Федеральный исследовательский центр
«Карельский научный центр Российской академии наук»
Институт прикладных математических исследований КарНЦ РАН

Н. Б. Крижановская, А. А. Крижановский

ВепКар:
руководство для пользователей

Учебное пособие

Учебное электронное издание

Петрозаводск
КарНЦ РАН
2023

ISBN 978-5-9274-0961-7

УДК 004:811.511.1(075)
ББК 32.81я7
К85

Научные редакторы:
И. П. Новак, Н. А. Пеллинен

Печатается по решению Ученого совета
Института прикладных математических исследований
КарНЦ РАН

Крижановская, Н. Б.

К85 ВепКар: руководство для пользователей : учебное пособие : учебное электронное издание / Н. Б. Крижановская, А. А. Крижановский ; научные редакторы И. П. Новак, Н. А. Пеллинен ; Федеральный исследовательский центр «Карельский научный центр Российской академии наук», Институт прикладных математических исследований КарНЦ РАН. — Петрозаводск : КарНЦ РАН, 2023. — 1 DVD-ROM. — Систем. требования: PC, MAC с процессором Intel 1,3 ГГц и выше; Microsoft Windows, MAC OSX; 256 Мб (RAM); видеосистема: разрешение экрана 800×600 и выше, графический ускоритель (опционально); мышь; Adobe Reader; дисковод DVD-ROM. — Загл. с титул. экрана. — Текст: электронный.

ISBN 978-5-9274-0961-7

В учебном пособии объясняются основные понятия корпусной лингвистики. Описана работа в Открытом корпусе вепсского и карельского языков ВепКар. Этот корпус разрабатывается в Карельском научном центре РАН и доступен в сети Интернет (http://dictorpus.krc.karelia.ru). В пособии разобрано большое количество простых и сложных запросов, позволяющих извлекать интересные языковые данные из корпуса текстов и из словаря ВепКар.

Учебное пособие позволит школьникам, студентам и учителям овладеть тонкостями корпусного менеджера ВепКар, а также позволит ученым проводить современные исследования прибалтийско-финских языков народов Карелии на высоком уровне.

УДК 004:811.511.1(075)
ББК 32.81я7

Текстовое (символьное) электронное издание

Системные требования: PC, MAC с процессором Intel 1,3 ГГц и выше; Microsoft Windows, MAC OSX;
256 Мб (RAM); от 500 Мб свободного пространства на жестком диске;
видеосистема: разрешение экрана 800×600 и выше, графический ускоритель (опционально);
мышь; Adobe Reader; дисковод DVD-ROM

Для создания электронного издания использованы
ПО Microsoft Word, Adobe Acrobat Pro

Редактор *М. А. Радостина*
Компьютерная верстка *Н. Н. Сабанцева*
Оформление обложки и этикетки диска *Т. В. Уткина*

Подписано к использованию 08.02.2023. 1 DVD-ROM. 9,33 Мб.
Тираж 100 экз. Заказ № 748

Федеральное государственное бюджетное учреждение науки
Федеральный исследовательский центр
«Карельский научный центр Российской академии наук»
185910, г. Петрозаводск, ул. Пушкинская, д. 11
Телефон (8142) 76-60-40. E-mail: krcras@krc.karelia.ru
URL: http://www.krc.karelia.ru

Изготовлено в Федеральном государственном бюджетном учреждении науки
Федеральный исследовательский центр
«Карельский научный центр Российской академии наук»
185910, г. Петрозаводск, ул. Пушкинская, д. 11
Телефон (8142) 76-60-40. E-mail: krcras@krc.karelia.ru
URL: http://www.krc.karelia.ru

Содержание

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 языками[1]. 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 et al., 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].

Наличие корпуса для языка, объем 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].

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

Для анализа языкового явления и проведения лингвистического исследования необходим языковой материал в электронной форме, т. е. *лингвистический корпус*, 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 века, 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,* т. е. 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,* т. е. 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;
- *экстралингвистические* 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:

- *Морфологическая разметка* отражает признаки части речи и грамматических категорий, свойственных данной части речи. Код привязан к отдельному 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.

Экстралингвистическая разметка

Дополнительная информация о текстах корпуса называется внешней или экстралингвистической разметкой (метаразметкой). Экстралингвистическая 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;
- другие свойства.

1.1.3. Типология корпусов

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

1. Относится ли корпус к языку в целом (возможно, к языку заданного периода), или корпус содержит тексты определенного 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.

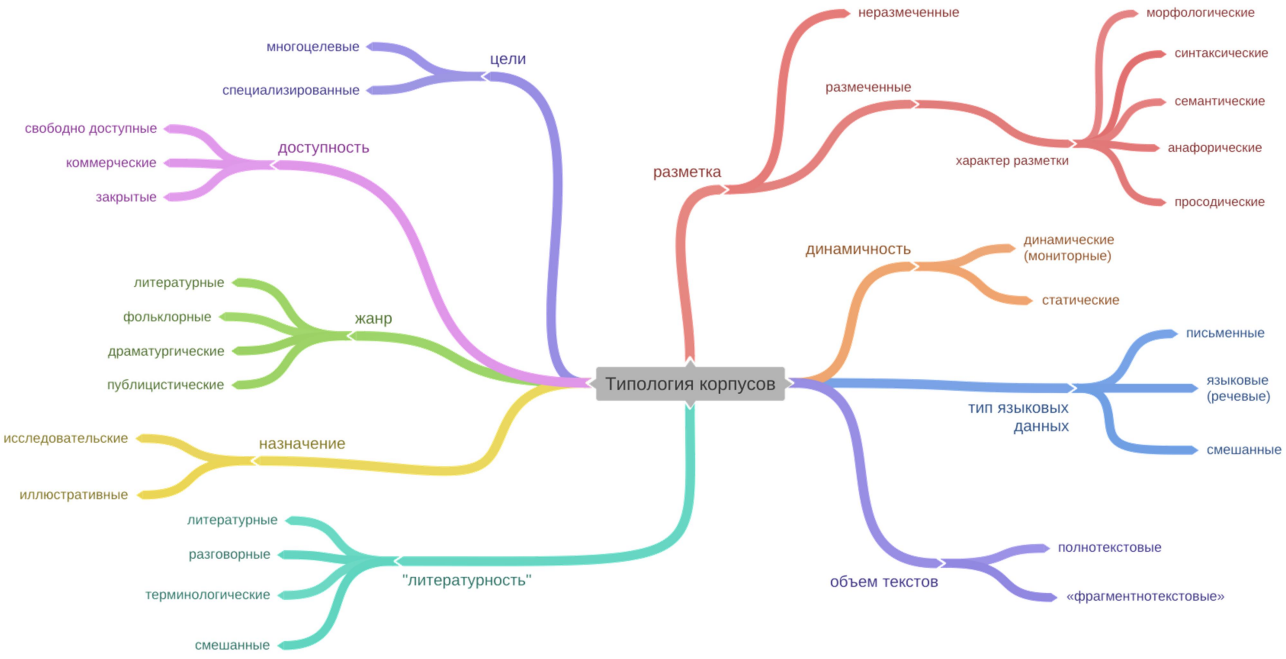


Рис. 1. *Типы корпусов по классификации В. П. Захарова*

1.1.4. Обзор существующих корпусов различных типов

В настоящее время для большинства основных языков мира существуют национальные общеязыковые корпусы. Число лингвистических корпусов измеряется тысячами и постоянно растет. Два крупнейших специализированных каталога CLARIN[2] и ELRA[3] содержат информацию о более чем 3000 корпусов. Список лингвистических ресурсов, включающий и корпусы, можно найти на сайте Ассоциации компьютерной лингвистики[4]. Корпусы используются как для лингвистических исследований, так и в прикладных целях.

Зарубежные корпусы

Британский национальный корпус[5] (кратко BNC) является большим (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 пленку свои обычные разговоры с друзьями, соседями, в семье, на работе. Была записана речь в самых разных контекстах — от разговоров на формальных деловых или правительственных 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 аудиокассет[6], 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.

---

[2] Common Language Resources and Technology Infrastructure, https://www.clarin.eu/resource-families
[3] European Language Resources Association, http://www.elra.info/en/
[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/

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 «подко́рпуса» 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 онлайн[8]. 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].

Таблица 1. **Примеры корпусов русского языка** (объем словоупотребления в млн)

Корпусы	Адрес	Объем
Упсальский и Тюбингенский корпуса	https://www.lingexp.uni-tuebingen.de/sfb441/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	1000
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

[8] См. https://ruscorpora.ru

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 мордовских, марийского, коми, удмуртского, мансийского языков. На сайте проекта «Малые языки Сибири» (Siberian Lang | Малые 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 литература, научные статьи, переводы Библии, статьи Википедии, официальные тексты и публичные записи в социальных сетях [Arkhangelskiy, 2020, с. 58–61].

Таблица 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

Таблица 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 Н. Г. Зайцевой был создан «Корпус вепсского языка»[9] [Зайцева, 2012]. Стоит задача возрождения вепсского и карельского языков и культуры. Для этого нужны большие электронные языковые ресурсы, которые позволят 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» был создан новый интернет-ресурс «Открытый корпус вепсского и карельского языков» (ВепКар)[10]. В корпус помимо вепсского вошли три карельских подкорпуса: собственно карельский, ливвиковский и людиковский. Объединенная 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 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 на 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 размечаются, и от слов в тексте идут отсылки на соответствующие значения в словарных статьях.

---

[9] См. http://vepsian.krc.karelia.ru/about/

[10] См. http://dictorpus.krc.karelia.ru

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 лицензией[11].

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

Подробнее об архитектуре (рис. 2) и возможностях ВепКар читайте в работе [Бойко и др., 2021], в докладе на конференции [Крижановская, Крижановский, 2020] и смотрите видеозапись на YouTube[12].

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

[11] См. страницу корпуса ВепКар с разрешениями авторов текстов: http://dictorpus.krc.karelia.ru/ru/page/permission
[12] См. видеозапись «Архитектура корпуса ВепКар и диалектные особенности» https://www.youtube.com/watch?v=cUpqM97LXGs

1.2.3. Страница с текстом в корпусе ВепКар

Страница с текстом (рис. 3) в зависимости от информации о тексте (т. е. 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):
    - автор перевода
    - заголовок
    - текст, выровненный по предложениям с исходным.

