> 187 million words

Veps corpus

Veps dictionary + corpus are available at:

http://vepsian.krc.karelia.ru/

Корпус вепсского языка

О проекте Поиск в словаре Поиск по текстам Тексты Подкорпус причитаний Ссылки

Поиск в словаре

Statistics 1

Статистика

Новые материалы: леммы, словоформы, переводы, омонимы

Число лемм в словаре: 2750

Число словоформ в словаре: 8408

Wild card 2: % _

Umlaut 3

◉ везде ○ в леммах ○ в словоформах

__lud% Найти

ä ö ü č š ž ’

Ваш запрос: **__lud%**

лемма: **kuluda**, POS: **V** (глагол)

№	словоформы	Грамматические признаки
1.	kulub	
2.	kulu	
3.	kului	

Перевод
русский: слышаться
английский: to be heard

лемма: **kuludas**, POS: **V** (глагол)

№	словоформы	Грамматические признаки
1.	kuluškanzihe	

Перевод
русский: послышаться
английский: to be heard

лемма: **polut**, POS: **N** (существительное)

№		Грамматические

Поиск по текстам

Статистика
Новые материалы: тексты, источники
Число текстов в корпусе: 1097
Число библиографических источников: 871

Пример: vasan, minä или ленточка, или %o_o_ни%

В текстах vasan Найти

ä ö ü č š ž ’

приближённый поиск лемм и словоформ в текстах,
то есть поиск подстрок (по умолчанию выполняется точный поиск)

Ваш запрос: **vasan**

Леммы в словаре: vas

Перевод лемм: vas (квас || bread juice | kvass | quass)

Словоформы в словаре: vasan, vas-se, vas

По запросу найдено текстов: 2

1. Kut tegiba vasan
   Язык и народ: тексты и комментарии, (2002), с. 67-68
   Карпов Никанор Леонтьевич, г.р. 1897, место записи: Ладва (Ladv), Подпорожский р-н, Ленинградская обл.

1. Kut tegiba vasan

   Язык и народ: тексты и комментарии, (2002), с. 67-68

   Карпов Никанор Леонтьевич, г.р. 1897, место записи: Ладва (Ladv), Подпорожский р-н, Ленинградская обл.

1.	Potom liiban paned sin’n’a da kipetkoižuu valad viluu, da vuu sepid paned dei vas tegese čoma.
2.	Ende tegiba vasan, ka rugihen ligotadas da id’atadas da päče pandas.
3.	Tegižihe imen, nu da siť kuivatas da jouhtas da pašttas liibeižiks, nu i vasan pandas.
4.	Otad, vasan valad, lukhiineižid da reť kašt da babušt.
5.	Uuged ezmäks paned mišto ii jokseiž liib, riigun kohtha, riig mi jokseb vas-se.

2. Van’ka-vor

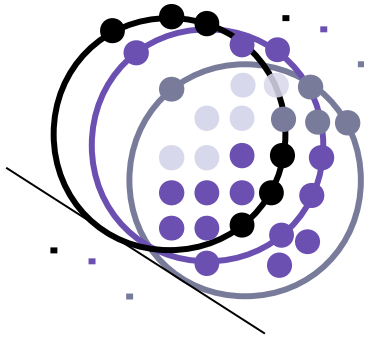
   Вепсские народные сказки, (1996), с. 122-124; ф/архив ИЯЛИ КарНЦ РАН: № 2624/2, НА КарНЦ, кол.83, ед.хр.146

   Микшина Марфа Захарьевна, г.р. 1910, место записи: Ладва (Ladv), Подпорожский р-н, Ленинградская обл., г. записи: 1980, записали: Онегина Нина Федоровна

По запросу найдено текстов: 2,

из этих текстов найдено цитат, содержащих искомые слова из словаря: 5

Veps corpus



User

- Search in dictionary
- Search in corpus

Editor (admin)

- Edit dictionary
- Edit texts in corpus

Все тексты

Выберите подкорпус

- ☐ диалектные тексты (199)
- ☐ библейские тексты (переводные) (431)
- ☐ младописьменный подкорпус (89)
- ☑ подкорпус вепсских причитаний (47)
- ☐ подкорпус вепсских сказок (55)

свойство текста

- ☑ северновепсский диалект (27)
- ☐ средневепсский диалект (214)
- ☐ южновепсский диалект (57)
- ☐ восточные говоры (32)
- ☐ западные говоры (31)
- ☐ свадебные причитания (29)
- ☐ похоронные и поминальные причитания (18)
- ☐ публицистические тексты (51)
- ☐ художественные тексты (21)
- ☐ тексты для детей (17)

выбрать

Подкорпус вепсских причитаний

Список текстов только для одного языка: вепсский

Северновепсский диалект

Найдено текстов: 5.

