

ВепКар: руководство для пользователей

Крижановская Н. Б., Новак И. П., Пеллинен Н. А.

Издание 1. Декабрь 2022

Содержание

Введение	4
1. Принципы работы в современных корпусах текстов	4
1.1. Корпусная лингвистика и виды корпусов	4
1.1.1 Основные понятия корпусной лингвистики	4
1.1.2. Разметка корпусов	6
Понятие разметки	6
Лингвистическая разметка	6
Экстравербальная разметка	6
1.1.3. Типология корпусов	7
1.1.4. Обзор существующих корпусов различных типов	7
Зарубежные корпусы	8
Корпусы русского языка	8
Корпусы уральских языков	9
1.2. Описание корпуса ВепКар	11
1.2.1. История создания	11
1.2.2. Архитектура ВепКар	11
1.2.3. Страница с текстом в корпусе ВепКар	13
1.2.4. Словарная статья в корпусе ВепКар	15
1.2.5. Открытость данных	17
2. Методические рекомендации по использованию интернет-ресурса ВепКар	17
2.1. Применение интернет-ресурса ВепКар	17
2.2. Рекомендации по использованию интернет-ресурса ВепКар	18
2.2.1. Общие вопросы	18
Шаг 1. Как найти сайт Открытого корпуса вепсского и карельского языков (ВепКар) в Интернете?	19

Шаг 2. Как переключить интерфейс сайта ВепКар на русский язык?	21
Шаг 3. Обязательно ли нужно регистрироваться на сайте ВепКар?	21
Шаг 4. Как ввести особые графические символы карельского и вепсского языков, если их нет на клавиатуре?	21
Стоит особо отметить, что в процессе поиска символ ѹ приравнен к символу у, w - к u / y / ѹ, а знак палатализации ' не учитывается.	21
Шаг 5. Как использовать простые шаблоны в текстовых полях?	21
Шаг 6. Как использовать специализированные шаблоны в текстовых полях?	21
2.2.2. Поиск в корпусе ВепКар	23
Шаг 7. Как осуществлять поиск текстов в корпусе ВепКар?	23
Шаг 8. Как найти текст на вепсском / карельском языке?	23
Шаг 9. Как осуществлять расширенный поиск по текстам ВепКар?	24
Шаг 10. Как найти озвученный текст на вепсском / карельском языке?	25
Шаг 11. Как найти вепсские сказки, записанные до 1950 года?	25
Шаг 12. Как осуществить поиск примеров употребления слова в корпусе?	26
Шаг 13. Как найти примеры употребления в корпусе людиковских имен в форме аппроксиматива?	26
Шаг 14. Как найти примеры употребления в корпусе ливвиковских глаголов в форме перфекта и плюсквамперфекта кондиционала?	27
Шаг 16. Как найти примеры употребления конкретной словоформы в вепсских текстах 30х годов?	30
Шаг 17. Как найти частотные словари корпуса?	31
Шаг 18. Как найти самые частотные слова в текстах?	31
Шаг 19. Как найти «Речевой корпус»?	32
Шаг 20. Как найти и прослушать образцы речи интересующего населённого пункта?	33
Шаг 21. Как ссылаться на материалы корпуса ВепКар?	34
2.2.3. Поиск в словаре ВепКар	34
Шаг 22. Как осуществлять поиск в словаре ВепКар?	34
Шаг 23. Как найти вепсское / карельское слово по начальной форме?	34
Шаг 24. Как указать язык/наречие при поиске леммы?	35
Шаг 25. Как осуществлять расширенный поиск в словаре ВепКар?	35
Шаг 26. Как найти вепсское (карельское) слово по русскому толкованию?	36
Шаг 27. Как найти вепсское (карельское) слово по словоформе?	36

Шаг 28. Как найти все глаголы новописьменного северокарельского языка, начинающиеся с буквы ‘m’, у которых начальная форма (инффинитив) оканчивается на ‘ио’ или ‘уё’, а форма 1 лица ед.ч. настоящего времени оканчивается на ‘ин’ или ‘ун’?	37
Шаг 29. Как найти обратный словарь?	37
Заключение	38
Тезаурус	39
Список использованной литературы	39

Введение

Сохранение уникальной языковой культуры имеет серьезное научное и социальное значение. В России и во всем мире уменьшается число носителей коренных языков и стоит задача сохранения языков малых народов. Из 250 языков России около 100 являются миноритарными языками¹. Стоит отметить, что даже языки, не относящиеся к меньшинствам, например, якутский (саха), могут находиться под угрозой исчезновения. Для большинства языков России, кроме русского, цифровых ресурсов (электронных словарей, лингвистических корпусов и текстов онлайн) либо нет, либо их относительно мало [Klyachko, 2019].

Для сохранения языкового богатства и последующего изучения языков создаются лингвистические корпусы текстов.

Электронные языковые корпусы в настоящее время являются одним из самых важных инструментов сохранения языков малых народов. Они представляют собой базу для решения миллионов задач прикладной лингвистики. Научное направление лингвистики, изучающее построение, анализ и использование языковых корпусов, называется *корпусной лингвистикой*.

Разработка открытых ресурсов, позволяющих хранить, систематизировать, осуществлять поиск информации по заданным параметрам, анализировать и интегрировать языковой материал в необходимый формат (печатный, электронный), представляется перспективным направлением в соответствии с требованиями современной науки.

1. Принципы работы в современных корпусах текстов

1.1. Корпусная лингвистика и виды корпусов

1.1.1 Основные понятия корпусной лингвистики

Корпусная лингвистика занимается разработкой общих принципов построения и использованием лингвистических корпусов. Существует много определений того, что составляет лингвистический корпус, одно из них гласит, что корпус представляет собой «совокупность текстов или частей текстов, по которым можно провести общий лингвистический анализ» [Meyer, 2002, xi].

Наличие корпуса для языка, объём корпуса и различные его характеристики крайне важны, поскольку «любое лингвистическое исследование в той или иной мере опирается на анализ языкового материала, языковых данных» [Баранов, 2001].

Стоит также отметить, что благодаря корпусам многократно повысились не только эффективность и скорость обработки языковых данных, но и достоверность результатов, поскольку на корпусе данных легче проверить гипотезу, предположения, выводы [Горина, 2014, с. 4].

Для анализа языкового явления и проведения лингвистического исследования необходим языковой материал в электронной форме, то есть *лингвистический корпус*,

¹ Миноритарный язык - это язык национального (этнического) меньшинства [ССТ: 129].

представляющий собой коллекцию текстов, специально отобранных, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска, которую называют *корпусным менеджером*.

Основные характеристики корпусов

Исходя из определений можно представить минимальные требования к корпусу текстов, которые выражаются в следующих параметрах [Копотев, 2014]:

- *репрезентативность* — свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового многообразия текстов;
- *сбалансированность* — параметр, определяющий насколько равномерно представлены тексты разных типов;
- *объем корпуса* — информация об общем объеме корпуса и о количестве извлеченных из текста примеров;
- *электронная форма хранения* обеспечивает быстрый поиск и извлечение материала;
- *разметка (аннотация)* — введенная вручную или автоматически лингвистическая или метатекстовая информация обо всех выбранных единицах текста (тексте, предложении, словосочетании, словоформе, морфеме).

Прагматическая ориентированность

Практика разработки и применения электронных корпусов текстов показала, что невозможно создать универсальный корпус, обеспечивающий решение всех задач. Задачи и цели любого исследования определяют тип корпуса, правила отбора текстов, а также способ и степень их обработки. Корпусы всегда создаются для решения определенной задачи или круга задач. Это определяет как наполнение корпуса текстами (например русская драма XIX в., тексты языка охотников), так и разметку корпуса [Захаров, Богданова, 2020].

Практика показывает также, что корпусная лингвистика оперирует как минимум тремя разными типами корпусов текстов:

1. *универсальными*, то есть отражающими все многообразие речевой деятельности;
2. *специфичными*, или представляющими бытование некоторого языкового или культурного явления в общественной речевой практике, например, корпус пословиц или корпус политических метафор в газетной речи;
3. *специальными*, то есть создаваемыми для решения специальной задачи, например, для обучения, для задач социолингвистики, для отладки систем машинного перевода.

1.1.2. Разметка корпусов

Понятие разметки

Разметка (аннотирование) корпуса заключается в приписывании текстам и их компонентам специальных атрибутов (тегов) [Захаров, Богданова, 2020]. Теги бывают двух типов:

- *лингвистические*, описывающие лексические, грамматические и прочие характеристики элементов текста;
- *экстравалингвистические* или метатекстовые (сведения об авторе или исполнителе и о тексте: название, год и место издания, жанр, тематика).

Лингвистическая разметка

Особое значение имеет лингвистическая разметка, т.е. любая аннотация, основанная на лингвистических характеристиках текста. Данные лингвистической разметки можно добавлять к текстовым элементам разных уровней. С точки зрения технологии, разметка может быть автоматической, ручной или автоматической с ручной правкой [Захаров, Богданова, 2020].

Среди лингвистических типов разметки можно выделить следующие типы:

- *Морфологическая разметка* отражает признаки части речи и грамматических категорий, свойственных данной части речи. Код привязан кциальному слову или группе слов. Это основной тип разметки, поскольку он используется для синтаксической и семантической разметки.
- *Синтаксическая разметка* описывает синтаксические связи между лексическими единицами и/или различные синтаксические конструкции (придаточное предложение, именное сказуемое и т. п.). Синтаксический код может быть закреплен за предложением или за синтаксическим отношением.
- *Семантическая разметка* обозначает семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, определяющие его значение. Семантический код закрепляется за словом или словосочетанием.
- *Анафорическая* разметка фиксирует референтные связи, например, местоименные.
- *Просодическая* разметка отражает ударение и интонацию.
- *Дискурсная* разметка служит для обозначения пауз, повторов, оговорок и т. п.

Экстравалингвистическая разметка

Дополнительная информация о текстах корпуса называется внешней или экстравалингвистической разметкой (метаразметкой). Экстравалингвистическая разметка может описывать как общие свойства текста, так и дополнительные технические данные:

- сведения об авторе;
- информацию о событии записи текста (характерно для диалектных текстов): год, место записи, собиратель (человек, записавший текст), информант (имя, год и место рождения, род занятий, национальность, пол);
- данные о публикации: автор, название, год, источник, страницы;

- типологические данные: художественный текст (например, рассказ, роман, басня) или нехудожественный (например, публицистический, научный, учебный, бытовой);
- тематическая (жанровая) классификация: историческая проза, детектив, драматургия и другие;
- технические данные: даты обработки, исполнители, кодировка, лицензия, даты загрузки;
- и другие свойства.

1.1.3. Типология корпусов

При разделении корпусов на классы рассматривают два подхода [Захаров, 2020]:

1. Относится ли корпус к языку в целом (возможно, к языку заданного периода) или корпус содержит тексты определённого стиля, жанра, отдельного писателя, социальной или возрастной группы и т.д.
2. Деление корпусов по типу разметки. В основном корпуса содержат только морфологическую разметку. Отдельно выделяют корпуса с синтаксической разметкой (treebanks).

Классификация корпусов по различным подходам представлена на Рис. 1.