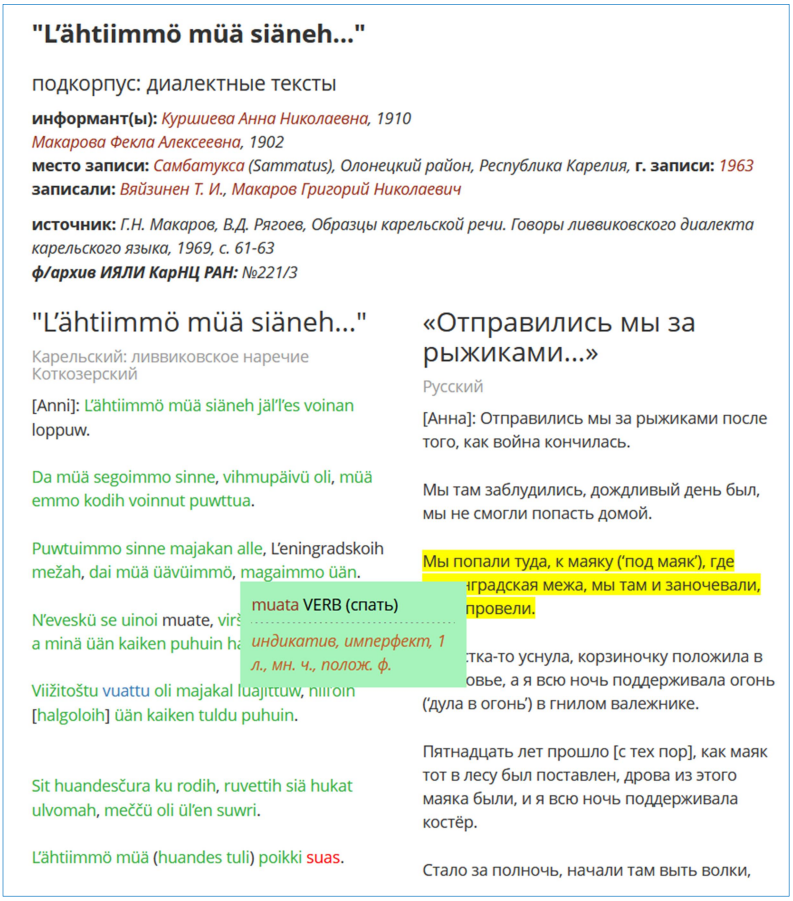


Рис. 3. *Фрагмент ливвиковского диалектного текста "L'ähtiimmö müä siäneh..."*[13]

[13] Полный текст см. http://dictorpus.krc.karelia.ru/ru/corpus/text/1514

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

Если кликнуть на размеченное слово в вепсском/карельском тексте, откроется окно с леммой, значением и грамматическими свойствами. Каждое автоматически размеченное 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. Черный цвет означает, что в словаре пока нет соответствия для слова в тексте.

1.2.4. Словарная статья в корпусе ВепКар

На сайте корпуса ВепКар организован словарь, по которому можно производить поиск (рис. 33). Щелкув на найденное слово, можно перейти к словарной статье. В словаре ВепКар 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 ВепКар. После ручной разметки (экспертной проверки) эти примеры помечены звездочками 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.

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

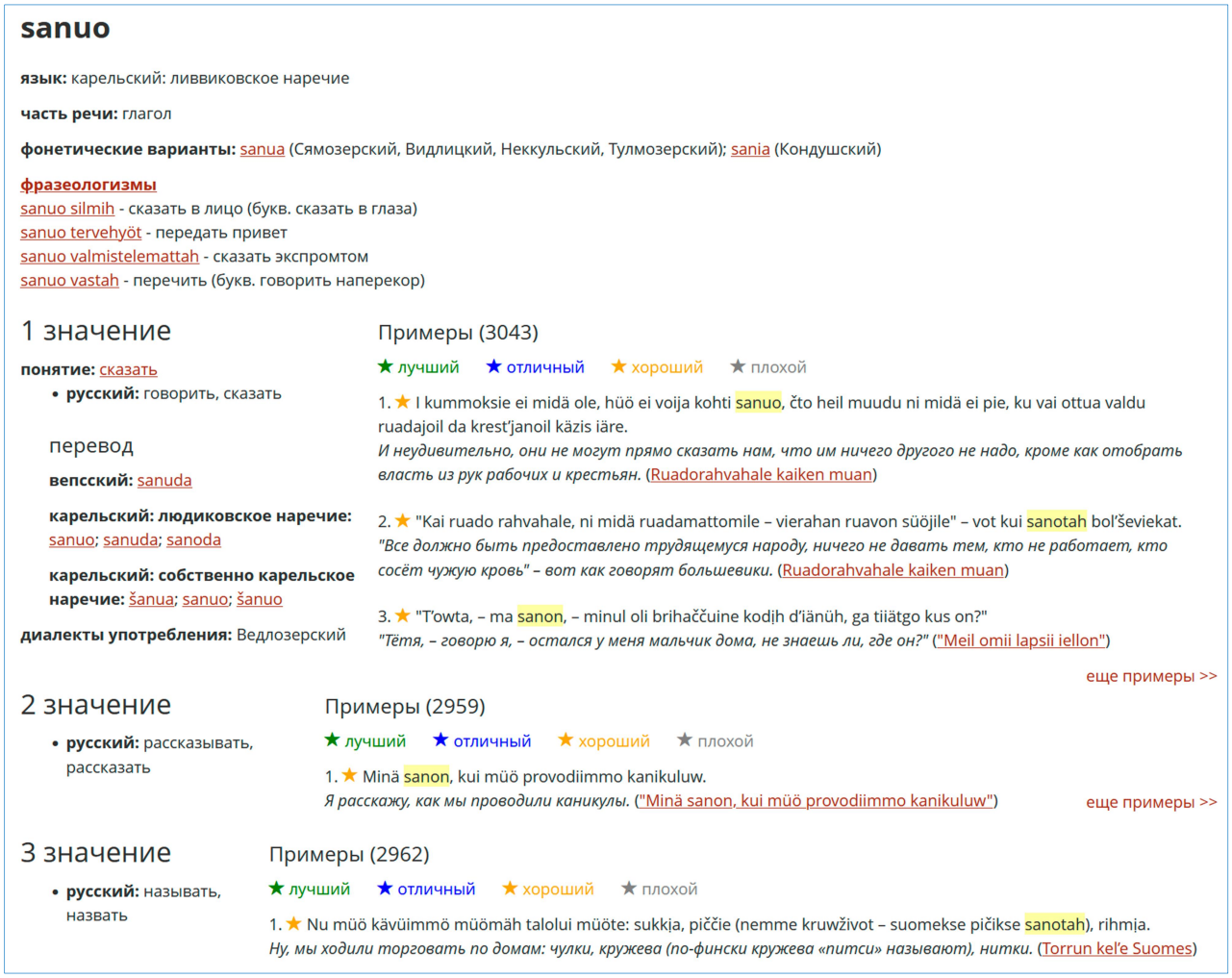


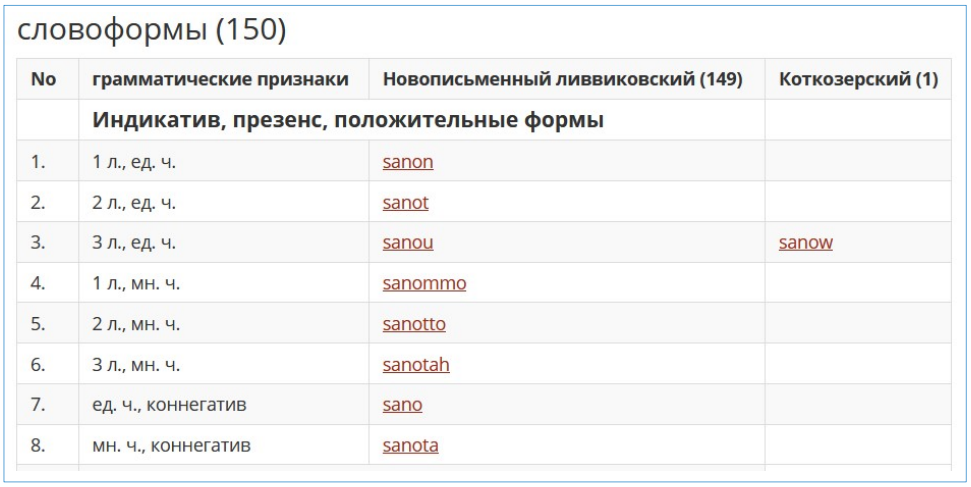
Рис. 5. *Фрагмент словарной статьи ливвиковского глагола 'sanuo' с тремя значениями*[14]

Словарная статья (см. рис. 5) в зависимости от степени ее 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;

[14] Полную статью см. http://dictorpus.krc.karelia.ru/ru/dict/lemma/15190

- семантические отношения с другими леммами в данном языке: синонимы, антонимы, гиперонимы, гипонимы, холонимы, миронимы и т. д.;
- диалекты, в которых употребляется данное значение;
- 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.	sanon	
2.	2 л., ед. ч.	sanot	
3.	3 л., ед. ч.	sanou	sanow
4.	1 л., мн. ч.	sanommo	
5.	2 л., мн. ч.	sanotto	
6.	3 л., мн. ч.	sanotah	
7.	ед. ч., коннегатив	sano	
8.	мн. ч., коннегатив	sanota	

Рис. 6. *Фрагмент словарной статьи ливвиковского глагола ‘sanuo’ с тремя значениями*[15]

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 дает возможность публиковать тексты (см. 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;

[15] Полную статью см. http://dictorpus.krc.karelia.ru/ru/dict/lemma/15190

- 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.

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 источниковой 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).

