Труды Карельского научного центра РАН
№ 7. 2021. С. 100–115
**DOI: 10.17076/them1415**

*УДК 811.511*

# ЛИНГВИСТИЧЕСКИЙ КОРПУС ВЕПКАР – «ЗАПОВЕДНИК» ПРИБАЛТИЙСКО-ФИНСКИХ ЯЗЫКОВ КАРЕЛИИ

**Т. П. Бойко[1], Н. Г. Зайцева[1], Н. Б. Крижановская[2], А. А. Крижановский[2], И. П. Новак[1], Н. А. Пеллинен[1], А. П. Родионова[1], Е. Д. Трубина[3]**

*[1] Институт языка литературы и истории КарНЦ РАН, ФИЦ «Карельский научный центр РАН», Петрозаводск, Россия*
*[2] Институт прикладных математических исследований КарНЦ РАН, ФИЦ «Карельский научный центр РАН», Петрозаводск, Россия*
*[3] Институт математики и информационных технологий, Петрозаводский государственный университет, Россия*

Целью создания природных заповедников является охрана исчезающих видов флоры и фауны. Для сохранения и последующего изучения языков такими «заповедниками» становятся большие, размеченные, разножанровые лингвистические корпусы. В статье описаны история, структура, возможности и перспективы развития «Открытого корпуса вепсского и карельского языков», являющегося одновременно как результатом многолетней междисциплинарной работы лингвистов и программистов Карельского научного центра РАН, так и уникальной источниковой базой для новых исследований. Электронный ресурс ведет свою историю с 2009 года, когда под руководством Н. Г. Зайцевой был создан «Корпус вепсского языка». С 2016 года в корпус входят три карельских подкорпуса: собственно карельский, ливвиковский и людиковский. Объединенная лингвистическая платформа получила название «Открытый корпус вепсского и карельского языков» (ВепКар). Корпус включает в себя тексты и словари, хранящиеся в базе данных, и компьютерную программу, обеспечивающую поиск и обработку текстов. Эта программа называется «корпусным менеджером», она написана 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 (dictorpus.krc.karelia.ru). Авторы проекта уделяют внимание популяризации корпуса ВепКар с помощью сайтов YouTube и Википедия. Особенностью базы данных и самого корпуса ВепКар является тесная взаимосвязь словарей и текстов. Многофункциональные словари вепсского и карельского языков содержат толкование, 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. Разработчики добавляют в корпусный менеджер новые полезные функции, призванные облегчить работу редакторов. Например, за последние три года сформулированы и запрограммированы правила именного и глагольного словоизменения для всех диалектов вепсского языка и его младописьменного варианта, а также для ливвиковского, северокарельского и тверского новописьменных вариантов карельского языка. Благодаря этому в системе ВепКар в полуавтоматическом режиме сгенерировано 2,1 млн словоформ. Кроме семантической разметки, представленной в корпусе (2,1 млн связей между словами из текста и значениями лемм в словаре), добавлена грамма-

тическая разметка, позволившая автоматически установить 1,1 млн связей между словами из текста и грамматическими характеристиками словоформ из словаря. Многоязычный корпус ВепКар делится на подкорпусы 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. Эти категории появились благодаря интеграции выдающегося «Сопоставительно-ономасиологического словаря диалектов карельского, вепсского, саамского языков» в словарную часть ВепКар. На базе ВепКар в 2021 году был создан электронный словарь Sanahelmi для телефонов с операционной системой Android. Разработка мобильных приложений на основе данных корпуса будет продолжена.

Ключевые слова: карельский язык; вепсский язык; корпусная лингвистика; Открытый корпус вепсского и карельского языков; корпусный менеджер; словоизменительная парадигма.

**T. P. Boyko, N. G. Zaitseva, N. B. Krizhanovskaya, A. A. Krizhanovsky, I. P. Novak, N. A. Pellinen, A. P. Rodionova, E. D. Trubina. THE LINGUISTIC CORPUS VEPKAR IS A LANGUAGE REFUGE FOR THE BALTIC-FINNISH LANGUAGES OF KARELIA**