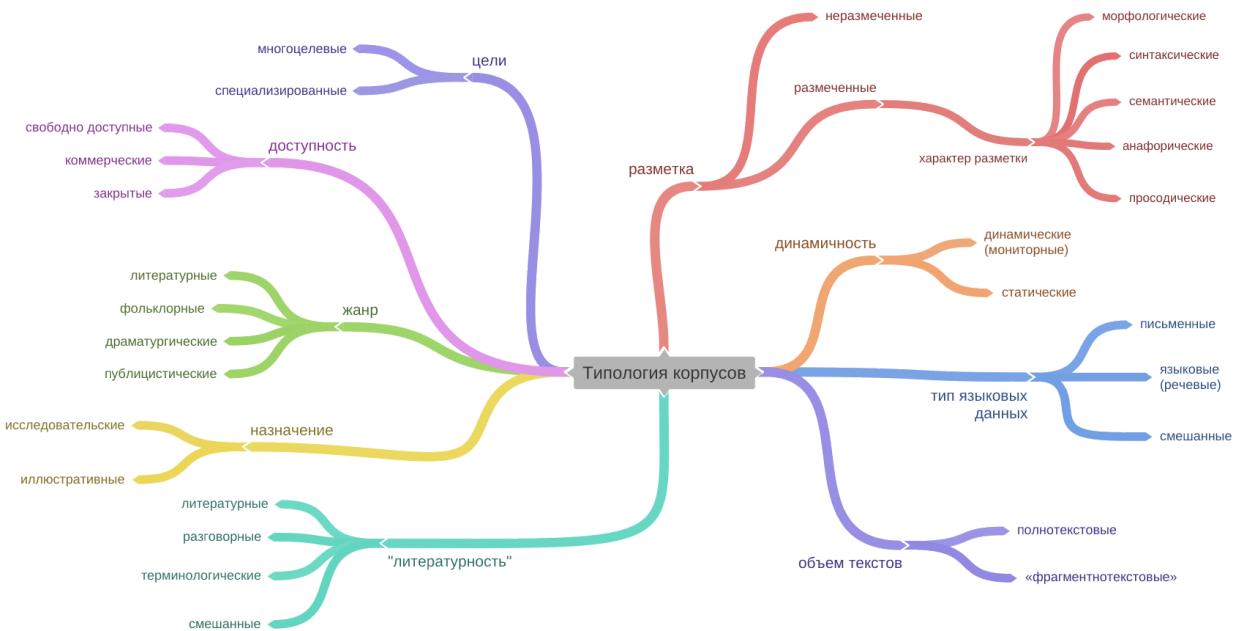


Рис. 1. Типы корпусов по классификации В.П. Захарова

1.1.4. Обзор существующих корпусов различных типов

В настоящее время для большинства основных языков мира существуют национальные общязыковые корпусы. Число лингвистических корпусов измеряется тысячами и постоянно растет. Два крупнейших специализированных каталога CLARIN² и ELRA³ содержат информацию о более чем 3000 корпусах. Список лингвистических ресурсов,

² Common Language Resources and Technology Infrastructure, <https://www.clarin.eu/resource-families>

³ European Language Resources Association, <http://www.elra.info/en/>

включающий и корпусы, можно найти на сайте Ассоциации компьютерной лингвистики⁴. Корпусы используются как для лингвистических исследований, так и в прикладных целях.

Зарубежные корпусы

*Британский национальный корпус*⁵ (кратко BNC) является最大的 (100 млн слов) корпусом британского варианта английского языка. Это *синхронический* корпус, он содержит актуальные для одного временного периода (конец XX века) примеры. Корпус разрабатывали два университета: Оксфордский и Ланкастерский в 1991-1994 годах.

Корпус содержит художественную и документальную прозу, тексты газет, периодических научных изданий и журналов, выпускаемых для различных возрастов, популярную научную фантастику, опубликованные и неопубликованные письма, школьные и университетские сочинения и др. 90% корпуса BNC составляют письменные тексты.

Устная речь в виде аудиозаписей с транскрипцией составляет 10% от объема корпуса BNC. Авторы BNC наняли 124 добровольца, мужчин и женщин разных возрастов, принадлежащих разным социальным группам, живущих в разных частях Великобритании, чтобы они записывали на плёнку свои обычные разговоры с друзьями, соседями, в семье, на работе. Была записана речь в самых разных контекстах — от разговоров на формальных деловых или правительственные встречах до радиошоу и телефонных разговоров. Было записано более 700 часов. К аудиозаписям были добавлены транскрипции (тексты с расшифровкой), тексты были аннотированы. Лаборатория фонетики Оксфордского университета создала звуковой подкорпус, для этого была выполнена оцифровка записей, каталогизация и анализ массива аудиокассет⁶, бывших недоступными в Звуковом архиве Британской библиотеки [Albert, 2012].

Одним из наиболее известных корпусов общего типа является *Чешский национальный корпус*⁷. Это синхронический морфологически размеченный корпус, представляющий современный чешский язык. Корпус был создан в конце 90-х годов в Карловом университете в Праге.

Корпусы русского языка

Первый русскоязычный корпус был создан в 1980-е годы в Университете Упсалы (Швеция). В Упсальский корпус отбирались специальные тексты с 1985 по 1989 годы и художественные тексты с 1960 по 1988 годы с целью представить современный литературный язык, поэтому в массиве нет образцов разговорной речи. Упсальский корпус русского языка состоит из 600 текстов, его объем составляет 1 млн словоупотреблений. Корпус имеет морфологическую разметку.

Первый общедоступный, представительный и размеченный корпус русского языка, с которым могли бы работать лингвисты, стал доступен в 2004 году [Плунгян, Сичинава, 2004]. *Национальный корпус русского языка (НКРЯ)* содержит тексты на русском языке, корпус является диахроническим и содержит тексты, созданные на протяжении нескольких веков, включая современные тексты. Авторы НКРЯ использовали понятие

⁴ Список ресурсов с разбивкой по языкам, https://aclweb.org/aclwiki>List_of_resources_by_language

⁵ British National Corpus, BNC, <http://sara.natcorp.ox.ac.uk/> или <https://www.english-corpora.org/bnc/>

⁶ См. примеры аудиозаписей BNC <http://www.phon.ox.ac.uk/SpokenBNCcontents>

⁷ Český národní korpus, <http://korpus.cz/>

«подкорпуса» для разделения текстов на следующие большие группы: основной, поэтический, диалектный, параллельные тексты, газетный, церковнославянский, исторический, синтаксический, акцентологический, мультимедийный и обучающий. Как все современные лингвистические корпуса, он доступен онлайн⁸. Отметим такие важные составляющие НКРЯ, как богатые инструменты поиска и статистические сведения о корпусе, делящиеся на статическую и динамическую статистику [Гришина, Плунгян, 2003, с. 335]. Сложности создания большого корпуса русского языка при одновременном сосуществовании разных классификаций языковых уровней, ведущие к неизбежным трудностям при автоматическом аннотировании на основе этих классификаций, описаны в работе «Национальный корпус русского языка» [Копотев, Янда, 2006].

Ряд особенностей интерфейса НКРЯ и поисковых возможностей описан в работе «Национальный корпус русского языка» [Копотев, Янда, 2006, с. 150].

Table 1. Примеры корпусов русского языка (объем словаупотребления в млн.)

Корпусы	Адрес	Объем
Упсальский и Тюбингенский корпуса	http://www.sfb441.uni-tuebingen.de/b1/rus/korpora.html	1
Машинный фонд русского языка	http://cfrl.ruslang.ru/	100
Национальный корпус русского языка (НКРЯ)	https://ruscorpora.ru/new/index.html	1 000
Хельсинкский аннотированный корпус русских текстов (ХАНКО)	http://h248.it.helsinki.fi/hanco/index.html	0,1
Открытый корпус русского языка	http://opencorpora.org/	1,5
Генеральный Интернет-корпус Русского Языка (ГИКРЯ)	http://www.webcorpora.ru/	20 000
Корпус русского литературного языка	https://narusco.ru/	1
Корпус русских учебных текстов (КРУТ)	http://web-corpora.net/learner_corpus	3,1
Русский учебный корпус	http://web-corpora.net/RLC	2

Корпусы уральских языков

В области уралистики известны, прежде всего, электронные ресурсы наиболее крупных финно-угорских языков, такие как Языковой банк Финляндии, Венгерский национальный корпус, Текстовый корпус Института эстонского языка. В России ведется работа по созданию и наполнению корпусов мордовского, марийского, коми, удмуртского, мансийского языков. На сайте проекта “Малые языки Сибири” ([Siberian Lang | Малые языки Сибири](https://siberianlang.ru))

⁸ См. <https://ruscorpora.ru>

[языки Сибири: наше культурное наследие \(msu.ru\)](#)) представлены также аннотированные тексты с видео на селькупском самодийском языке.

Большой теоретический и практический интерес представляет собой корпус уральских языков Поволжья, разработанный Т. Архангельским. Для построения веб-корпуса пяти уральских языков (коми-зырянского, лугового марийского, мокшанского, удмуртского и эрзянского) обрабатываются тексты сети Интернет, корпус содержит только автоматическую разметку. В него включены современная художественная литературу, научные статьи, переводы Библии, статьи Википедии, официальные тексты и публичные записи в социальных сетях [Arkhangelskiy, 2020, с. 58-61].

Table 2. Примеры корпусов языков России (объем словоупотребления в млн)

Корпусы	Адрес	Объем
Башкирский поэтический корпус	http://web-corpora.net/bashcorpus/search/index.php?interface_language=ru	1,8
Калмыцкий корпус	http://web-corpora.net/KalmykCorpus/search/index.php?interface_language=ru	0,8
Бурятский корпус	http://web-corpora.net/BuryatCorpus/search/index.php?interface_language=ru	2,2
Татарский национальный корпус «Туган тел»	http://web-corpora.net/TatarCorpus/search/index.php?interface_language=ru	26

Table 3. Корпусы уральских языков мира (объем словоупотребления в млн.)

Корпусы	Адрес	Объем
Языковой банк Финляндии	https://sanat.csc.fi	
Венгерский национальный корпус	http://mnsz.nytud.hu/index_hun.html	188
Текстовый корпус Института эстонского языка	https://portaal.eki.ee/corpus	10,4
Корпус коми-зырянского языка	http://komi-zyrian.web-corpora.net/	1,76
Корпус лугового марийского языка	http://meadow-mari.web-corpora.net/	5,53
Корпус мокшанского языка	http://moksha.web-corpora.net/	1,74
Корпус эрзянского языка	http://erzya.web-corpora.net/	2,3
Корпус удмуртского языка	http://udmurt.web-corpora.net/	9,57

Национальный корпус удмуртского языка	http://udmcorpus.udman.ru/home	
Аннотированный корпус мансийских текстов	http://mansi.pro/corpus/	0,05

1.2. Описание корпуса ВепКар

Открытый корпус вепсского и карельского языков как результат многолетней междисциплинарной работы лингвистов и программистов Карельского научного центра РАН является уникальной источниковой базой для новых исследований.

1.2.1. История создания

Электронный ресурс ведет свою историю с 2009 года, когда под руководством Зайцевой Н. Г. был создан «Корпус вепсского языка»⁹ [Зайцева, 2012]. Стоит задача возрождения вепсского и карельского языка и культуры. Для этого нужны большие электронные языковые ресурсы, которые позволяют исследователям создавать и развивать, например, правила орфографии, вводить в научный и общественный оборот значительное количество материалов, востребованных авторами учебников по языку и истории родного края, учителями, писателями. В связи с этим, в целях сохранения, развития и популяризации не только вепсского, но и карельского языка в 2016 году сотрудники Института языка, литературы и истории и Института прикладных математических исследований КарНЦ РАН приступили к созданию многоязычного корпуса. В 2016 году на базе «Вепсского корпуса» была создан новый интернет-ресурс «Открытый корпус вепсского и карельского языков» (ВепКар)¹⁰. В корпус помимо вепсского вошли три карельских подкорпуса: собственно карельский, ливвицкий и людиковский. Объединённая лингвистическая платформа получила название «Открытый корпус вепсского и карельского языков» (ВепКар). В 2022 году ВепКар «заговорил»: был разработан речевой модуль, позволяющий услышать тексты или леммы, озвученные носителями языка. Редакторами ВепКара осуществляется постепенное наполнение звукового корпуса аудиотекстами с их расшифровками.