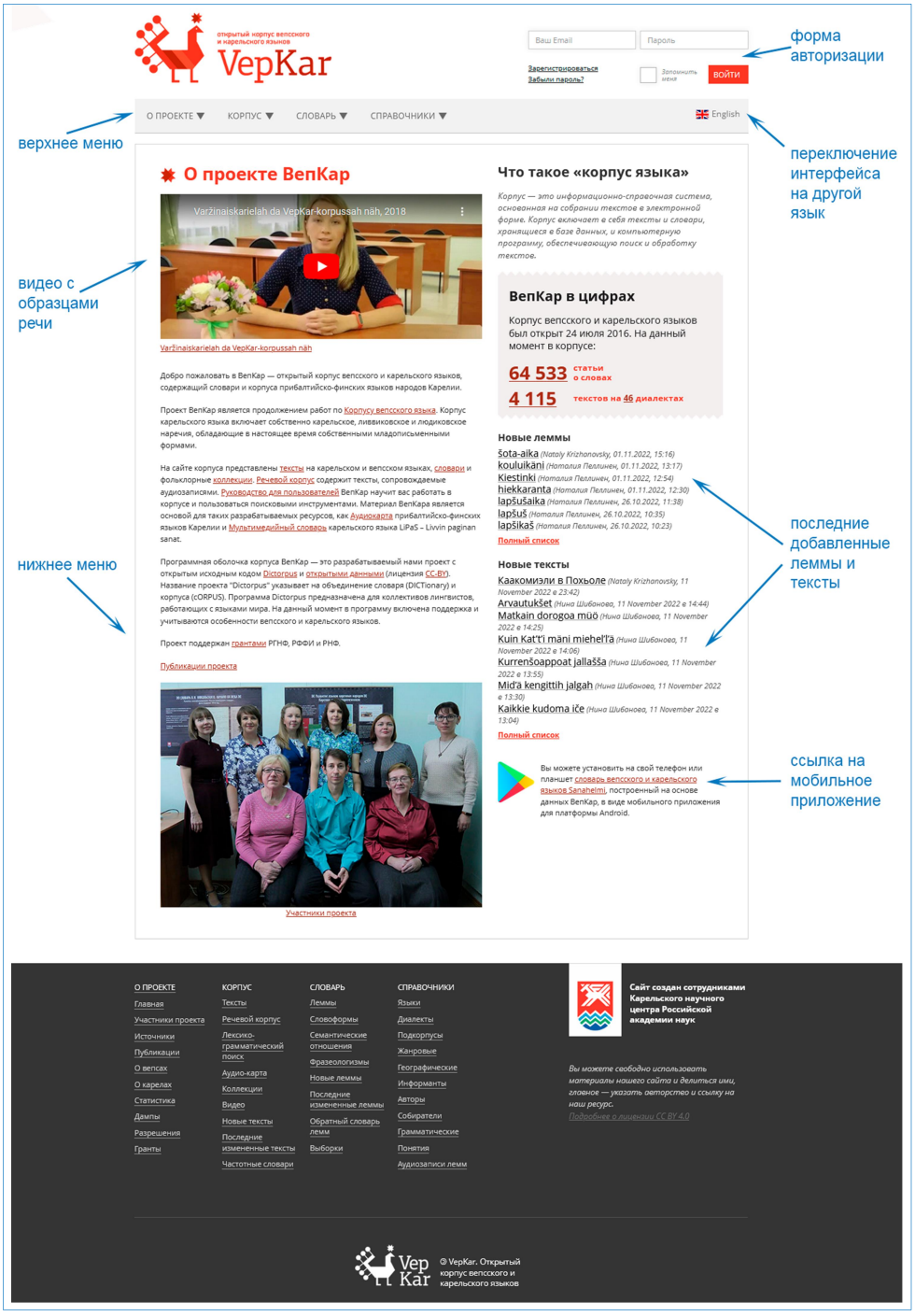


Рис. 7. *Стартовая страница сайта Открытого корпуса вепсского и карельского языков (ВепКар)*

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):

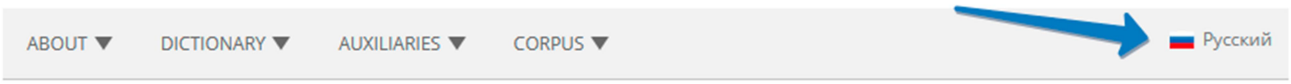


Рис. 8. *Переключение на русский язык ВепКар*

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 «тяжелых» 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 (ä ö ü č š ž ’), справа размещена иконка *«специальные символы»*: ä. 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.

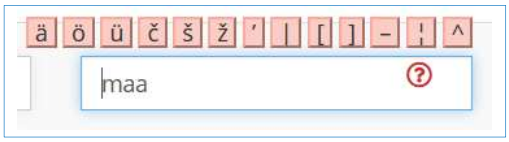


Рис. 9. *Открытая панель «специальные символы»*

> Стоит особо отметить, что в процессе поиска символ ü приравнен к символу y, w — к u / y / ü, 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 ‘el**änd**y’ или ‘vier**änd**ü’.

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 табл. 4 приведены шаблоны и примеры их использования для поиска строк.

Таблица 4. **Специализированные шаблоны для поиска текста**

Шаблон	Описание	Пример	Пояснение
^	начало строки	**^c**	все слова, начинающиеся на букву *c* (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?$**	*a* на конце может быть, а может не быть (Поиск лемм по словоформам)
*****	ноль или более символов, идущих перед звездочкой	**ab*c**	*abc* ИЛИ *ac* ИЛИ *abbbc*
+	один или более символов, идущих перед плюсом	**ab+c**	*abc* ИЛИ *abbc* ИЛИ *abbbc*
{n}	n символов, стоящих перед скобками	**b{3}**	*bbb*
{m,n}	от m до n символов, стоящих перед скобками	**a{1,3}**	*a* ИЛИ *aa* ИЛИ *aaa*
{m,}	предыдущий символ может встретиться m и более раз	**^pert{1,}i?$**	pert ИЛИ pertt ИЛИ perti ИЛИ pertti (Поиск лемм по словоформам)
(...)	круглые скобки задают группировку символов	**(abc){1,3}**	*abc* ИЛИ *abcabc* ИЛИ *abcabcabc*
p1|p2	p1 или p2	**^rein|^rien|^rejen|^regen**	все варианты слабоступенной основы существительного regi (Поиск лемм по словоформам)
V	любая гласная буква	**ttVV$**	все собственно карельские инфинитивы одноосновных глаголов, содержащих в основе чередование tt : t (Поиск лемм по словоформам)
C	любая согласная буква	**^kaCrV$**	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 щелкните по ссылке *«Тексты»* (рис. 10).

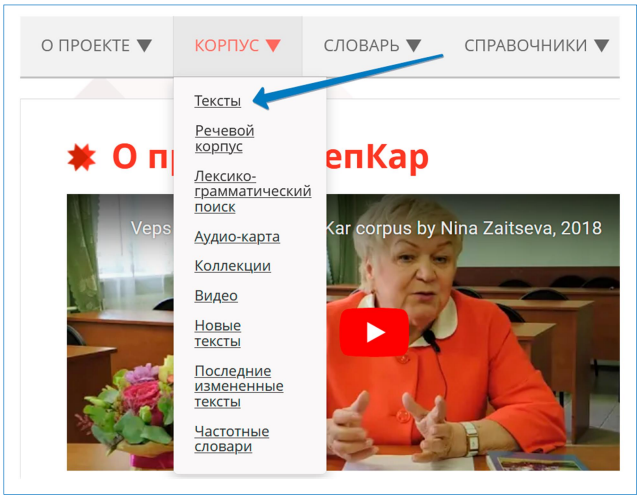


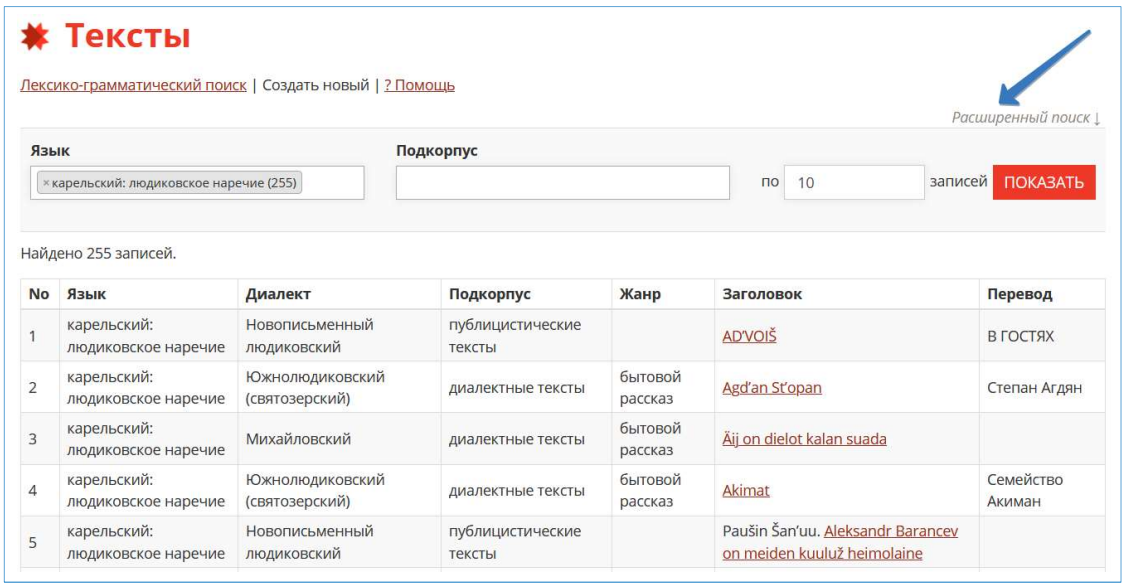
Рис. 10. *Поиск в корпусе ВепКар*

Шаг 8. Как найти текст на вепсском / 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.



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	Agd'an 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 ВепКар. Пример найденных людиковских текстов*

Шаг 9. Как осуществлять расширенный поиск по текстам ВепКар?

- Выполните шаг 7.
- Щелкните на ссылку *«расширенный поиск»* справа вверху над поисковой формой (см. 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.