1. Nevestan voik svad'ban päiväl
   северновепсский диалект, свадебные причитания

Nevestan voik svad’ban päiväl

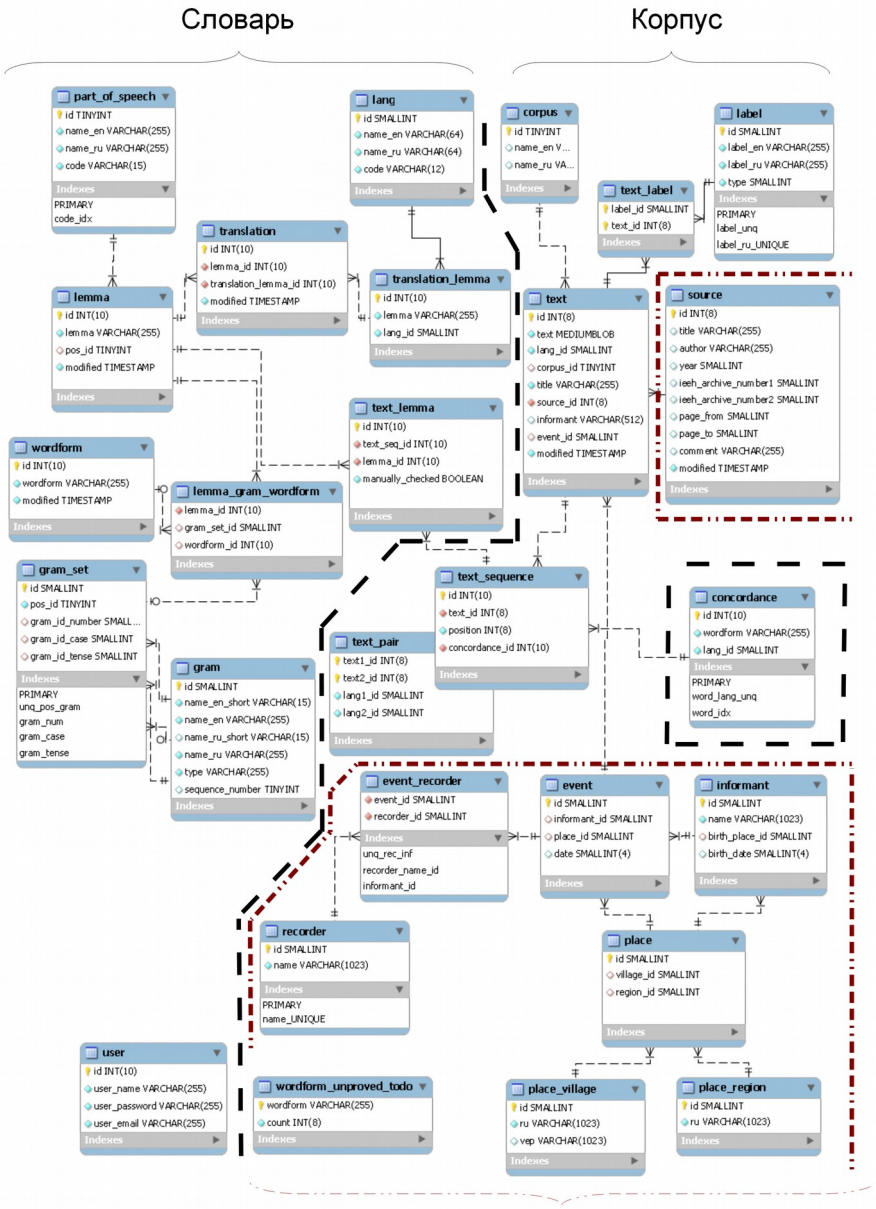
подкорпус вепсских причитаний

северновепсский диалект, свадебные причитания

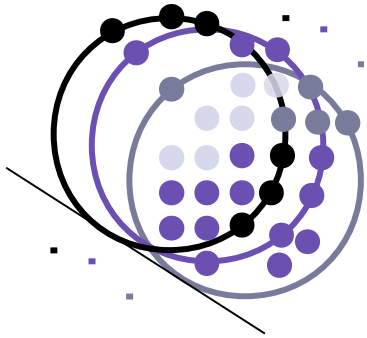
Информант: Ермолаева Анастасия, г.р. 1899, урожд. Шелтозеро (Šoutarv), Прионежский р-н, Республика Карелия, место записи: Шелтозеро (Šoutarv), Прионежский р-н, Республика Карелия, г. записи: 1962, записали: Лонин Рюрик Петрович