1.2.2. Архитектура ВепКар

Открытый корпус вепсского и карельского языков — это многоязычный полнотекстовый лингвистический корпус, содержащий морфологическую, семантическую и метатекстовую разметку.

Корпус включает в себя хранящиеся в базе данных тексты и словари на вепсском и карельском языках, часть из которых сопровождена аудиофайлами, а также компьютерную программу (корпусный менеджер), обеспечивающую поиск и обработку текстов. Корпусный менеджер написан на языке программирования PHP в системе разработки веб-сайтов Laravel. Данные хранятся в базе данных MySQL. Словари и тексты корпуса

⁹ См. <http://vepsian.krc.karelia.ru/about/>

¹⁰ См. <http://dictorus.krc.karelia.ru>

вместе с поисковой системой доступны онлайн. Авторы проекта уделяют внимание популяризации корпуса ВепКар с помощью сайтов YouTube и Википедия.

Особенностью корпуса ВепКар является тесная взаимосвязь словарей и текстов. Многофункциональные словари вепсского и карельского языков содержат толкование русском и частично английском языках, перевод, диалектные пометы, семантические отношения (синонимы, антонимы и др.), примеры словоупотреблений со ссылкой на тексты, а также полные словоизменительные парадигмы. Все тексты автоматически размечиваются и от слов в тексте идут ссылки на соответствующие значения в словарных статьях.

Разработчики добавляют в корпусный менеджер новые полезные функции, призванные облегчить работу редакторов. Например, были сформулированы и запрограммированы правила именного и глагольного словоизменения [Krizhanovskaya et al., 2022] для всех диалектов вепсского языка и его младописьменного варианта, а также для ливвиковского, севернокарельского и тверского новописьменных вариантов карельского языка. Благодаря этому в системе ВепКар в полуавтоматическом режиме было сгенерировано 2.1 млн словоформ. Кроме семантической разметки, представленной в корпусе (2.1 млн связей между словами из текста и значениями лемм в словаре), была добавлена грамматическая разметка, позволившая автоматически установить 1.1 млн связей между словами из текста и грамматическими характеристиками словоформ из словаря.

Многоязычный корпус ВепКар делится на подкорпуса по языкам и наречиям, также есть стилевая и жанровая классификация текстов. В корпусе организована развитая система поиска с фильтрацией текстов по языковой, стилистической и диалектной принадлежности, по информанту, собирателю или автору, году записи или году публикации. Поиск лемм возможен по диалектам, частям речи, грамматическим признакам и даже по лексико-семантическим категориям.

Корпус ВепКар можно рассматривать как электронную библиотеку, предоставляющую пользователям доступ к полным текстам документов. Материалами для ВепКар могут служить тексты только с открытой лицензией¹¹.

¹¹ См. страницу корпуса ВепКар с разрешениями авторов текстов:
<http://dictorus.krc.karelia.ru/ru/page/permission>



Рис. 2. Архитектура корпуса VenKar. Связь словаря и корпуса текстов. Данные на ноябрь 2022 г.

Таким образом, интернет-ресурс ВенКар объединяет в себе три составляющие: размеченный корпус, многофункциональный словарь и библиотеку.

Подробнее об архитектуре (Рис. 2) и возможностях ВенКар читайте в работе [Бойко и др., 2021], в докладе на конференции [Крижановская, Крижановский, 2020] и смотрите видеозапись на YouTube¹².

1.2.3. Страница с текстом в корпусе ВенКар

Страница с текстом (Рис. 3) в зависимости от информации о тексте (то есть от метаданных) содержит:

- информацию об авторе;
- заголовок;
- название подкорпуса / коллекции;
- языковую принадлежность;
- диалектную принадлежность;
- для диалектных текстов сведения о месте и времени записи, информанте, собирателе;
- сведения о публикации;
- информацию о хранении исходного материала в Научном архиве Карельского научного центра РАН или Фонограммархиве Института языка, литературы и истории КарНЦ;
- аудиофайл;
- размеченный текст;
- перевод текста (в основном на русский язык):

¹² См. видеозапись «Архитектура корпуса ВенКар и диалектные особенности» <https://www.youtube.com/watch?v=cUpqM97LXGs>

- автор перевода
- заголовок
- текст, выровненный по предложениям с исходным.

При наведении мышки на предложение исходного текста, жёлтым цветом выделяется соответствующее ему предложение-перевод. При наведении мышки на предложение-перевод, голубым цветом выделяется соответствующее ему предложение в исходном тексте.

Если кликнуть на размечено слово в вепсском/карельском тексте, откроется окно с леммой, значением и грамматическими свойствами. Каждое автоматически размечено слово проверяется экспертом. Зеленый цвет означает, что слово проверено, синий и красный — не проверено. Если система нашла одно соответствие в словаре, то слово помечено синим, если несколько — красным. Чёрный цвет означает, что в словаре пока нет соответствия для слова в тексте.

"Lähtiimmö müä siäneh..."

подкорпус: диалектные тексты

информант(ы): Куршиева Анна Николаевна, 1910

Макарова Фекла Алексеевна, 1902

место записи: Самбатукса (Sammatus), Олонецкий район, Республика Карелия, г. записи: 1963

записали: Вайзинен Т. И., Макаров Григорий Николаевич

источник: Г.Н. Макаров, В.Д. Рягоев, Образцы карельской речи. Говоры ливвиковского диалекта карельского языка, 1969, с. 61-63

ф/архив ИЯЛИ КарНЦ РАН: №221/3

"Lähtiimmö müä siäneh..."

Карельский: ливвиковское наречие
Коткозерский

[Anni]: Lähtiimmö müä siäneh jäl'les voinan loppuu.

Da müä segoimmo sinne, vihmupäivü oli, müä emmo kodih voinnut puwtua.

Puwtuimmo sinne majakan alle, Leningradskoih mežah, dai müä üävüimmö, magaimmo üän.

N'veeskü se uinoi muate, virs muata VERB (спать)

a minä üän kaiken puhuin h indикатив, имперфект, 1 л., мн. ч., полож. ф.

Viižitoštu vuattu oli majakal luajittuw, nii'roin [halgoloih] üän kaiken tuldu puhuin.

Sit huandesčura ku rodih, ruvettih siä hukat ulvomah, mečü oli ül'en suwri.

Lähtiimmö müä (huandes tuli) poikki suas.

«Отправились мы за рыжиками...»

Русский

[Анна]: Отправились мы за рыжиками после того, как война кончилась.

Мы там заблудились, дождливый день был, мы не смогли попасть домой.

Мы попали туда, к маяку ('под маяк'), где ленинградская межа, мы там и заночевали, провели.

Скотка-то уснула, корзиночку положила в овье, а я всю ночь поддерживала огонь ('дула в огонь') в гнилом валежнике.

Пятнадцать лет прошло [с тех пор], как маяк тот в лесу был поставлен, дрова из этого маяка были, и я всю ночь поддерживала костёр.

Стало за полночь, начали там выть волки,

Рис. 3. Фрагмент ливвиковского диалектного текста "Lähtiimmö müä siäneh..."¹³

¹³ Полный текст см. <http://dictorus.krc.karelia.ru/ru/corpus/text/1514>

1.2.4. Словарная статья в корпусе ВенКар

На сайте корпуса ВенКар организован словарь, по которому можно производить поиск (Рис. 33). Щелкнув на найденное слово, можно перейти к словарной статье. В словаре ВенКар словарные статьи называются “леммами” (лемма - начальная форма слова). Если слово из одного языка (наречия) употребляется в разных частях речи, то для каждой части речи имеется отдельная словарная статья.

После автоматической разметки каждому значению слова привязаны примеры из полнотекстового корпуса ВенКар. После ручной разметки (экспертной проверки) эти примеры помечены звёздочками с оценками (“лучший”, “хороший” и т.д.) (Рис. 4). Если у примера нет оценки, то этот пример возможно не проверен экспертом, а значит может относиться к другому значению или даже другой словарной статье.

ahven

язык: карельский: ливвицкое наречие

часть речи: существительное



1 значение

понятие: [окунь](#)

- **русский:** окунь
- **английский:** European perch, redfin perch, perch



перевод

венесский: [ahven; ahr'](#)

карельский: людиковское наречие: [ahven](#)

карельский: собственно карельское наречие: [ahven](#)

диалекты употребления: Ведлозерский, Видлицкий, Кондушский, Неккульский, Сямозерский, Тулмозерский

Примеры (17)

★ лучший ★ отличный ★ хороший ★ плохой

1. ★ Ahvenet ollah ei ylen suuret, vie pojat.
Окунь еще не крупные, еще малки. (["Minä sanon, kui müö provodimmo kanikuluw"](#))

2. ★ Ahven kudou hätken, tedri kiimuičou hätken.
Окунь нерестится долго, тетерев токует долго. ([Kalua pyvvimmö ijän kaiken](#))

3. ★ Vie sanotah: ahven ahalav kudou, a särgi siäl lämmäl.
Еще говорят: окунь в холодную сухую погоду нерестится, а плотва в теплую. ([Kalua pyvvimmö ijän kaiken](#))

4. ★ Sit nowzow, tämä joi, ahven: sit ahvenet kuvotetah kalakkahat järvet.
Потом поднимается, так сказать, окунь, в рыбных озерах во время нереста ловят окуня. ([Minä olen rodinuh Čil'miel'e](#))

5. ★ Erilastu sanondua on kalastajih näh, što vaiku pattii ei kergie ahvenen kuduh da teedrin kiimah.
Различные выражения есть о рыбаках, что только ленивый не успевает на окуневый нерест и тетеревиный ток. ([Kalua pyvvimmö ijän kaiken](#))

Рис. 4. Словарная статья для ливвицкого существительного *ahven* с тремя «лучшими», одним «отличным» и одним «хорошим», цитатами из корпуса ВенКар, проверенными лингвистами

sanuo

язык: карельский: ливвиковское наречие

часть речи: глагол

фонетические варианты: [sanua](#) (Сямозерский, Видлицкий, Неккульский, Тулмозерский); [sania](#) (Кондушский)

фразеологизмы

[sanuo silmih](#) - сказать в лицо (букв. сказать в глаза)

[sanuo tervehyöt](#) - передать привет

[sanuo valmistelemattah](#) - сказать экспромтом

[sanuo vastah](#) - перечить (букв. говорить наперекор)

1 значение

понятие: [сказать](#)

- **русский:** говорить, сказать

перевод

вепсский: [sanuda](#)

карельский: людиковское наречие:

[sanuo](#); [sanuda](#); [sanoda](#)

карельский: собственно карельское наречие:

[šanua](#); [sanuo](#); [šanuo](#)

диалекты употребления: Ведлозерский

Примеры (3043)

★ лучший ★ отличный ★ хороший ★ плохой

1. ★ I kummoksie ei midä ole, hüö ei voija kohti [sanuo](#), čto heil muudu ni midä ei pie, ku vai ottua valdu ruadajoil da krest'janoil käzis iäre.