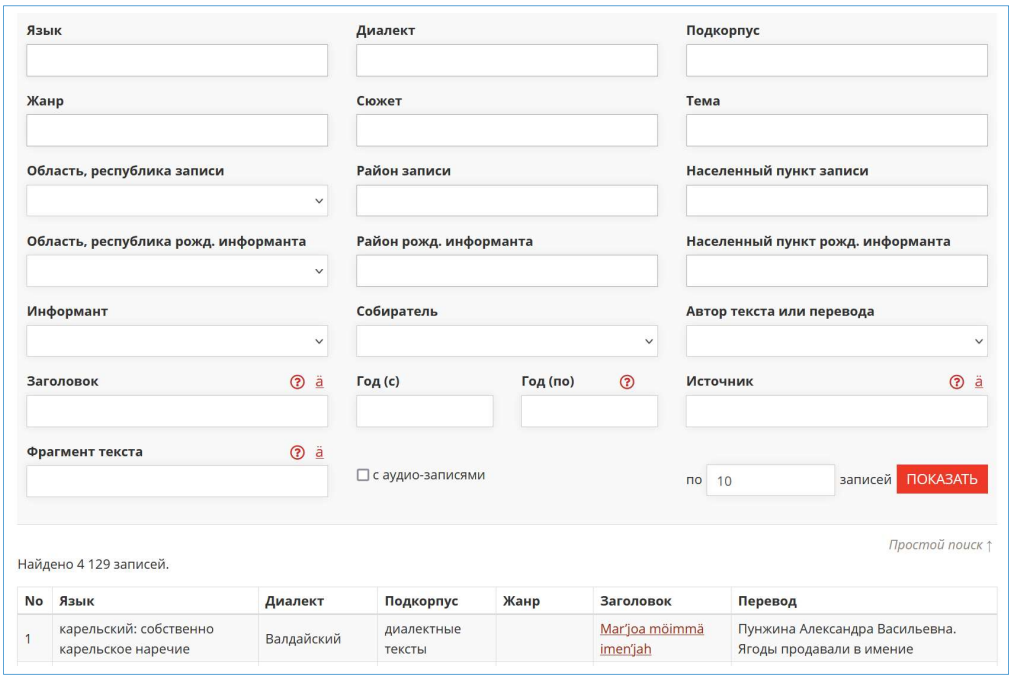


Рис. 12. *Расширенный поиск текстов в корпусе ВепКар*

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 аудиозаписями»*.
- Нажмите кнопку ПОКАЗАТЬ.

Загрузится таблица с текстами (рис. 13).

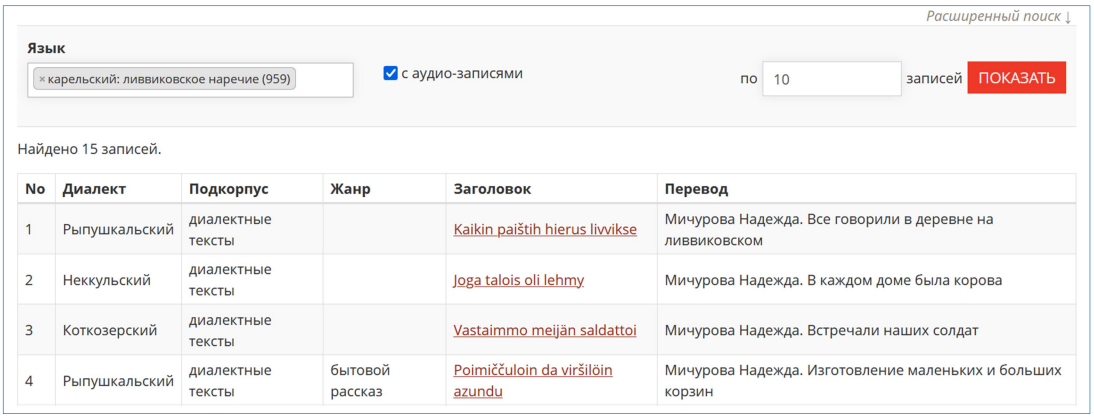


Рис. 13. *Поиск в корпусе ВепКар. Пример найденных ливвиковских текстов с аудиозаписями*

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.

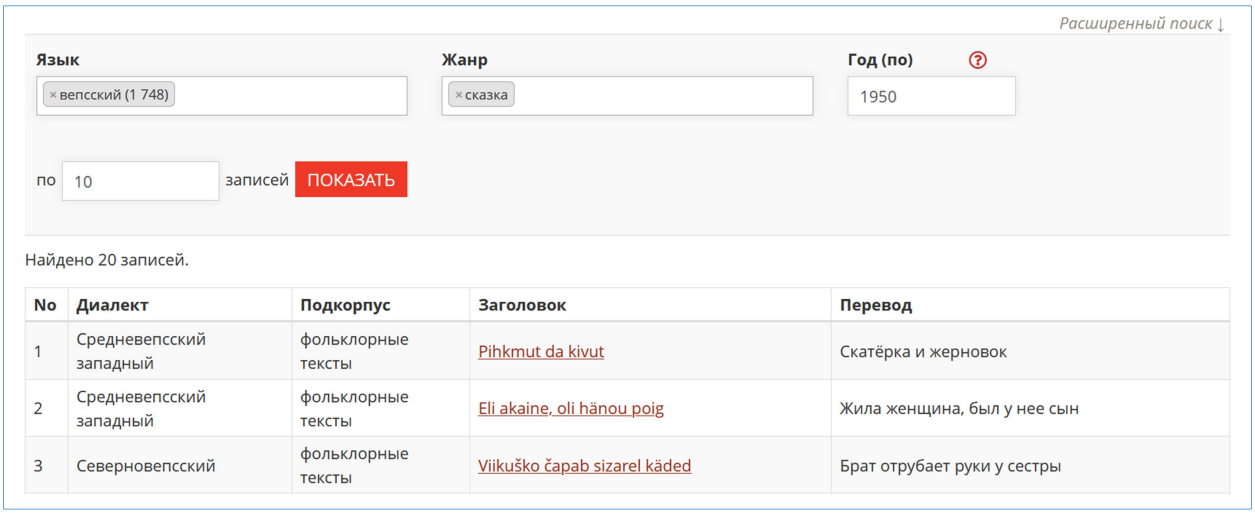


Рис. 14. *Поиск вепсских сказок, записанных до 1950 года*

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

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

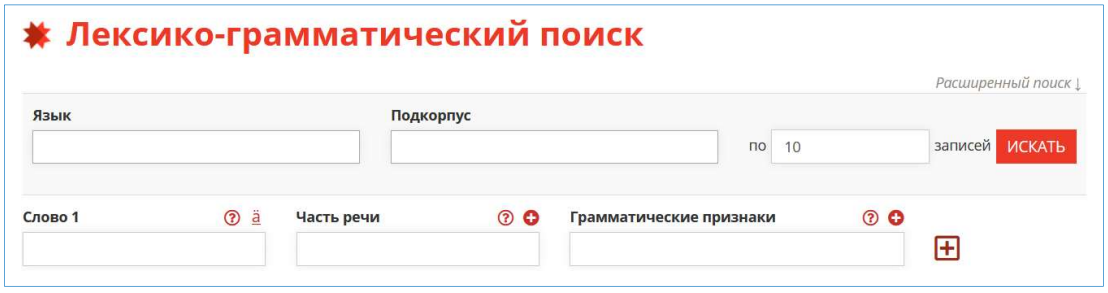


Рис. 15. *Поиск в корпусе заданной последовательности лемм и/или словоформ, обладающих определенными грамматическими характеристиками*

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 *^vuozi$* будут найдены примеры вхождения *vuvvel, vuodel,* и т. д. Если нужно найти только *vuvvel*, задайте эту форму в кавычках: *"^vuvvel$"*.

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).

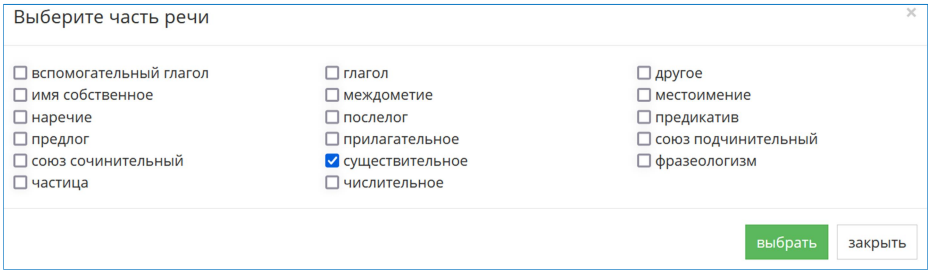


Рис. 16. *Выбор части речи «существительное» в дополнительном окне*

- 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.

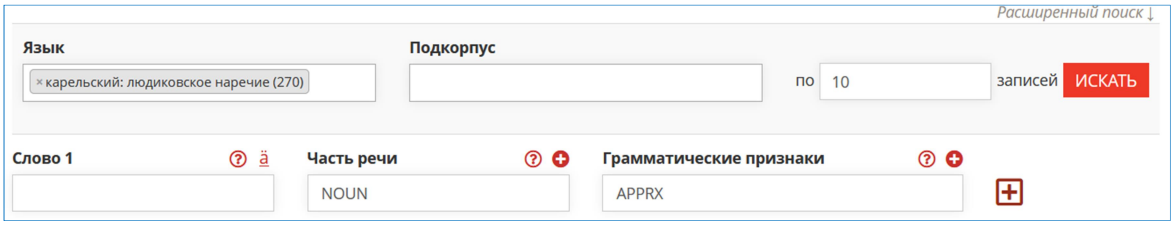


Рис. 17. *Поиск в корпусе примеров употребления людиковских имен в форме аппроксиматива*

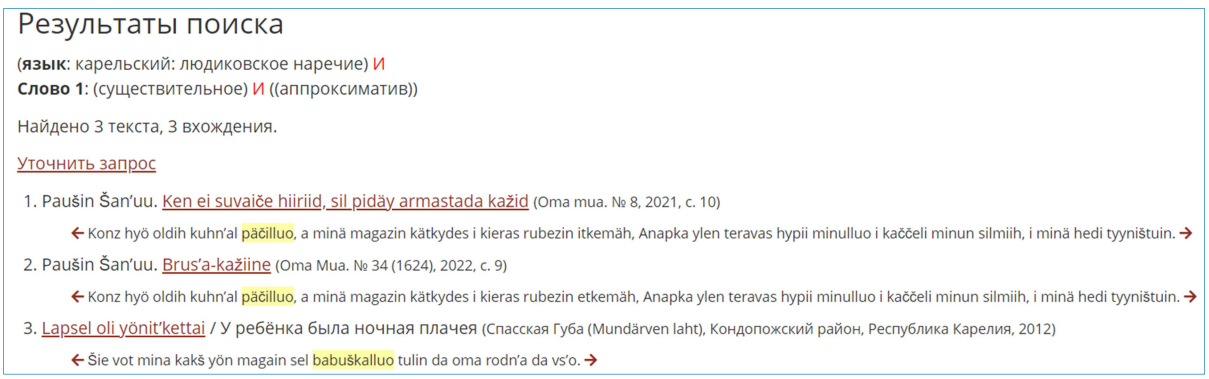


Рис. 18. *Результаты поиска примеров людиковских имен в форме аппроксиматива*

Шаг 14. Как найти примеры употребления в корпусе ливвиковских 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 *выбрать*.
- В поле *«Грамматические признаки»* первого слова нажмите на иконку ⊕.
- В дополнительном окне отметьте галочкой наклонение *«кондиционал»*.
- Нажмите кнопку *выбрать*.
- 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 *выбрать*.
- В поле *«Грамматические признаки»* второго слова нажмите на иконку ⊕.
- В дополнительном окне отметьте галочками залог *«актив»* и *«2-е причастие»*.
- Нажмите кнопку *выбрать*.
- В основной форме (рис. 19) нажмите кнопку ИСКАТЬ.