Nevestan voik svad’ban päiväl	**Плач невесты в день свадьбы**
Tänambeižel päivaižel om milei-d’o svad’baine.	В сегодняшний денечек у меня уже свадьба.
Kogozihe kaik minun rodn’aine.	Собралась вся моя родня.
Milei vaise üht ii täudu bat’uškod-se rodnijad.	Не хватает мне одного батюшки родного.
Armhad tö minun kaik läheližed,	Любимые вы мои родственники,
armhad tö minun susedaižed,	любимые вы мои соседушки,
avaikat tö mili dorogaine,	откройте вы мне дороженьку,

dictorpus


Словарь
Корпус
part_of_speech
id TINYINT
name_en VARCHAR(255)
name_ru VARCHAR(255)
code VARCHAR(15)
Indexes
PRIMARY
code_idx
lang
id SMALLINT
name_en VARCHAR(64)
name_ru VARCHAR(64)
code VARCHAR(12)
Indexes
corpus
id TINYINT
name_en V...
name_ru VA...
Indexes
label
id SMALLINT
label_en VARCHAR(255)
label_ru VARCHAR(255)
type SMALLINT
Indexes
PRIMARY
label_unq
label_ru_UNIQUE
text_label
label_id SMALLINT
text_id INT(8)
Indexes
translation
id INT(10)
lemma_id INT(10)
translation_lemma_id INT(10)
modified TIMESTAMP
Indexes
translation_lemma
id INT(10)
lemma VARCHAR(255)
lang_id SMALLINT
Indexes
lemma
id INT(10)
lemma VARCHAR(255)
pos_id TINYINT
modified TIMESTAMP
Indexes
text
id INT(8)
text MEDIUMBLOB
lang_id SMALLINT
corpus_id TINYINT
title VARCHAR(255)
source_id INT(8)
informant VARCHAR(512)
event_id SMALLINT
modified TIMESTAMP
Indexes
source
id INT(8)
title VARCHAR(255)
author VARCHAR(255)
year SMALLINT
ieeh_archive_number1 SMALLINT
ieeh_archive_number2 SMALLINT
page_from SMALLINT
page_to SMALLINT
comment VARCHAR(255)
modified TIMESTAMP
Indexes
text_lemma
id INT(10)
text_seq_id INT(10)
lemma_id INT(10)
manually_checked BOOLEAN
Indexes
wordform
id INT(10)
wordform VARCHAR(255)
modified TIMESTAMP
Indexes
lemma_gram_wordform
lemma_id INT(10)
gram_set_id SMALLINT
wordform_id INT(10)
Indexes
gram_set
id SMALLINT
pos_id TINYINT
gram_id_number SMALL...
gram_id_case SMALLINT
gram_id_tense SMALLINT
Indexes
PRIMARY
unq_pos_gram
gram_num
gram_case
gram_tense
text_sequence
id INT(10)
text_id INT(8)
position INT(8)
concordance_id INT(10)
Indexes
concordance
id INT(10)
wordform VARCHAR(255)
lang_id SMALLINT
Indexes
PRIMARY
word_lang_unq
word_idx
text_pair
text1_id INT(8)
text2_id INT(8)
lang1_id SMALLINT
lang2_id SMALLINT
Indexes
gram
id SMALLINT
name_en_short VARCHAR(15)
name_en VARCHAR(255)
name_ru_short VARCHAR(15)
name_ru VARCHAR(255)
type VARCHAR(255)
sequence_number TINYINT
Indexes
event_recorder
event_id SMALLINT
recorder_id SMALLINT
Indexes
unq_rec_inf
recorder_name_id
informant_id
event
id SMALLINT
informant_id SMALLINT
place_id SMALLINT
date SMALLINT(4)
Indexes
informant
id SMALLINT
name VARCHAR(1023)
birth_place_id SMALLINT
birth_date SMALLINT(4)
Indexes
recorder
id SMALLINT
name VARCHAR(1023)
Indexes
PRIMARY
name_UNIQUE
place
id SMALLINT
village_id SMALLINT
region_id SMALLINT
Indexes
user
id INT(10)
user_name VARCHAR(255)
user_password VARCHAR(255)
user_email VARCHAR(255)
Indexes
wordform_unproved_todo
wordform VARCHAR(255)
count INT(8)
Indexes
place_village
id SMALLINT
ru VARCHAR(1023)
vep VARCHAR(1023)
Indexes
place_region
id SMALLINT
ru VARCHAR(1023)
Indexes
Паспорт, источники в корпусе текстов

Future work



Open corpus of Veps and Karelian languages (VepKar) (Открытый корпус вепсского и карельского языков)

- Open software license,
- Open (Creative-commons cc-by) data

Thank you

Veps corpus:

http://vepsian.krc.karelia.ru/

The paper “Корпус вепсского языка”:

http://scipeople.com/publication/121149/