И неудивительно, они не могут прямо сказать нам, что им ничего другого не надо, кроме как отобрать власть из рук рабочих и крестьян. ([Ruadorahyahale kaiken muan](#))

2. ★ "Kai ruado rahvahale, ni midä ruadamattomile – vierahan ruavon süöjile" – vot kui [sanotah](#) bol'seviekat.

"Все должно быть предоставлено трудащемуся народу, ничего не давать тем, кто не работает, кто сосёт чужую кровь" – вот как говорят большевики. ([Ruadorahyahale kaiken muan](#))

3. ★ "Towta, – ma [sanon](#), – minul oli brihaččuine kodjh d'iänüh, ga tiiätgo kus on?"

"Тётя, – говорю я, – остался у меня мальчик дома, не знаешь ли, где он?" ("[Meil omii lapsii iellon](#)")

[еще примеры >>](#)

2 значение

Примеры (2959)

★ лучший ★ отличный ★ хороший ★ плохой

1. ★ Minä [sanon](#), kui műö provodimmo kanikuluw.

Я расскажу, как мы проводили каникулы. ("[Minä sanon, kui műö provodimmo kanikuluw](#)")

[еще примеры >>](#)

3 значение

Примеры (2962)

★ лучший ★ отличный ★ хороший ★ плохой

1. ★ Nu műö käväymmö műömäh talolui müöte: sukkia, pičcie (немме kruwživot – suomekse pičikse [sanotah](#)), rihmia.

Ну, мы ходили торговать по домам: чулки, кружева (по-фински кружева «пимтси» называю), нитки. ([Torrunkel'e Suomes](#))

Рис. 5. Фрагмент словарной статьи ливвиковского глагола 'sanuo' с тремя значениями¹⁴

Словарная статья (Рис. 5) в зависимости от степени её проработки содержит:

- **язык;**
- **грамматическую информацию**
 - часть речи,
 - возвратность, безличность, переходность у глагола,
 - одушевленность у существительного,
 - степень сравнения у прилагательного и наречия,
 - тип наречия,
 - разряд местоимения и т.д.;
- **фонетические варианты в других диалектах;**
- **фразеологизмы или устойчивые словосочетания с этим словом;**
- **значения с толкованиями на русском, английском, вепсском, наречиях карельского и финском языках;**
- **переводы значений на вышеперечисленные языки;**
- **семантические отношения с другими леммами в данном языке: синонимы, антонимы, гиперонимы, гипонимы, холонимы, миранимы и т.д.;**
- **диалекты, в которых употребляется данное значение;**

¹⁴ Полную статью см. <http://dictorus.krc.karelia.ru/ru/dict/lemma/15190>

- примеры из корпуса с оценками эксперта;
- список словоформ (Рис. 6) со ссылками на тексты корпуса.

словоформы (150)

No	грамматические признаки	Новописьменный ливвиковский (149)	Коткозерский (1)
Индикатив, презенс, положительные формы			
1.	1 л., ед. ч.	<u>sanon</u>	
2.	2 л., ед. ч.	<u>sanot</u>	
3.	3 л., ед. ч.	<u>sanou</u>	<u>sanow</u>
4.	1 л., мн. ч.	<u>sanommo</u>	
5.	2 л., мн. ч.	<u>sanotto</u>	
6.	3 л., мн. ч.	<u>sanotah</u>	
7.	ед. ч., коннегатив	<u>sano</u>	
8.	мн. ч., коннегатив	<u>sanota</u>	

Рис. 6. Фрагмент словарной статьи ливвиковского глагола ‘sanuo’ с тремя значениями¹⁵

1.2.5. Открытость данных

ВепКар включает в себя только те тексты, на которые получено разрешение авторов на публикацию под открытой лицензией CC-BY 4.0. При такой политике не каждый текст можно включить в корпус, однако наличие разрешения даёт возможность публиковать тексты (Рис. 3) и свободно передавать третьим лицам для научных исследований и разработки коммерческих приложений.

2. Методические рекомендации по использованию интернет-ресурса ВепКар

2.1. Применение интернет-ресурса ВепКар

Лингвистические корпусы используются студентами и преподавателями разных дисциплин, связанных с анализом употреблений слов, с поиском типичных или необычных словоформ и оттенков значений.

С помощью ВепКар преподаватели могут

- подбирать примеры и формировать задания по грамматике, фразеологии, фольклору, литературе и разговорной практике по карельскому и вепсскому языкам;
- использовать текстовые материалы корпуса для проведения курса домашнего чтения;
- привлекать диалектные тексты в курсе диалектологии в целях выявления фонетических и грамматических особенностей карельской и вепсской диалектной речи;
- использовать материалы корпуса в процессе разработки новых учебных пособий;
- привлекать материалы параллельных словарей для обучения основам перевода.

¹⁵ Полную статью см. <http://dictorus.krc.karelia.ru/ru/dict/lemma/15190>

С помощью ВепКар студенты могут

- находить полные словоизменительные парадигмы изменяемых слов, т.е. использовать материалы словарного модуля корпуса в качестве электронного грамматико-орфографического словаря карельского и вепсского языков;
- использовать словарь корпуса в качестве переводного словаря карельского и вепсского языков;
- определять управление (глагольное управление, употребление послелогов и предлогов) с помощью контекстных примеров из корпуса текстов;
- выявлять фонетические и грамматические особенности карельской и вепсской диалектной речи на основе диалектных текстов;
- подбирать контекстные примеры в качестве источников базы для курсовых и дипломных работ.

С помощью ВепКар исследователи могут

- находить примеры, демонстрирующие различные фонетические явления, словоизменительные и словообразовательные особенности карельской и вепсской диалектной речи (расширенный поиск по заданным параметрам);
- получать выборки заданных словоформ из размеченных текстов с целью анализа грамматических категорий;
- анализировать и сравнивать тексты на новописьменных вариантах языка с целью выявления грамматических правил, нуждающихся в редактировании или доработке;
- проводить лингвостатистический анализ при помощи частотных словарей;
- отслеживать динамику изменения лексического состава анализируемых языков за вековой период;
- анализировать язык определенных жанров или конкретных авторов;
- использовать материалы корпуса в процессе разработки новых словарей.

С помощью ВепКар журналисты и писатели имеют возможность

- пользоваться большими объемами разнородных и разножанровых текстов для детального знакомства с творчеством коллег, разработки тем, образов и др.;
- уточнять и подбирать подходящие для текстов лексемы, словоформы и фразеологизмы, пользуясь линейкой словарей ВепКар (например, при подборе синонимов);
- привлекать данные корпуса для создания текстов и произведений на диалектах вепсского и карельского языков;
- использовать платформу ВепКар для популяризации своего творчества.

2.2. Рекомендации по использованию интернет-ресурса ВепКар

2.2.1. Общие вопросы

Шаг 1. Как найти сайт Открытого корпуса вепсского и карельского языков (ВепКар) в Интернете?

Наберите адрес <http://dictorpus.krc.karelia.ru> и Вы попадете на стартовую страницу сайта корпуса ВепКар (Рис. 7).

форма авторизации

переключение интерфейса на другой язык

верхнее меню

видео с образцами речи

нижнее меню

ссылка на мобильное приложение

О ПРОЕКТЕ	КОРПУС	СЛОВАРЬ	СПРАВОЧНИКИ
Главная	Тексты	Леммы	Языки
Участники проекта	Речевой корпус	Словоформы	Диалекты
Источники	Лексико-грамматический поиск	Семантические отношения	Подкорпусы
Публикации	Фразеологизмы	Новые леммы	Жанровые
О вепсах	Новые тексты	Последние измененные леммы	Географические
О карелах	Измененные тексты	Обратный словарь лемм	Информанты
Статистика	Выборки	Выборки	Авторы
Гранты	Частотные словари	Понятия	Собиратели
		Аудиозаписи лемм	Грамматические

 © VepKar. Открытый корпус вепсского и карельского языков

Сайт создан сотрудниками Карельского научного центра Российской академии наук

Рис. 7. Стартовая страница сайта Открытого корпуса вепсского и карельского языков (BenKar)

Шаг 2. Как переключить интерфейс сайта ВепКар на русский язык?

На главной странице сайта найдите иконку с флагом России и ссылку «*Русский*» и щелкните по ссылке мышкой (Рис. 8):



Рис. 8. Переключение на русский язык BenKap.

Шаг 3. Обязательно ли нужно регистрироваться на сайте ВепКар?

Регистрация не обязательна. Пока она необходима только редакторам ВепКар. Обычные пользователи на настоящий момент не получают никаких привилегий при регистрации. В будущем планируются дополнительные функции для зарегистрированных пользователей, например, сохранение типичных запросов, подобранных данных, организация «тяжёлых» запросов.

Шаг 4. Как ввести особые графические символы карельского и вепсского языков, если их нет на клавиатуре?

В тех полях, где могут понадобиться отличающиеся от английской раскладки буквы вепсского или карельского алфавита (äöüčšž’), справа размещена иконка «специальные символы»:  . Щелкните по этой иконке, иконка исчезнет, сверху откроется панель с буквами-кнопками (Рис. 9). Щелкните по нужному символу, символ вставится в поле на месте курсора.

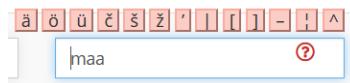


Рис. 9. Открытая панель «специальные символы»

Стоит особо отметить, что в процессе поиска символ ѹ приравнен к символу у, w - к u / y / ѹ, а знак палатализации ' не учитывается.

Шаг 5. Как использовать простые шаблоны в текстовых полях?

В текстовых полях, если Вам нужен неточный поиск (по фрагменту слова), используйте процент % для замены любого количества символов и подчеркивание _ для замены одного символа.

Например, по шаблону %ta будут найдены все строки, оканчивающиеся на 'ta'. А по шаблону %änd_ будут найдены все строки, оканчивающиеся на 'änd' и один любой символ. Например, будут найдены 'eländy' или 'vierändü'.

Шаг 6. Как использовать специализированные шаблоны в текстовых полях?

В некоторых текстовых полях для построения более сложных запросов можно использовать специализированные шаблоны с регулярными выражениями (Regex и собственные). В table 4 приведены шаблоны и примеры их использования для поиска строк .

Table 4. Специализированные шаблоны для поиска текста

шаблон	описание	пример	пояснение
^	начало строки	^с	все слова, начинающиеся на букву с (Поиск лемм по словоформам)
\$	конец строки	in\$	все слова, оканчивающиеся на in (в словарной форме имен с.к. наречия: Поиск лемм по словоформам)
.	любой один символ	h.\$	все глаголы на ha/hä в ливвиковском новописьменном варианте (Поиск лемм по словоформам)
[...]	любой один символ из перечисленных в квадратных скобках	läht[öe]m[äy]	все фонетические варианты понятия ‘нетель’ в словарной форме (Поиск лемм по словоформам)
[a-z]	любая латинская буква от a до z	[a-zA-Z]	любая латинская буква вне зависимости от регистра, например, нужно исключить символ палатализации
[^...]	любой один символ, кроме перечисленных в квадратных скобках	[^0-9]	любой символ, кроме цифры
?	идущий перед знаком вопроса символ может встретиться, а может и нет	^ikkuna?*	a на конце может быть, а может не быть (Поиск лемм по словоформам)
*	ноль или более символов, идущих перед звёздочкой	ab*c	abc ИЛИ ac ИЛИ abbc
+	один или более символов, идущих перед плюсом	ab+c	abc ИЛИ abbc ИЛИ abbbc
{n}	n символов, стоящих перед скобками	b{3}	bbb
{m,n}	от m до n символов, стоящих перед скобками	a{1,3}	a ИЛИ aa ИЛИ aaa
{m,}	предыдущий символ может встретиться m и более раз	^pert{1,}i?*	pert ИЛИ pertt ИЛИ perti ИЛИ pertti (Поиск лемм по словоформам)

(...)	круглые скобки задают группировку символов	$(abc)\{1,3\}$	abc ИЛИ $ab\overset{c}{a}bc$ ИЛИ $a\overset{b}{c}a\overset{b}{c}abc$
p1 p2	p1 или p2	$\overset{\wedge}{rein}\overset{\mid}{rien}\overset{\wedge}{rejen}\overset{\mid}{regen}$	все варианты слабоступенной основы существительного regi (Поиск лемм по словоформам)
V	любая гласная буква	$ttVV\$$	все собственно карельские инфинитивы одноосновных глаголов, содержащих в основе чередование tt : t (Поиск лемм по словоформам)
C	любая согласная буква	$\overset{\wedge}{kaCrV\$}$	фонетические варианты с глухим и звонким согласным для kagra / kagru / kakra (Поиск лемм по словоформам)

2.2.2. Поиск в корпусе ВепКар

Шаг 7. Как осуществлять поиск текстов в корпусе ВепКар?