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.

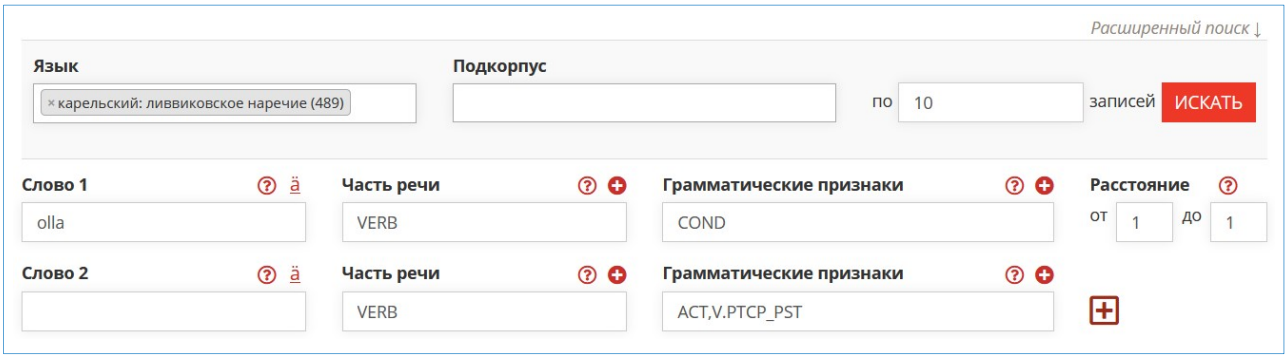


Рис. 19. *Поиск в корпусе примеров употребления ливвиковских глаголов в форме перфекта и плюсквамперфекта кондиционала*

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

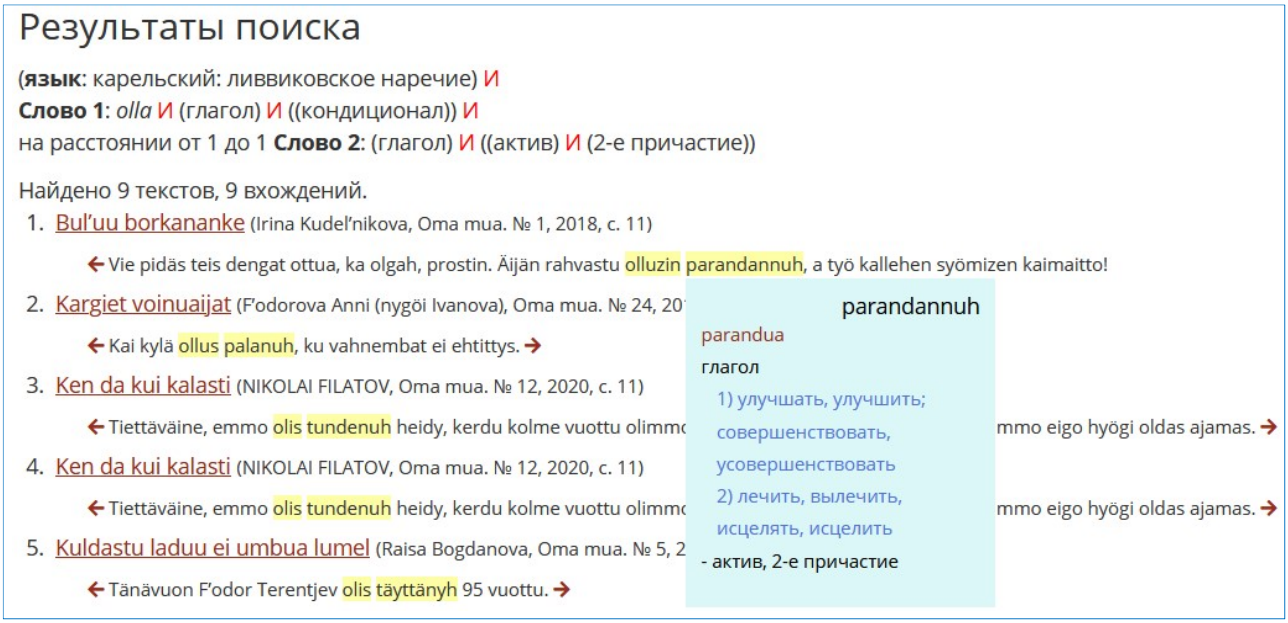


Рис. 20. *Результаты поиска примеров ливвиковских глаголов в форме перфекта и плюсквамперфекта кондиционала*

Шаг 15. Как найти примеры использования глаголов с существительными в определенных падежах, например, в партитиве?

- Выполните шаг 12.
- В поле *«Часть речи»* первого слова нажмите на иконку ⊕.
- В дополнительном окне (рис. 16) отметьте галочкой *«глагол»*.
- Нажмите кнопку *выбрать*.
- 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 *«существительное»*.
- Нажмите кнопку *выбрать*.
- В поле *«Грамматические признаки»* второго слова нажмите на иконку ⊕.
- В дополнительном окне отметьте галочкой падеж *«партитив»*.
- Нажмите кнопку *выбрать*.
- В основной форме нажмите кнопку ИСКАТЬ.

Такой запрос является достаточно «тяжелым», и если загрузка результата затянулась на несколько минут, попробуйте уточнить запрос. Выбор языка и других параметров тексту ускоряет процесс поиска примеров. Например, нас интересуют только ливвиковские фольклорные тексты.

- В поле *«Язык»* выберите *«карельский: ливвиковское наречие»*.
- В поле *«Подкорпус»* выберите *«фольклорные тексты»*.
- В основной форме (рис. 21) нажмите кнопку ИСКАТЬ.

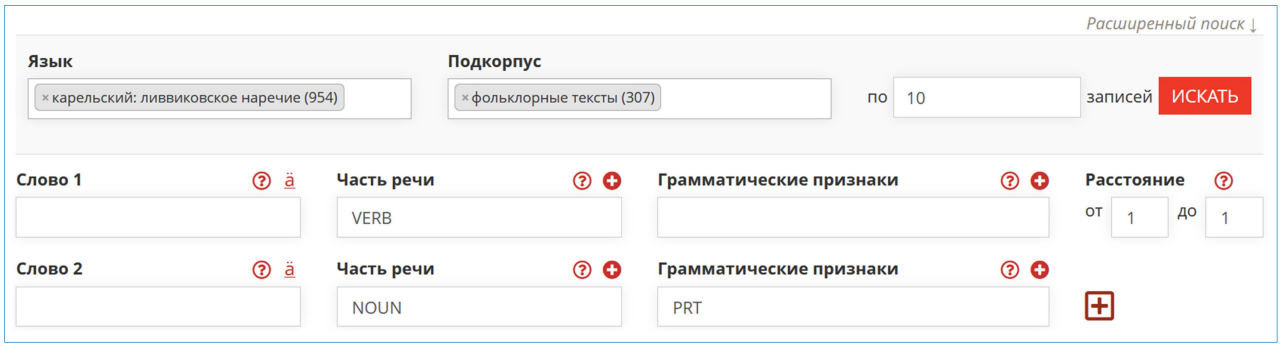


Рис. 21. *Поиск в корпусе примеров употребления глаголов с существительными в партитиве*

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

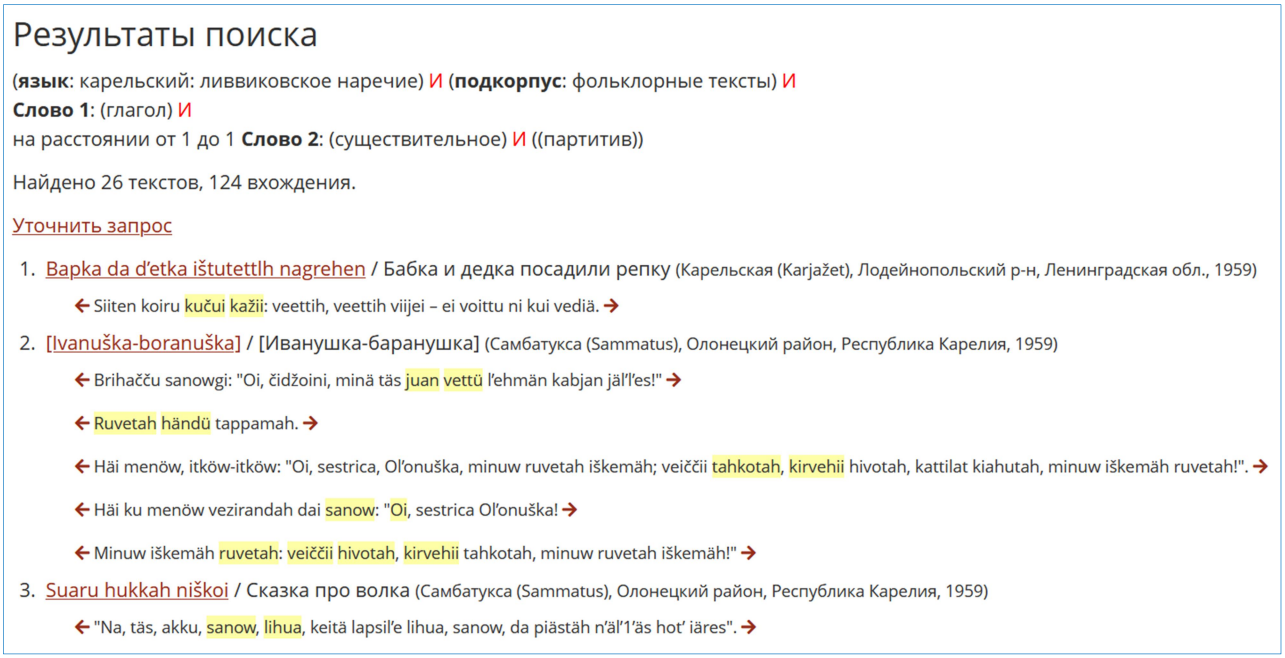


Рис. 22. *Результаты поиска в корпусе примеров употребления глаголов с существительными в партитиве*

Шаг 16. Как найти примеры употребления конкретной словоформы в вепсских текстах 1930-х годов?

- Выполните шаг 12.
- В поле *«Выберите язык»* выберите *«вепсский»*.
- Нажмите *«Расширенный поиск»* (см. 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.

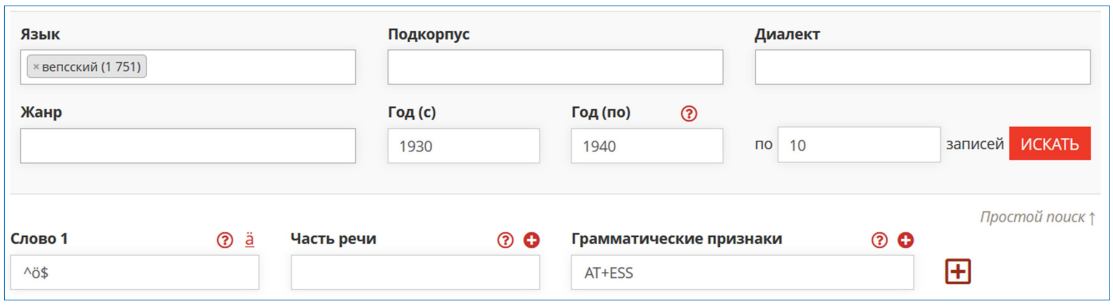


Рис. 23. *Поиск в корпусе примеров употребления вепсской лексемы ‘ö’ в форме адессива в текстах 1930-х годов*

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

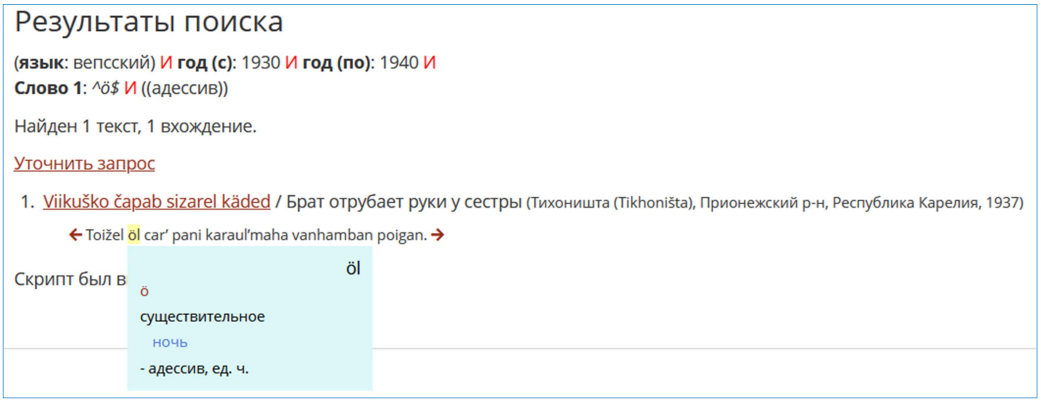


Рис. 24. *Результаты поиска в корпусе примеров употребления вепсской лексемы ‘ö’ в форме адессива в текстах 1930-х годов*

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 щелкните по ссылке *«Частотные словари»*.

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

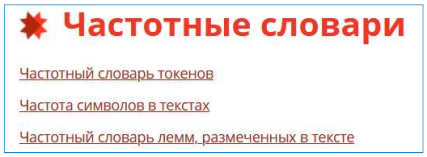


Рис. 25. *Страница с частотными словарями корпуса*

[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 табл. 4.

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).

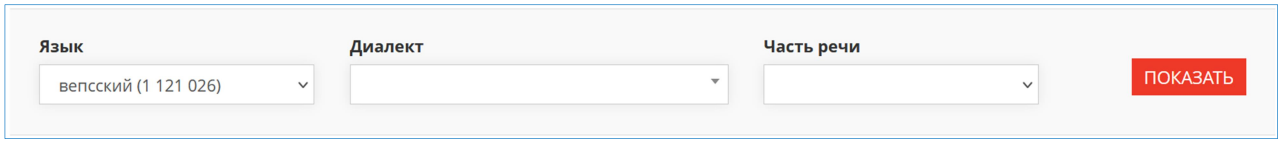


Рис. 26. *Частотный словарь лемм, размеченных в тексте*

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

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

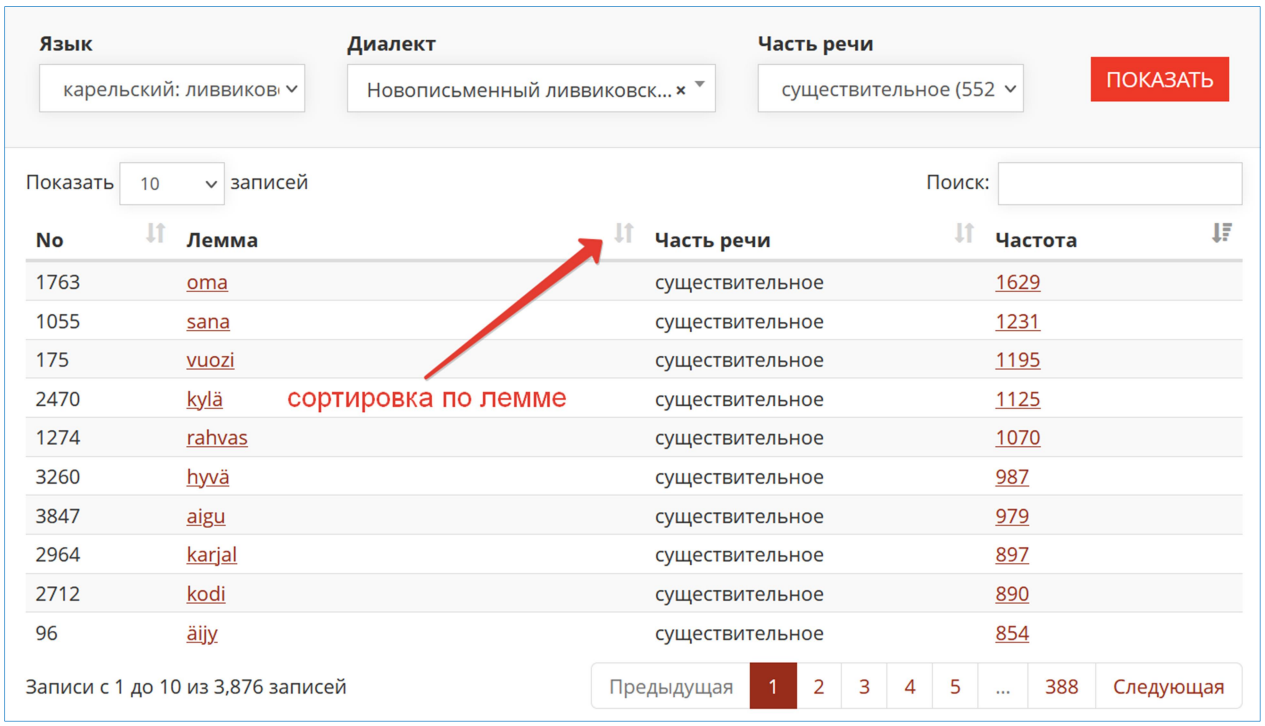


Рис. 27. *Частотный словарь ливвиковских существительных, размеченных в новописьменных текстах*

По умолчанию список отсортирован по частоте встречаемости слова в текстах. Можно менять сортировку, нажимая стрелочки СПРАВА от заголовка колонки (см. 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 щелкните по ссылке *«Речевой корпус»*.

Выведется список озвученных текстов (рис. 28).

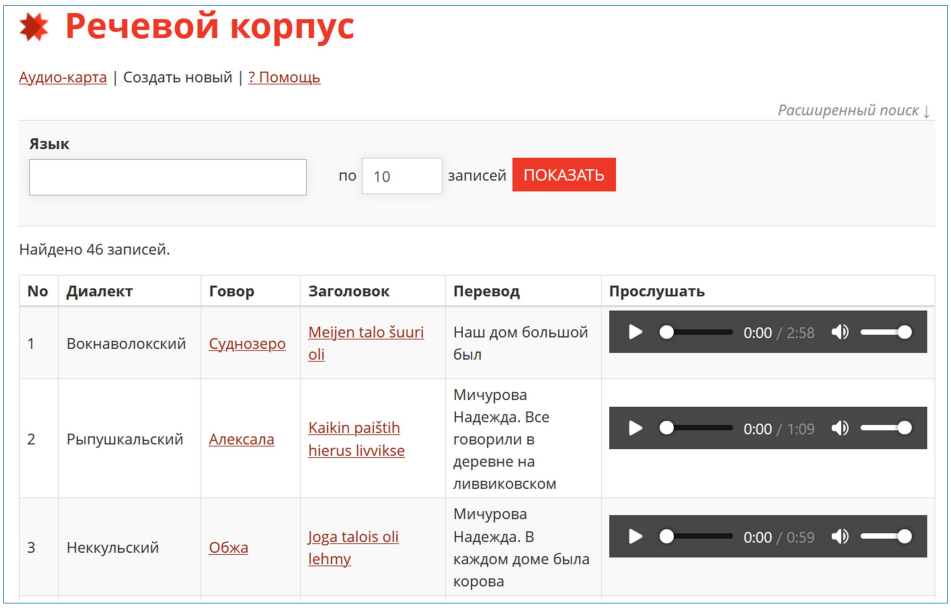


Рис. 28. *Речевой корпус прибалтийско-финских языков Карелии*

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

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

- Выполните шаг 19.
- Нажмите на ссылку *«Расширенный поиск»* (см. 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).