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «КОРПУС» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Тексты» (Рис. 10).

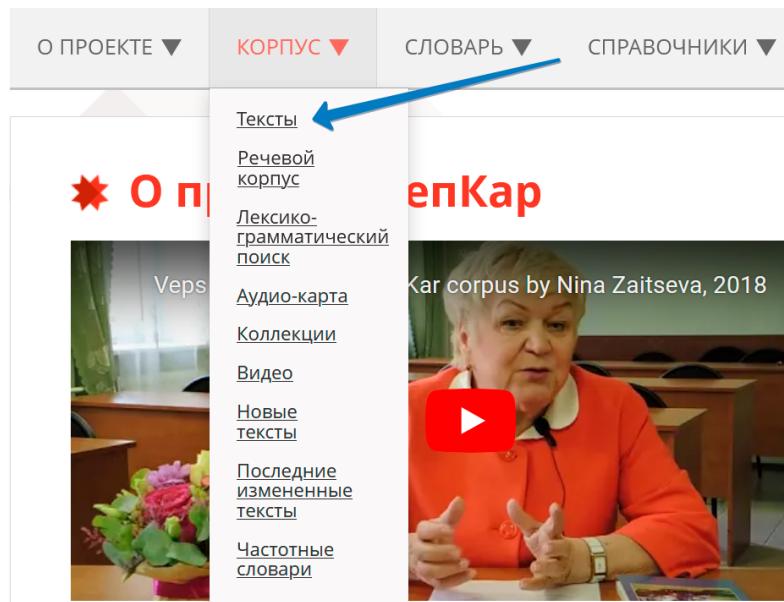


Рис. 10. Поиск в корпусе BenKar

Шаг 8. Как найти текст на вепском / карельском языке?

- Выполните шаг 7.
- В поле «Язык» выберите нужный язык/наречие.

- Нажмите кнопку **ПОКАЗАТЬ**.

Загрузится таблица с текстами (Рис. 11).

- Чтобы перейти на страницу с текстом, щелкните на заголовок текста.

★ Тексты

[Лексико-грамматический поиск](#) | Создать новый | [? Помощь](#)



Расширенный поиск ↓

Язык		Подкорпус				
No	Язык	Диалект	Подкорпус	Жанр	Заголовок	Перевод
1	карельский: людиковское наречие	Новописьменный людиковский	публицистические тексты		AD'VOIŠ	В ГОСТЯХ
2	карельский: людиковское наречие	Южнолюдиковский (святозерский)	диалектные тексты	бытовой рассказ	Agđan St'opan	Степан Агдян
3	карельский: людиковское наречие	Михайловский	диалектные тексты	бытовой рассказ	Äij on dielot kalan suada	
4	карельский: людиковское наречие	Южнолюдиковский (святозерский)	диалектные тексты	бытовой рассказ	Akimat	Семейство Акиман
5	карельский: людиковское наречие	Новописьменный людиковский	публицистические тексты		Paušin Šan'uu. Aleksandr Barancev on meiden kuuluž heimolaine	

Рис. 11. Поиск в корпусе BenKar. Пример найденных людиковских текстов

Шаг 9. Как осуществлять расширенный поиск по текстам ВенКар?

- Выполните шаг 7.
- Щелкните на ссылку «расширенный поиск» справа вверху над поисковой формой (Рис. 11).

В форме откроются дополнительные поля (Рис. 12). Под поисковой формой выводится количество найденных текстов и таблица с данными. Если сформирован поиск по языку, подкорпусу, диалекту или жанру, то соответствующие колонки в таблице не дублируются.

Язык	Диалект	Подкорпус	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Жанр	Сюжет	Тема	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Область, республика записи	Район записи	Населенный пункт записи	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Область, республика рожд. информанта	Район рожд. информанта	Населенный пункт рожд. информанта	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Информант	Собиратель	Автор текста или перевода	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Заголовок ? ä	Год (с) ?	Год (по) ?	Источник ? ä
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Фрагмент текста ? ä	<input type="checkbox"/> с аудио-записями		
<input type="text"/>	по <input type="text"/> записей ПОКАЗАТЬ		

Простой поиск ↑

Найдено 4 129 записей.

No	Язык	Диалект	Подкорпус	Жанр	Заголовок	Перевод
1	карельский: собственно карельское наречие	Валдайский	диалектные тексты		Mar'joa mōimmä imen'jah	Пунжина Александра Васильевна. Ягоды продавали в имение

Рис. 12. Расширенный поиск текстов в корпусе BenKar

Шаг 10. Как найти озвученный текст на вепсском / карельском языке?

- Выполните шаг 9.
- В поле «Выберите язык» выберите нужный язык/наречие.
- Кликните в поле «с аудио-записями».
- Нажмите кнопку ПОКАЗАТЬ.

Загрузится таблица с текстами (Рис. 13).

Язык					<input type="text"/> карельский: ливвиковское наречие (959)	<input checked="" type="checkbox"/> с аудио-записями	по <input type="text"/> записей	ПОКАЗАТЬ																														
<p>Найдено 15 записей.</p> <table border="1"> <thead> <tr> <th>No</th> <th>Диалект</th> <th>Подкорпус</th> <th>Жанр</th> <th>Заголовок</th> <th>Перевод</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Рыпушкальский</td> <td>диалектные тексты</td> <td></td> <td>Kaikin paistih hierus livvikse</td> <td>Мичурова Надежда. Все говорили в деревне на ливвиковском</td> </tr> <tr> <td>2</td> <td>Неккульский</td> <td>диалектные тексты</td> <td></td> <td>Joga talois oli lehmy</td> <td>Мичурова Надежда. В каждом доме была корова</td> </tr> <tr> <td>3</td> <td>Котозерский</td> <td>диалектные тексты</td> <td></td> <td>Vastaimmo meijän saldottoi</td> <td>Мичурова Надежда. Встречали наших солдат</td> </tr> <tr> <td>4</td> <td>Рыпушкальский</td> <td>диалектные тексты</td> <td>бытовой рассказ</td> <td>Poimičuloin da viršilöin azundu</td> <td>Мичурова Надежда. Изготовление маленьких и больших корзин</td> </tr> </tbody> </table>									No	Диалект	Подкорпус	Жанр	Заголовок	Перевод	1	Рыпушкальский	диалектные тексты		Kaikin paistih hierus livvikse	Мичурова Надежда. Все говорили в деревне на ливвиковском	2	Неккульский	диалектные тексты		Joga talois oli lehmy	Мичурова Надежда. В каждом доме была корова	3	Котозерский	диалектные тексты		Vastaimmo meijän saldottoi	Мичурова Надежда. Встречали наших солдат	4	Рыпушкальский	диалектные тексты	бытовой рассказ	Poimičuloin da viršilöin azundu	Мичурова Надежда. Изготовление маленьких и больших корзин
No	Диалект	Подкорпус	Жанр	Заголовок	Перевод																																	
1	Рыпушкальский	диалектные тексты		Kaikin paistih hierus livvikse	Мичурова Надежда. Все говорили в деревне на ливвиковском																																	
2	Неккульский	диалектные тексты		Joga talois oli lehmy	Мичурова Надежда. В каждом доме была корова																																	
3	Котозерский	диалектные тексты		Vastaimmo meijän saldottoi	Мичурова Надежда. Встречали наших солдат																																	
4	Рыпушкальский	диалектные тексты	бытовой рассказ	Poimičuloin da viršilöin azundu	Мичурова Надежда. Изготовление маленьких и больших корзин																																	

Рис. 13. Поиск в корпусе BenKar. Пример найденных ливвиковских текстов с аудио-записями.

Шаг 11. Как найти вепсские сказки, записанные до 1950 года?

- Выполните шаг 9.
- В поле «Язык» выберите «вепсский».
- В поле «Жанр» выберите «сказка».
- В поле «Год (по)» укажите ‘1950’.

- Нажмите кнопку **ПОКАЗАТЬ**.

The screenshot shows a search form with three filters: 'Язык' (Language) set to 'вепсский (1 748)', 'Жанр' (Genre) set to 'сказка', and 'Год (по)' (Year) set to '1950'. Below the filters, there is a dropdown for 'по' (from) with '10' selected, and a button labeled 'ПОКАЗАТЬ' (Show). A link 'Расширенный поиск' (Advanced search) is visible at the top right.

Найдено 20 записей.

No	Диалект	Подкорпус	Заголовок	Перевод
1	Средневепсский западный	фольклорные тексты	Pihkmut da kivut	Скатёрка и жерновок
2	Средневепсский западный	фольклорные тексты	Eli akaine, oli hänou poig	Жила женщина, был у нее сын
3	Северновепсский	фольклорные тексты	Viikuško čapab sizarel kåded	Брат отрубает руки у сестры

Рис. 14. Поиск вепсских сказок, записанных до 1950 года.

Шаг 12. Как осуществить поиск примеров употребления слова в корпусе?

- В верхнем меню найдите пункт «КОРПУС» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Лексико-грамматический поиск» (Рис. 15).

★ Лексико-грамматический поиск

The screenshot shows the 'Lexico-grammatical search' interface. It includes fields for 'Язык' (Language), 'Подкорпус' (Subcorpus), 'по' (from) with '10' selected, and a 'записей' (records) button. Below these are three search parameters: 'Слово 1' (Word 1) with a placeholder 'á', 'Часть речи' (Part of speech) with a placeholder 'é', and 'Грамматические признаки' (Grammatical features) with a placeholder 'é'. Each parameter has a '+' icon to its right, which opens a dropdown menu. A red 'ИСКАТЬ' (Search) button is located to the right of the search parameters.

Рис. 15. Поиск в корпусе заданной последовательности лемм и/или словоформ, обладающих определенными грамматическими характеристиками

В поле «Слово» можно задать начальную форму (лемму) и получить все формы этой лексемы, найденные в корпусе. Если нужно найти одну определенную форму слова, укажите ее в кавычках.

Например, для поисковой строки `^vuozis$` будут найдены примеры вхождения `vuvel`, `vuodel`, и т.д. Если нужно найти только `vuvel`, задайте эту форму в кавычках: `"^vuvel$"`.

Иконки открывают дополнительные окна для полей «Часть речи» и «Грамматические признаки». Галочками отметьте нужные значения.

Шаг 13. Как найти примеры употребления в корпусе людиковских имен в форме аппроксиматива?