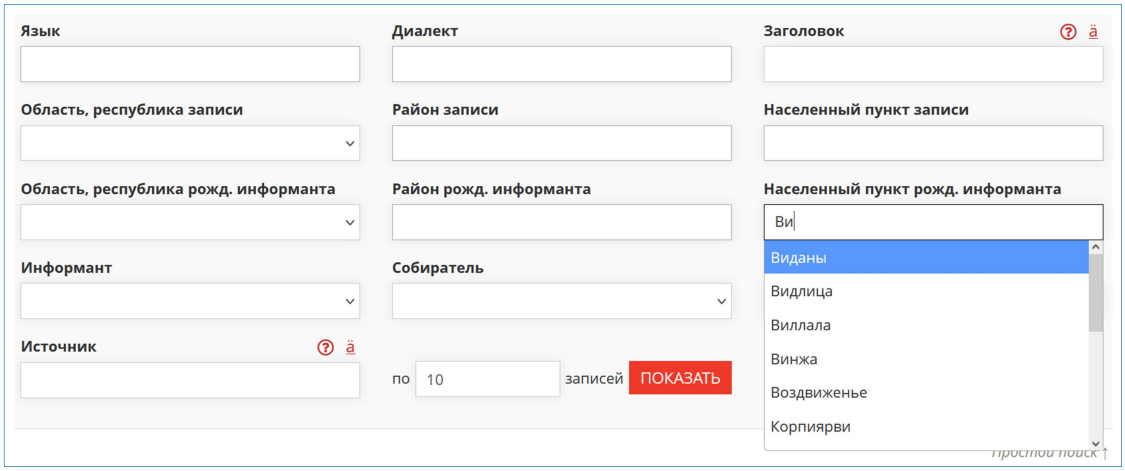


Рис. 29. *Выбор населенного пункта места рождения информанта*

Шаг 21. Как ссылаться на материалы корпуса ВепКар?

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

1. В тексте или в сноске указать:

   Исследование проведено на материале Открытого корпуса вепсского и карельского языков (dictorpus.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: dictorpus.krc.karelia.ru

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

   Lehme № 2051 // Открытый корпус вепсского и карельского языков ВепКар. URL: http://dictorpus.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://dictorpus.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 щелкните по ссылке *«Леммы»* или *«Словоформы»* для поиска в словаре соответственно лемм или словоформ (рис. 30).

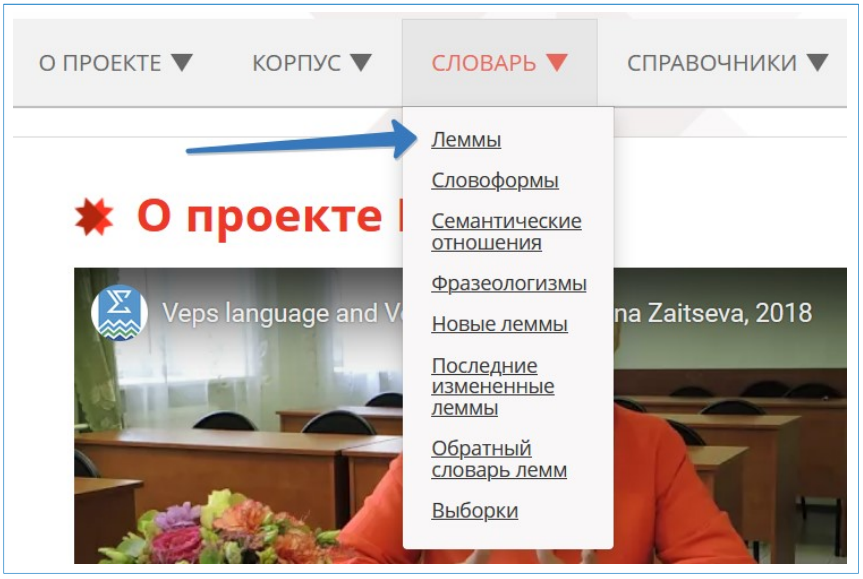


Рис. 30. *Поиск лемм в словаре ВепКар*

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).

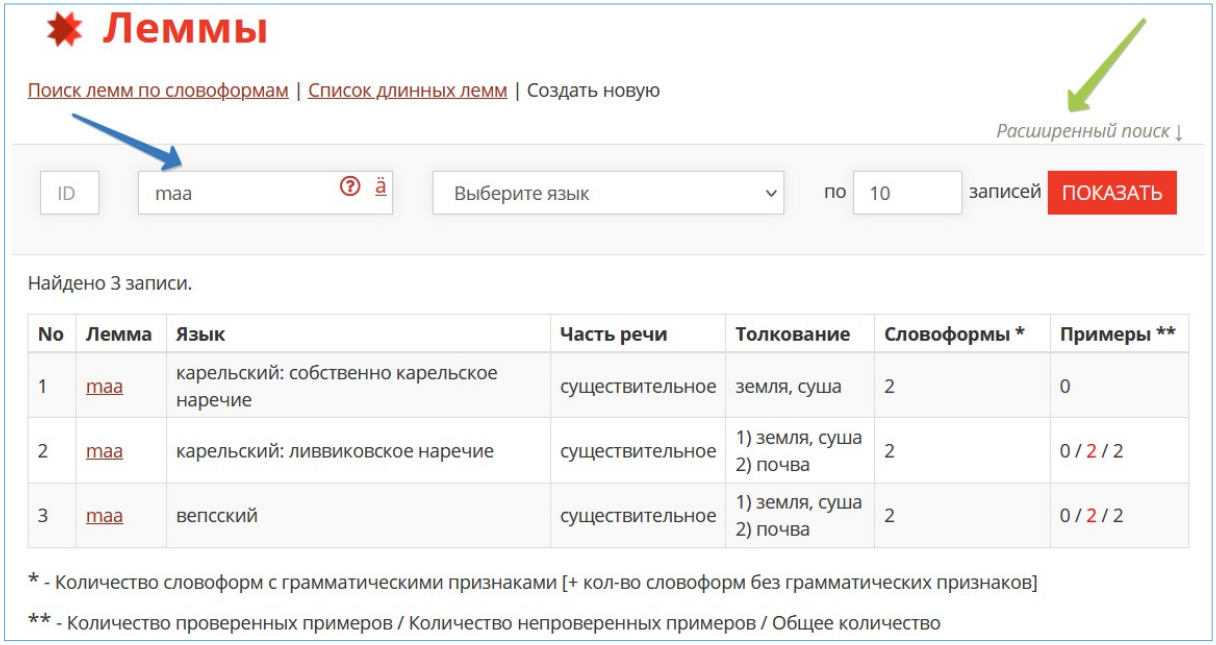


Рис. 31. *Поиск лемм в словаре ВепКар по начальной форме слова. Поле «лемма» с иконками «помощь» и «специальные символы». Ссылка «расширенный поиск» — для открытия полной формы*

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).

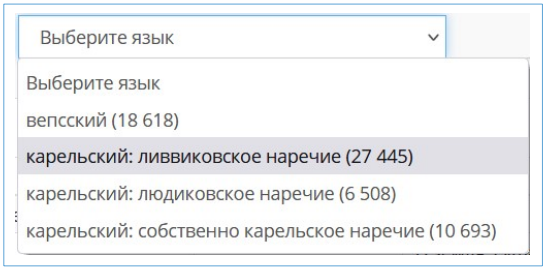


Рис. 32. *Выбор языка из выпадающего списка*

Шаг 25. Как осуществлять расширенный поиск в словаре ВепКар?

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

В форме откроются дополнительные поля (рис. 33).

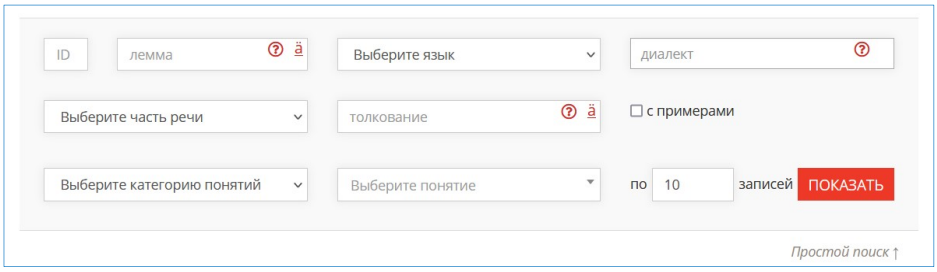
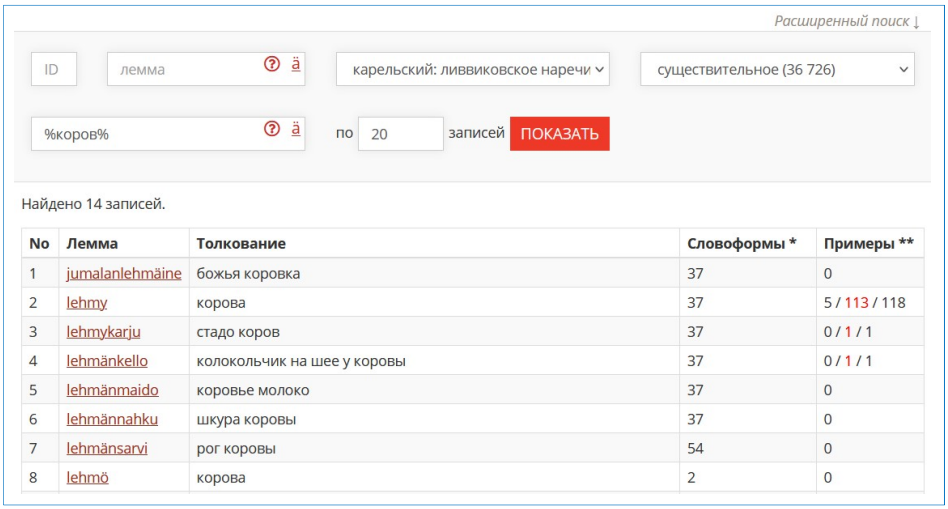


Рис. 33. *Расширенный поиск лемм в словаре ВепКар*

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

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



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 ВепКар 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] (см. 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 1»*.
- Если известны грамматические признаки слова, выберите нужное значение в списке *«грамматические признаки для словоформы»*.
- Нажмите кнопку ПОКАЗАТЬ.

[18] Поиск лемм по словоформам: http://dictorpus.krc.karelia.ru/ru/dict/lemma/by_wordforms

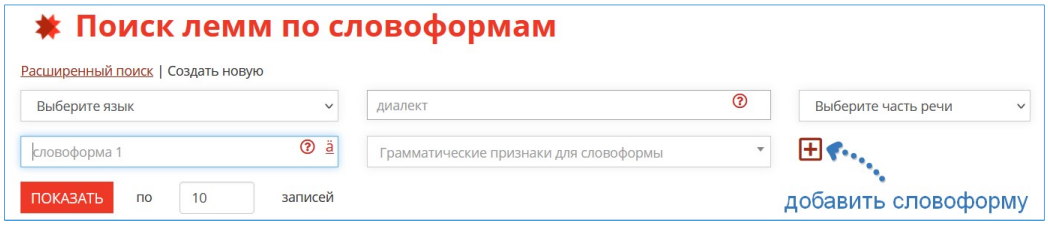
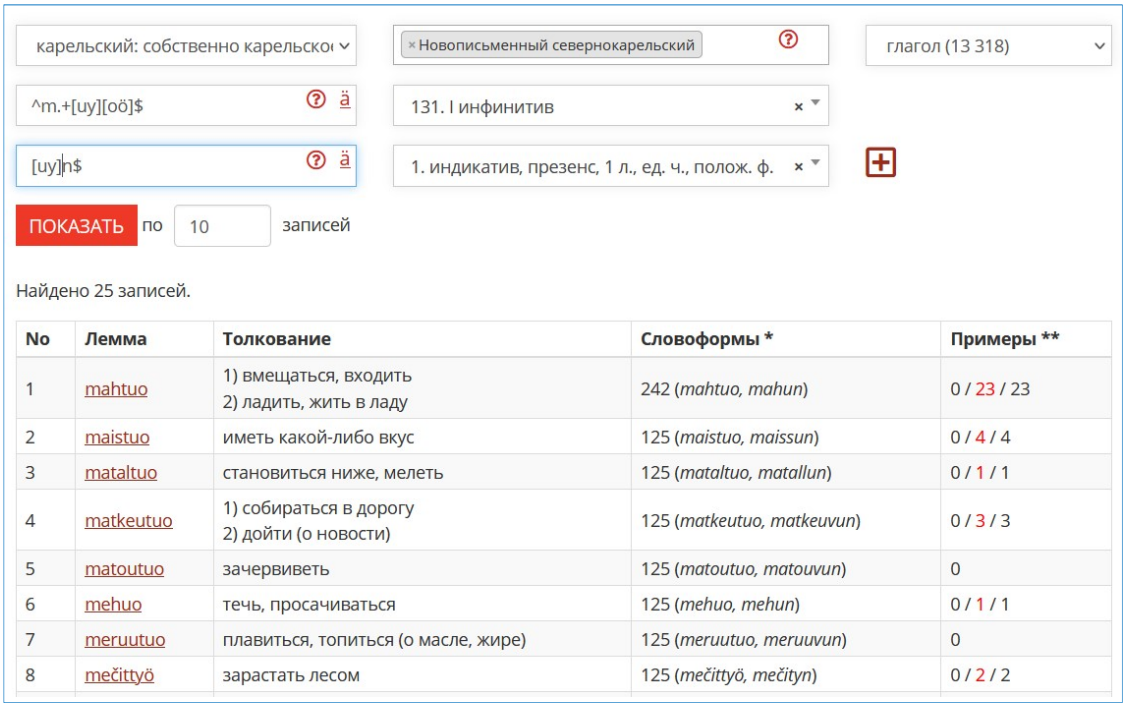


Рис. 35. *Поиск лемм в словаре ВепКар по словоформам*

Шаг 28. Как найти все глаголы новописьменного севернокарельского языка, начинающиеся с буквы ‘m’, у которых начальная форма (инфинитив) оканчивается на ‘uo’ или ‘yö’, а форма 1 лица ед. ч. настоящего времени оканчивается на ‘un’ или ‘yn’?

- 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]$»*.
- В поле справа выберите *«I инфинитив»*.
- Нажмите иконку *«добавить словоформу»* (см. рис. 35).
- В новом поле *«словоформа 2»* введите шаблон *«[uy]n$»*.
- В поле справа выберите *«индикатив, презенс, 1 л., ед. ч., полож. ф.»*.
- Нажмите кнопку ПОКАЗАТЬ (рис. 36).



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, meruuvun*)	0
8	mečittyö	зарастать лесом	125 (*mečittyö, mečityn*)	0 / 2 / 2

Рис. 36. *Пример поиска всех глаголов новописьменного севернокарельского языка, начинающихся с буквы ‘m’, у которых начальная форма (инфинитив) оканчивается на ‘uo’ или ‘yö’, а форма 1 лица ед. ч. настоящего времени оканчивается на ‘un’ или ‘yn’*

Шаг 29. Как найти обратный словарь?

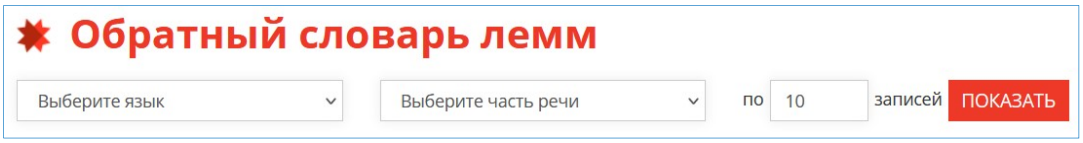
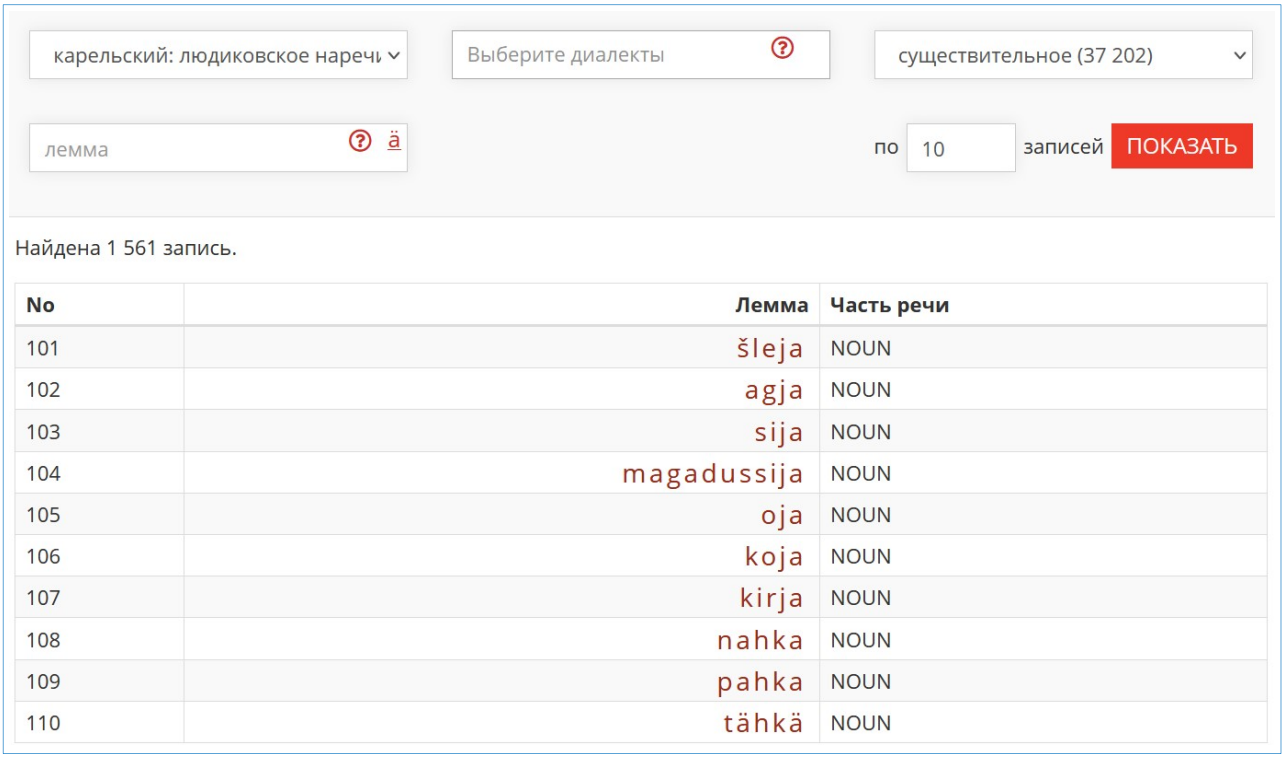


Рис. 37. *Поиск по обратному словарю лемм*

- Выполните шаги 1 и 2.
- В верхнем меню найдите пункт «СЛОВАРЬ» и щелкните на него мышкой.
- В выпадающем меню щелкните по ссылке «Обратный словарь лемм».
- В загрузившейся форме (см. 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.



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. *Результаты поиска по обратному словарю лемм*

Заключение

В этом руководстве для пользователей приведены основные понятия корпусной лингвистики, описаны 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 сфере 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.

Благодарности

Данное пособие было подготовлено в рамках проекта Российского научного фонда № 22-28-20215 «Создание речевого корпуса прибалтийско-финских языков Карелии», проводимого совместно с органами власти Республики Карелия с финансированием из Фонда венчурных инвестиций Республики Карелия (ФВИ РК).

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.

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

Лемма — 1) начальная форма слова; 2) словарная статья в словаре ВепКар.

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.

Литература

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

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

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, 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. 2022. — Vol. 13(2). — P. 47–78 DOI: 1012697/jeful.2022.13.2.02.

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, А. Н. Введение в прикладную лингвистику. — М., 2001. — Т. 2.

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. No 7. Комплексные научные исследования КарНЦ РАН. 2021. — С. 100–115 DOI: 10.17076/them1415

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).

Горина, О. Г. Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: дисс. … канд. пед. наук. — М., 2014. — Т. 13.

Зайцева, Н. Г. Вепсские 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, В. П., Богданова, С. Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.: Изд-во С.-Петерб. ун-та, 2020. — 234 с.

Копотев, М. В., Янда, Л. Национальный корпус русского языка // Вопросы языкознания. 2006. — № 5. — С. 149–155.

Копотев, Михаил. Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. — Прага, 2014.

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).

Плунгян, В. А., Сичинава, Д. В. Национальный корпус русского языка: опыт создания корпуса текстов современного русского языка // Труды Международной конференции «Корпусная лингвистика-2004». — СПб.: Изд-во С.-Петербургского ун-та, 2004. — С. 216–238. — URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2004/Sitchinava_art.pdf (дата обращения: 07.06.2021).

Словарь социолингвистических терминов. — М., 2006.