- Выполните шаг 12.
- В поле «Выберите язык» выберите «карельский: людиковское наречие».
- В поле «Часть речи» первого слова нажмите на иконку .

Откроется дополнительное окно для выбора частей речи (Рис. 16).

Выберите часть речи

<input type="checkbox"/> вспомогательный глагол	<input type="checkbox"/> глагол	<input type="checkbox"/> другое
<input type="checkbox"/> имя собственное	<input type="checkbox"/> междометие	<input type="checkbox"/> местоимение
<input type="checkbox"/> наречие	<input type="checkbox"/> послелог	<input type="checkbox"/> предикатив
<input type="checkbox"/> предлог	<input type="checkbox"/> прилагательное	<input type="checkbox"/> союз подчинительный
<input type="checkbox"/> союз сочинительный	<input checked="" type="checkbox"/> существительное	<input type="checkbox"/> фразеологизм
<input type="checkbox"/> частица	<input type="checkbox"/> числительное	

выбрать закрыть

Рис. 16. Выбор части речи «существительное» в дополнительном окне.

- Отметьте галочкой «существительное».
- Нажмите кнопку **выбрать**.
- В поле «Грамматические признаки» первого слова нажмите на иконку

Откроется дополнительное окно для выбора грамматических признаков.

- Отметьте галочкой падеж «аппроксиматив».
- Нажмите кнопку **выбрать**.
- В основной форме (Рис. 17) нажмите кнопку **ИСКАТЬ**.

В новой вкладке загрузится результат поиска (Рис. 18). Стрелочки слева и справа предложения добавляют левый и правый контекст соответственно. Если щелкнуть на слово в предложении, откроется окошко с дополнительной грамматической и семантической информацией о слове.

The screenshot shows a search interface with the following fields:

- Язык:** карельский: людиковское наречие (270)
- Подкорпус:** (empty)
- по:** 10 **записей** **ИСКАТЬ**
- Слово 1:** (empty)
- Часть речи:** NOUN
- Грамматические признаки:** APPRX

Рис. 17. Поиск в корпусе примеров употребления людиковских имен в форме аппроксиматива

Результаты поиска

(язык: карельский: людиковское наречие) И

Слово 1: (существительное) И ((аппроксиматив))

Найдено 3 текста, 3 вхождения.

Уточнить запрос

1. Paušin Šan'uu. [Ken ei suvaiče hiirid, sil pidäy armastada kažid](#) (Oma mua. № 8, 2021, c. 10)

← Konz hyö oldih kuhn'al pääilluo, a minä magazin kätkydes i kieras rubezin itkemäh, Anapka ylen teravas hypiä minulluo i kaččeli minun silmiih, i minä hedi tyyništuin. →
2. Paušin Šan'uu. [Brus'a-kažiine](#) (Oma Mua. № 34 (1624), 2022, c. 9)

← Konz hyö oldih kuhn'al pääilluo, a minä magazin kätkydes i kieras rubezin etkemäh, Anapka ylen teravas hypiä minulluo i kaččeli minun silmiih, i minä hedi tyyništuin. →
3. [Lapsei oli yönit'kettai](#) / У ребёнка была ночная плачая (Спасская Губа (Mundärven laht), Кондопожский район, Республика Карелия, 2012)

← Šie vot mina kakš yön magain sel babuškalluo tulin da oma rodn'a da vs'o. →

Рис. 18. Результаты поиска примеров людиковских имен в форме аппроксиматива

Шаг 14. Как найти примеры употребления в корпусе ливвицких глаголов в форме перфекта и плюсквамперфекта кондиционала?

- Выполните шаг 12.

- В поле «Язык» выберите «карельский: ливвиковское наречие».
- В поле «Слово 1» укажите лемму ‘olla’¹⁶.
- В поле «Часть речи» первого слова нажмите на иконку
- В дополнительном окне (Рис. 16) отметьте галочкой «глагол».
- Нажмите кнопку
- В поле «Грамматические признаки» первого слова нажмите на иконку
- В дополнительном окне отметьте галочкой наклонение «кондиционал».
- Нажмите кнопку
- Справа нажмите на иконку

На месте плюса появятся поля «Расстояние». Ниже откроются аналогичные поля для второго слова.

- В поле «Часть речи» второго слова нажмите на иконку
- В дополнительном окне отметьте галочкой «глагол».
- Нажмите кнопку
- В поле «Грамматические признаки» второго слова нажмите на иконку
- В дополнительном окне отметьте галочками залог «актив» и «2-е причастие».
- Нажмите кнопку
- В основной форме (Рис. 19) нажмите кнопку

В новой вкладке браузера загрузится результат поиска (Рис. 20). Если щелкнуть на слово в предложении, откроется окошко с дополнительной грамматической и семантической информацией о слове.

The screenshot shows a search interface with two search queries. The first query is for 'карельский: ливвиковское наречие (489)' with 'Verb' selected as the part of speech and 'COND' as the grammatical feature. The second query is for 'olla' with 'Verb' selected as the part of speech and 'ACT,V,PTCP,PST' as the grammatical feature. Both queries have '1' selected for 'from' and '1' for 'to' in the 'Distance' field. A red plus sign icon is visible on the right side of the interface.

Рис. 19. Поиск в корпусе примеров употребления ливвиковских глаголов в форме перфекта и плюсквамперфекта кондиционала

¹⁶ Чтобы отсечь варианты лемм, которые содержат подстроку ‘olla’, можно использовать шаблон ^olla\$.

Результаты поиска

(**язык**: карельский: ливвиковское наречие) И

Слово 1: olla И (глагол) И ((кондиционал)) И

на расстоянии от 1 до 1 **Слово 2:** (глагол) И ((актив) И (2-е причастие))

Найдено 9 текстов, 9 вхождений.

1. [Bul'uu borkananke](#) (Irina Kudel'nikova, Oma mua. № 1, 2018, с. 11)

← Vie pidäs teis dengat ottua, ka olgah, prostin. Äijän rahvastu **olluzin parandannuh**, a työ kallehen syömisen kaimaittol!

2. [Kargiet vainuajat](#) (Fodorova Anni (nygoi Ivanova), Oma mua. № 24, 20

parandannuh

← Kai kylä ollus **palanuh**, ku vahnembat ei ehtittys. →

parandua

3. [Ken da kui kalasti](#) (NIKOLAI FILATOV, Oma mua. № 12, 2020, с. 11)

← Tiettäväine, emmo **olis tundenuh** heidy, kerdu kolme vuotti olimm

1) улучшать, улучшить;

совершенствовать,

усовершенствовать

4. [Ken da kui kalasti](#) (NIKOLAI FILATOV, Oma mua. № 12, 2020, с. 11)

← Tiettäväine, emmo **olis tundenuh** heidy, kerdu kolme vuotti olimm

2) лечить, вылечить,

исцелять, исцелить

5. [Kuldastu laduu ei umbua lumel](#) (Raisa Bogdanova, Oma mua. № 5, 2

← Tänävuon Fodor Terentjev **olis täytänyh** 95 vuottu. →

- актив, 2-е причастие

mmo ego hyögi oldas ajamas. →

mmo ego hyögi oldas ajamas. →

Расширенный поиск ↓

Язык × карельский: ливвиковское наречие (954)	Подкорпус × фольклорные тексты (307)	по 10 записей ИСКАТЬ								
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 25%; padding: 5px;">Слово 1 <input type="text"/></td> <td style="width: 25%; padding: 5px; text-align: center;">Часть речи ? ä VERB</td> <td style="width: 25%; padding: 5px;">Грамматические признаки <input type="text"/></td> <td style="width: 25%; padding: 5px; text-align: center;">Расстояние ? + от 1 до 1</td> </tr> <tr> <td>Слово 2 <input type="text"/></td> <td style="text-align: center;">Часть речи ? ä NOUN</td> <td>Грамматические признаки <input type="text"/></td> <td style="text-align: center;">+</td> </tr> </table>			Слово 1 <input type="text"/>	Часть речи ? ä VERB	Грамматические признаки <input type="text"/>	Расстояние ? + от 1 до 1	Слово 2 <input type="text"/>	Часть речи ? ä NOUN	Грамматические признаки <input type="text"/>	+
Слово 1 <input type="text"/>	Часть речи ? ä VERB	Грамматические признаки <input type="text"/>	Расстояние ? + от 1 до 1							
Слово 2 <input type="text"/>	Часть речи ? ä NOUN	Грамматические признаки <input type="text"/>	+							

Рис. 21. Поиск в корпусе примеров употребления глаголов с существительными в партитиве

В новой вкладке браузера загрузится результат поиска (Рис. 22).

Результаты поиска

(язык: карельский: ливвиковское наречие) И (подкорпус: фольклорные тексты) И

Слово 1: (глагол) И

на расстоянии от 1 до 1 Слово 2: (существительное) И ((партитив))

Найдено 26 текстов, 124 вхождения.

Уточнить запрос

1. Barka da d'etka ištutettlh nagrehen / Бабка и дедка посадили репку (Карельская (Karjažet), Лодейнопольский р-н, Ленинградская обл., 1959)
 - ← Siien koiru kučui kažii: veettih, veettih viijei – ei voittu ni kui vediä. →
2. [Ivanuška-boranuška] / [Иванушка-баранушка] (Самбатука (Sammatus), Олонецкий район, Республика Карелия, 1959)
 - ← Brihaču sanowgi: "Oi, čidžoini, minä täs juan vettü l'ehmän kabjan jäl'l'es!" →
 - ← Ruvetah händü tappamah. →
 - ← Häi menöw, itkow-itkow: "Oi, sestrica, Ol'önüška, minuw ruvetah iškemäh; veiččii tahkotah, kirvehii hivotah, kattilat kiahutah, minuw iškemäh ruvetah!". →
 - ← Häi ku menöw vezirandah dai sanow: "Oi, sestrica Ol'önüška! →
 - ← Minuw iškemäh ruvetah: veiččii hivotah, kirvehii tahkotah, minuw ruvetah iškemäh!" →
3. Suaru hukkah niškoj / Сказка про волка (Самбатука (Sammatus), Олонецкий район, Республика Карелия, 1959)
 - ← "Na, täs, akku, sanow, lihua, keitä lapsil'e lihua, sanow, da piästäh n'al'l'as hot' iäres". →

Рис. 22. Результаты поиска в корпусе примеров употребления глаголов с существительными в партитиве

Шаг 16. Как найти примеры употребления конкретной словоформы в вепсских текстах 30х годов?

- Выполните шаг 12.
- В поле «Выберите язык» выберите «вепсский».
- Нажмите «Расширенный поиск» (Рис. 15).
- В поле «Год (с)» впишите '1930'.
- В поле «Год (по)» впишите '1940'.
- В поле «Слово 1» укажите лемму или шаблон, например '^ö\$'¹⁷.
- В поле «Грамматические признаки» нажмите на иконку **+**
- В дополнительном окне отметьте галочкой нужные грамматические категории, например, падеж «адессив».
- Нажмите кнопку **выбрать**.
- В основной форме (Рис. 23) нажмите кнопку **ИСКАТЬ**.

¹⁷ По шаблону ^ö\$ будут найдены формы слова ö. Если написать просто 'ö', то будут найдены формы всех слов с входящей буквой ö: söda, hö, löuta и т.д. Подробнее о шаблонах в таблице 4.

Язык
вепсский (1 751)

Подкорпус

Диалект

Жанр

Год (с) 1930

Год (по) 1940

по 10 записей ИСКАТЬ

Слово 1 ~ö\$

Часть речи ~ä

Грамматические признаки AT+ESS

Простой поиск ↑

Рис. 23. Поиск в корпусе примеров употребления вепсской лексемы ‘ö’ в форме адессива в текстах 30-х годов

В новой вкладке браузера загрузится результат поиска (Рис. 24).

Результаты поиска

(язык: вепсский) И год (с): 1930 И год (по): 1940 И

Слово 1: ^ö\$ И ((адессив))

Найден 1 текст, 1 вхождение.

[Уточнить запрос](#)

1. [Viikuško čapab sizarel kåded](#) / Брат отрубает руки у сестры (Tikhoništa), Прионежский р-н, Республика Карелия, 1937

← Toizel öl car' pani karauł'maha vanhamban poigan. →

Скрипт был в

öl

ö

существительное

ночь

- адессив, ед. ч.

Рис. 24. Поиск в корпусе примеров употребления вепсской лексемы ‘ö’ в форме адессива в текстах 30-х годов

Шаг 17. Как найти частотные словари корпуса?

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «КОРПУС» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Частотные словари».

Загрузится страница со списком частотных словарей (Рис. 25).

★ Частотные словари

- [Частотный словарь токенов](#)
- [Частота символов в текстах](#)
- [Частотный словарь лемм, размеченных в тексте](#)

Рис. 25. Страница с частотными словарями корпуса

Шаг 18. Как найти самые частотные слова в текстах?

- Выполните шаг 17.
- Щелкните по ссылке «Частотный словарь лемм, размеченных в текстах».

Откроется страница с поисковой формой (Рис. 26).

Язык
вепсский (1 121 026)

Диалект

Часть речи

ПОКАЗАТЬ

Рис. 26. Частотный словарь лемм, размещенных в тексте

- В полях «Язык», «Диалект», «Части речи» выберите нужные значения (или оставьте пустыми поля).
- Нажмите кнопку ПОКАЗАТЬ.

Выведется список слов, встречающихся в текстах (Рис. 27).

The screenshot shows a search interface with three dropdown menus: Язык (carельский: ливвикови), Диалект (Новописьменный ливвиковск...), and Часть речи (существительное (552)). A red button labeled 'ПОКАЗАТЬ' is visible. Below the search bar, there are filters for 'Показать' (10 записей) and 'Поиск:' (empty). The main area is a table with columns: №, Лемма, Часть речи, Частота. The 'Лемма' column is highlighted with a red arrow. The data shows 10 entries from 3,876 total, sorted by frequency. The first entry is 'ома' (1629). The table includes a footer with page navigation (Предыдущая, 1, 2, 3, 4, 5, ..., 388, Следующая).

№	Лемма	Часть речи	Частота
1763	ома	существительное	1629
1055	sana	существительное	1231
175	yuozi	существительное	1195
2470	kylä	существительное	1125
1274	rahvas	существительное	1070
3260	hyvä	существительное	987
3847	aigu	существительное	979
2964	karjal	существительное	897
2712	kodi	существительное	890
96	äijy	существительное	854

Записи с 1 до 10 из 3,876 записей

Предыдущая 1 2 3 4 5 ... 388 Следующая

Рис. 27. Частотный словарь ливвиковских существительных, размещенных в новописьменных текстах

По умолчанию список отсортирован по частоте встречаемости слова в текстах. Можно менять сортировку, нажимая стрелочки СПРАВА от заголовка колонки (Рис. 27).

Стоит отметить, что считаются все связи слова в тексте со словарной статьей (в том числе и непроверенные), кроме связей с меткой "совсем не подходит". Если слово в тексте автоматически было "привязано" к лемме с несколькими значениями и эти значения еще не проверены (не снята семантическая омонимия), то будут посчитаны все связи с этими значениями.

Шаг 19. Как найти «Речевой корпус»?

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «КОРПУС» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Речевой корпус».

Выведется список озвученных текстов (Рис. 28).

★ Речевой корпус

[Аудио-карта](#) | Создать новый | ? Помощь

Расширенный поиск ↓

Язык		по 10 записей		ПОКАЗАТЬ
------	--	---------------	--	----------

Найдено 46 записей.

No	Диалект	Говор	Заголовок	Перевод	Прослушать
1	Вокнаволокский	Суднозеро	Meijen talo šuuri oli	Наш дом большой был	0:00 / 2:58
2	Рыпушкальский	Алексала	Kaikin paisti hierus livvikse	Мичурова Надежда. Все говорили в деревне на ливвиковском	0:00 / 1:09
3	Неккульский	Обжа	Joga talois oli lehmy	Мичурова Надежда. В каждом доме была корова	0:00 / 0:59

Рис. 28. Речевой корпус прибалтийско-финских языков Карелии

В колонке «Говор» выводится населенный пункт - место рождения информанта.

Шаг 20. Как найти и прослушать образцы речи интересующего населённого пункта?

- Выполните шаг 19.
- Нажмите на ссылку «Расширенный поиск» (Рис. 28).
- Если Вас интересует населенный пункт, где были собраны образцы, то выберите нужное значение в списке «Населенный пункт записи».
- Если Вас интересует населенный пункт, в котором родились информанты, то выберите нужное значение в списке «Населенный пункт рожд. информанта» (Рис. 29).

Язык	Диалект	Заголовок
		?
Область, республика записи	Район записи	Населенный пункт записи
		?
Область, республика рожд. информанта	Район рожд. информанта	Населенный пункт рожд. информанта
		?
Информант	Собиратель	
Источник	?	
по 10 записей		ПОКАЗАТЬ

Рис. 29. Выбор населенного пункта рождения информанта

Шаг 21. Как ссылаться на материалы корпуса ВепКар?

Если вы используете материалы корпуса в своей исследовательской работе, то вы можете выбрать следующие варианты.

1. В тексте или в сноске указать:

Исследование проведено на материале Открытого корпуса вепсского и карельского языков (dictorus.krc.karelia.ru).

2. В списке литературы в конце работы привести ссылку на Корпус:

Открытый корпус вепсского и карельского языков (ВепКар). 2009—2022. Доступен по адресу: dictorus.krc.karelia.ru

3. Ссылку на конкретный текст оформляйте, например, так:

Lehme № 2051 // Открытый корпус вепсского и карельского языков ВепКар. URL: <http://dictorus.krc.karelia.ru/ru/corpus/text/2051>

Отметим, что вместо ссылки на весь корпус, вы можете поставить ссылку на какую-либо из наших публикаций о корпусе. Этот список есть на сайте ВепКар: <http://dictorus.krc.karelia.ru/ru/page/publ>

2.2.3. Поиск в словаре ВепКар

Шаг 22. Как осуществлять поиск в словаре ВепКар?

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «СЛОВАРЬ» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Леммы» или «Словоформы» для поиска в словаре соответственно лемм или словоформ (Рис. 30).

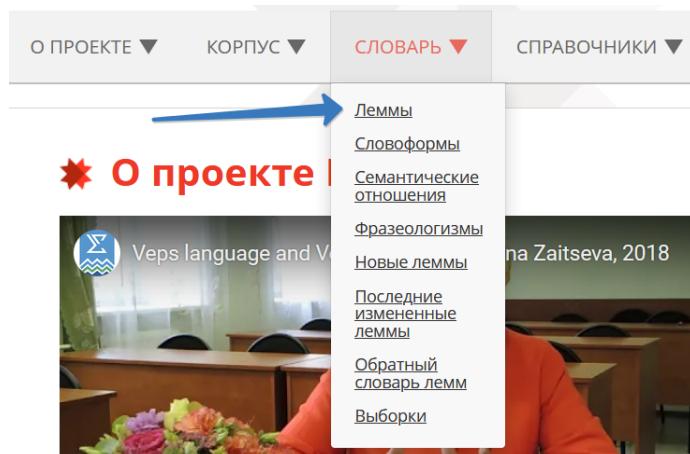


Рис. 30. Поиск лемм в словаре BenKar

Шаг 23. Как найти вепсское / карельское слово по начальной форме?

- Выполните шаг 22, щелкнув по ссылке «Леммы».
- Укажите вепсское (карельское) слово в поле «лемма».
- Нажмите кнопку **ПОКАЗАТЬ** (Рис. 31).

При поиске Вы можете задавать фрагменты леммы, используя простые шаблоны (см. Шаг 5).

★ Леммы

Поиск лемм по словоформам | Список длинных лемм | Создать новую

ID maa ⓘ Выберите язык по 10 записей ПОКАЗАТЬ

Найдено 3 записи.

No	Лемма	Язык	Часть речи	Толкование	Словоформы *	Примеры **
1	maa	карельский: собственно карельское наречие	существительное	земля, суша	2	0
2	maa	карельский: ливвиковское наречие	существительное	1) земля, суша 2) почва	2	0 / 2 / 2
3	maa	вепсский	существительное	1) земля, суша 2) почва	2	0 / 2 / 2

* - Количество словоформ с грамматическими признаками [+ кол-во словоформ без грамматических признаков]

** - Количество проверенных примеров / Количество непроверенных примеров / Общее количество

Рис. 31. Поиск лемм в словаре BenKar по начальной форме слова. Поле «лемма» с иконками «помощь» и «специальные символы». Ссылка «расширенный поиск» для открытия полной формы

Шаг 24. Как указать язык/наречие при поиске леммы?

- Выполните шаг 22, щелкнув по ссылке «Леммы».
- Щелкните на поле «Выберите язык».
- Выберите нужный язык/наречие (Рис. 32).

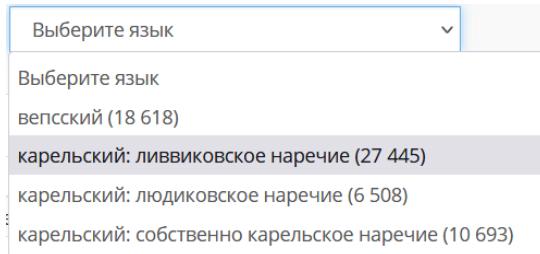


Рис. 32. Выбор языка из выпадающего списка

Шаг 25. Как осуществлять расширенный поиск в словаре BenKar?

- Выполните шаг 22, щелкнув по ссылке «Леммы».
- Щелкните на ссылку «расширенный поиск» справа вверху над поисковой формой (Рис. 31).

В форме откроются дополнительные поля (Рис. 33).

The screenshot shows the search interface for the BenKap dictionary. At the top, there are four input fields: 'ID' (disabled), 'лемма' (lemma) with a help icon and a red 'ä' symbol, 'Выберите язык' (Select language) with a dropdown arrow, and 'диалект' (dialect) with a help icon and a red 'ä' symbol. Below these are two more dropdowns: 'Выберите часть речи' (Select part of speech) and 'Выберите понятие' (Select concept). To the right of these dropdowns are checkboxes for 'с примерами' (with examples) and 'записей' (records). A red button labeled 'ПОКАЗАТЬ' (SHOW) is positioned at the bottom right. At the very bottom right of the interface is a link 'Простой поиск ↑' (Simple search ↑).

Рис. 33. Расширенный поиск лемм в словаре BenKap

Шаг 26. Как найти вепсское (карельское) слово по русскому толкованию?

- Выполните шаг 25.
- Введите текст в поле «**толкование**» (Рис. 33).
- Нажмите кнопку **ПОКАЗАТЬ** (Рис. 34).

The screenshot shows the search results page for the BenKap dictionary. At the top, there are several search parameters: 'ID' (disabled), 'лемма' (lemma) with a help icon and a red 'ä' symbol, 'карельский: ливвиковское наречие' (Karelian: Livvian dialect) with a dropdown arrow, and 'существительное (36 726)' (Noun (36 726)) with a dropdown arrow. Below these are the search terms '%коров%' (cow%) with a help icon and a red 'ä' symbol, and a search button 'ПОКАЗАТЬ'. The results section starts with the message 'Найдено 14 записей.' (Found 14 records). A table follows, listing 14 noun entries related to cows, each with a 'No' column, a 'Лемма' (Lemma) column, a 'Толкование' (Definition) column, a 'Словоформы *' (Wordforms *) column, and a 'Примеры **' (Examples **) column.

No	Лемма	Толкование	Словоформы *	Примеры **
1	jumalanlehmäine	божья коровка	37	0
2	lehmy	корова	37	5 / 113 / 118
3	lehmykarju	стадо коров	37	0 / 1 / 1
4	lehmänkello	колокольчик на шее у коровы	37	0 / 1 / 1
5	lehmänmaido	коровье молоко	37	0
6	lehmännahku	шкура коровы	37	0
7	lehmänsarvi	рог коровы	54	0
8	lehmö	корова	2	0

Рис. 34. Поиск лемм в словаре BenKap по языку (ливвиковское наречие карельского языка), части речи (существительное) и толкованию (%коров%)

Шаг 27. Как найти вепсское (карельское) слово по словоформе?

Можно искать слово не только по начальной, но и по любой его грамматической форме - словоформе.

- Выполните шаг 22, щелкнув по ссылке «Леммы».
- Щелкните на ссылку «Поиск лемм по словоформам»¹⁸ (Рис. 31).

Откроется страница с новой поисковой формой (Рис. 35). В этом поиске Вы можете указать словоформу (слово или специализированный шаблон, см. шаг 6).

- Если известна форма слова, заполните поле «*словоформа I*».
- Если известны грамматические признаки слова, выберите нужное значение в списке «*грамматические признаки для словоформы*».
- Нажмите кнопку **ПОКАЗАТЬ**.

¹⁸ Поиск лемм по словоформам: http://dictorus.krc.karelia.ru/ru/dict/lemma/by_wordforms

★ Поиск лемм по словоформам

Рис. 35. Поиск лемм в словаре VenKar по словоформам

Шаг 28. Как найти все глаголы новописьменного севернокарельского языка, начинающиеся с буквы ‘m’, у которых начальная форма (инфinitiv) оканчивается на ‘ио’ или ‘ёö’, а форма 1 лица ед.ч. настоящего времени оканчивается на ‘ип’ или ‘ун’?

- Выполните шаг 27.
- В поле «язык» выберите «карельский: собственно карельское наречие».
- В поле «диалект» выберите «Новописьменный севернокарельский».
- В поле «часть речи» выберите «глагол».
- В поле «словоформа 1» введите шаблон «^m.+[uy][oö]\$»
- В поле справа выберите «I инфинитив»
- Нажмите иконку «добавить словоформу» (Рис. 35).
- В новом поле «словоформа 2» введите шаблон «[uy]n\$»
- В поле справа выберите «индикатив, презенс, I л., ед. ч., полож. ф.»
- Нажмите кнопку ПОКАЗАТЬ (Рис. 36).

Найдено 25 записей.

No	Лемма	Толкование	Словоформы *	Примеры **
1	mahtuo	1) вмещаться, входить 2) ладить, жить в ладу	242 (mahtuo, mahun)	0 / 23 / 23
2	maistuo	иметь какой-либо вкус	125 (maistuo, maissun)	0 / 4 / 4
3	mataltuo	становиться ниже, мелеть	125 (mataltuo, matallun)	0 / 1 / 1
4	matkeutuo	1) собираться в дорогу 2) дойти (о новости)	125 (matkeutuo, matkeuvun)	0 / 3 / 3
5	matoutuo	зачервиветь	125 (matoutuo, matouvun)	0
6	mehuo	течь, просачиваться	125 (mehuo, mehun)	0 / 1 / 1
7	meruutuo	плавиться, топиться (о масле, жире)	125 (meruutuo, meruuuvun)	0
8	mečittyö	зарастать лесом	125 (mečittyö, mečtyn)	0 / 2 / 2

Рис. 36. Пример поиска всех глаголов новописьменного севернокарельского языка, начинающихся с буквы ‘m’, у которых начальная форма (инфинитив) оканчивается на ‘ио’ или ‘ёö’, а форма 1 лица ед.ч. настоящего времени оканчивается на ‘ип’ или ‘ун’

Шаг 29. Как найти обратный словарь?

* Обратный словарь лемм

Выберите язык Выберите часть речи по 10 записей ПОКАЗАТЬ

Рис. 37. Поиск по обратному словарю лемм.

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «СЛОВАРЬ» и щелкните на него мышкой.
- В выпадающем меню щёлкните по ссылке «Обратный словарь лемм».
- В загрузившейся форме (Рис. 37) щелкните на поле «Выберите язык» и выберите нужный язык/наречие.
- В поле «Выберите часть речи» выберите нужную часть речь (если нужно).
- Нажмите кнопку ПОКАЗАТЬ.

В загрузившейся таблице (Рис. 38) будут выведены слова, отсортированные по алфавиту по последней букве.

карельский: людиковское наречи ▾ Выберите диалекты ? существительное (37 202) ▾

лемма ? ä по 10 записей ПОКАЗАТЬ

Найдена 1 561 запись.

No	Лемма	Часть речи
101	šleja	NOUN
102	agja	NOUN
103	sija	NOUN
104	magadussija	NOUN
105	oja	NOUN
106	koja	NOUN
107	kirja	NOUN
108	nahka	NOUN
109	pahka	NOUN
110	tähkä	NOUN

Рис. 38. Поиск по обратному словарю лемм

Заключение

В этом руководстве для пользователей приведены основные понятия корпусной лингвистики, описана разметка, типология и обзор современных лингвистических корпусов. Показаны основные виды работы с лингвистическим корпусом ВепКар: регистрация, переключение языка интерфейса, поиск в словаре и текстах с помощью простых и специализированных шаблонов. Приведены примеры сложных запросов, в которых указывается сразу несколько (много) параметров поиска.

Надеемся, что предложенное пособие будет востребовано в сферах среднего и высшего образования в процессе преподавания вепсского и карельского языков, а также корпусной лингвистики. Замечания и предложения по улучшению руководства можно отправлять разработчикам корпуса ВепКар по адресу: nataly@krc.karelia.ru .

Тезаурус

Корпус (лингвистический) — коллекция текстов, специально отобранных, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска.

Корпус аннотированный/размеченный — корпус текстов, в котором содержатся специальные метки, позволяющие получать из корпуса данные (статистику, языковые примеры и др.) по каким-либо лингвистическим параметрам (части речи, грамматической форме, синтаксической функции и т. п.) [Захаров, Богданова, 2020].

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий [Захаров, Богданова, 2020].

Корпусный менеджер (корпус-менеджер) — специальная информационно-поисковая система, включающая программные средства для поиска данных в корпусе, получения статистической информации.

Лемма - 1) начальная форма слова; 2) словарная статья в словаре ВенКар.

Лексико-грамматический поиск — поиск заданной последовательности лемм и/или словоформ, обладающих определенными грамматическими характеристиками.

Миноритарный язык — язык национального (этнического) меньшинства [ССТ: 129].

Морфологический анализ — определение леммы и ее грамматических характеристик.

Объем корпуса — информация об общем объеме корпуса и о количестве извлеченных из текста примеров.

Разметка (аннотация) текста — введенная вручную или автоматически лингвистическая или метатекстовая информация обо всех выбранных единицах текста (тексте, предложении, словосочетании, словоформе, морфеме).

Репрезентативность — свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового многообразия текстов [Копотев, 2014].

Сбалансированность — свойство корпуса, определяющее насколько равномерно представлены тексты разных типов.

Список использованной литературы

[Albert, 2012] Albert, Saul. The Audio BNC. 2012. URL: <https://saulalbert.net/blog/the-audio-bnc/> (дата обращения: 04.11.2021).

[Arkhangelskiy, 2020] Arkhangelskiy, T., 2020. Web Corpora of Volga-Kama Uralic Languages // Finno-Ugric Languages and Linguistics. Vol. 9. No. 1-2. (2020), P. 58–66.

- [Klyachko, 2019] Klyachko E. L., Sorokin A. A., Krizhanovskaya N. B., Krizhanovsky A. A., Ryazanskaya G. M. LowResourceEval-2019: a shared task on morphological analysis for low-resource languages // Computational Linguistics and Intellectual Technologies : papers from the Annual conference “Dialogue”. — М.: РГГУ, 2019. — Вып. 18 (25). — С. 45—62.
- [Krizhanovskaya et al., 2022] Krizhanovskaya, N., Novak, I., Krizhanovsky, A., & Pellinen, N. Morphological inflectional rules for Karelian Proper verbs. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(2). 2022. Pp. 47–78. <https://doi.org/10.12697/jeful.2022.13.2.02>
- [Meyer, 2002] Meyer, Charles. English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 2002. URL: http://assets.cambridge.org/97805218/08798_frontmatter/9780521808798_frontmatter.pdf (дата обращения: 07.06.2021).
- [Баранов, 2001] Баранов А. Н. Введение в прикладную лингвистику. – М, 2001. – Т. 2.
- [Бойко и др., 2021] Бойко Т. П., Зайцева Н. Г., Крижановская Н. Б., Крижановский А. А., Новак И. П., Пеллинен Н. А., Родионова А. П., Трубина Е. Д. Лингвистический корпус ВепКар — «заповедник» прибалтийско-финских языков Карелии // Труды КарНЦ РАН. № 7. Комплексные научные исследования КарНЦ РАН. 2021. С. 100-115 DOI: 10.17076/them1415
- [Герд, 2008] Герд А. С. Национальный корпус русского языка – Словарная картотека – Академический словарь // Тр. Междунар. конф. «Корпусная лингвистика–2008» – 2008. – с. 143–148. URL: http://www.project.phil.spbu.ru/corpora2011/Works2008/Gerd_143_148.pdf (дата обращения: 21.06.2021).
- [Зайцева, 2012] Зайцева Н.Г. Вепские причитания в фокусе корпусной лингвистики и лингвофольклористики // Материалы XLI Международной филологической конференции. 26-31 марта 2012 г. Секция "Уралистика". СПб.: Филологический факультет СПбГУ. 2012. С. 16-26.
- [Захаров, Богданова, 2020] Захаров В. П., Богданова С. Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.: Изд-во С.-Петерб. ун-та, 2020. — 234 с.
- [Копотев, Янда, 2006] Копотев М. В., Янда Л. Национальный корпус русского языка // Вопросы языкознания. № 5. 2006. С. 149–155.
- [Копотев, 2014] Копотев, Михаил. Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. Прага, 2014
- [Крижановская, Крижановский, 2020] Крижановская Н.Б., Крижановский А.А. От корпуса ВепКар к лингвистической платформе // Материалы Всероссийской научной конференции с международным участием «Бубриховские чтения: задокументированное народное слово» (Петрозаводск, 27–28 октября 2020 г.) [Электронный ресурс]. 2020. С. 157–159. URL: <http://mathem.krc.karelia.ru/publ.php?id=19787> (дата обращения: 07.06.2021).
- [ССТ] Словарь социолингвистических терминов. М., 2006.

[Плунгян, Сичинава, 2004] Плунгян В. А., Сичинава Д. В. Национальный корпус русского языка: опыт создания корпуса текстов современного русского языка // Труды международной конференции «Корпусная лингвистика-2004». — СПб.: Издательство Санкт-Петербургского университета, 2004. С. 216–238. URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2004/Sitchinava_art.pdf