The purpose of creating conservation areas is to protect endangered plant and animal species. Large, tagged linguistic corpora with a great variety of genres are used for the preservation and research of safe and endangered languages. The article describes the history, structure and development of the Open Corpus of the Veps and Karelian languages. The Veps language corpus was created in 2009 under the leadership of Nina Zaitseva. Three Karelian subcorpora (Karelian proper, Livvi and Ludian) were included in the linguistic corpus in 2016. The united linguistic platform was named “The Open Corpus of the Veps and Karelian languages” (VepKar). This linguistic corpus includes texts and dictionaries stored in a database, and a computer program (corpus manager) for searching and processing the data. This corpus manager was written in the PHP programming language in the Laravel framework. The data are stored in a MySQL database. Corpus and dictionaries data are available online (dictorpus.krc.karelia.ru). YouTube and Wikipedia are used by VepKar authors to popularize the corpus. Dictionaries and corpus texts are strongly interrelated. Multifunctional dictionaries of the Veps and Karelian languages contain definition, translation, dialect labels, semantic relations (synonyms, antonyms, etc.), examples of word usage with reference to texts, as well as complete inflectional paradigms. All texts are automatically marked up and there are references from words in the text to the corresponding meanings in the dictionary entries. The developers continue adding useful new features to the corpus manager to make the work of editors easier. For example, over the past three years, nominal and verbal inflection rules have been formulated and programmed for all dialects of the Veps language and its newly-written version, as well as for the Livvi-Karelian, North Karelian and Tver newly-written versions of the Karelian language. Thanks to this, 2.1 million word forms were generated in the VepKar system in a semi-automatic mode. The semantic markup in the corpus is 2.1 million links between words from the text and the meanings of lemmas in the dictionary. The grammatical markup was added, namely, 1.1 million links between words from the text and the grammatical features of word forms from the dictionary were automatically established. The multilingual VepKar corpus is divided into subcorpora according to languages and dialects, and the texts are also classified into styles and genres. The corpus has a sophisticated search system (with filtering of texts by language, style and dialect, by informant, collector or author, by year of recording or year of publication). It is possible to search for lemmas by dialects, parts of speech, grammatical features, and even by lexical-semantic categories. These categories appeared due to the integration of the data of the outstanding “Comparative and Onomasiological Dictionary of the Dialects of the Karelian, Veps and Sami Languages” into the vocabulary part of VepKar. In 2021, the Sanahelmi electronic dictionary was created on the basis of VepKar for Android phones. The development of mobile applications based on corpus data is our bright future.

Keywords: Karelian language; Veps language; corpus linguistics; Open corpus of Veps and Karelian languages; corpus manager; inflectional paradigm.

## Введение

Как известно, заповедники предназначены для охраны природы и сохранения генетического фонда и имеют культурное и научное значение. Так же, как редкие и малочисленные виды животных и растений, многие языки мира находятся под угрозой исчезновения. Для сохранения языкового богатства и последующего изучения языков создаются лингвистические корпусы. Исследованием вопросов построения корпусов и методами их обработки занимается такой раздел языкознания, как корпусная лингвистика.

В мире существуют десятки больших лингвистических корпусов, например, Национальный корпус русского языка[1], Британский национальный корпус[2], Чешский национальный корпус[3]. Известны три наиболее крупных корпуса финно-угорских языков: Языковой банк Финляндии[4], Сводный корпус эстонского языка[5] и Венгерский национальный корпус[6].

Для долговременной и плодотворной работы над корпусом требуется научный коллектив и включение корпусных исследований в план научной работы. Например, для работ над Чешским национальным корпусом был организован одноименный институт, над пополнением и развитием Национального корпуса русского языка работают сотрудники нескольких университетов и институтов РАН. Наличие больших коллективов лингвистов в процессе работы над корпусом необходимо для выполнения его тонкой настройки. К такой настройке следует отнести ручную разметку, создание так называемого «золотого стандарта» или размеченной вручную части корпуса, которая в дальнейшем будет использоваться в различных экспериментах. Примером может служить глубоко аннотированный корпус текстов русского языка СинТагРус, где каждое слово в тексте привязано к какой-либо словарной статье комбинаторного словаря [Иншакова и др., 2019].

Большой теоретический и практический интерес представляет собой веб-корпус уральских языков Поволжья[7], построенный 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. Корпус включает современную художественную литературу, научные статьи, переводы Библии, статьи Википедии, официальные тексты и публичные записи в социальных сетях [Arkhangelskiy, 2020, с. 58–61].

Для каждого из пяти уральских языков разработан морфологический анализатор на основе правил. Анализаторы работают по данным словарей, поэтому в корпусах не распознаны несловарные слова или слова с опечатками. Проанализировано от 80 до 96 % слов в корпусах. Контекст при морфологическом анализе практически не учитывается, и анализатор выдает все возможные леммы для заданной словоформы. Для малоресурсных языков как раз важно, чтобы при автоматическом морфологическом анализе в разметке сохранялись все возможные формы слова для последующей проверки и выбора правильной формы лингвистом. Т. Архангельский называет корпусные менеджеры, сохраняющие все формы разборов, «дружественными к неоднозначности», такой “ambiguity-friendly” платформой является при разметке корпус ВепКар (рис. 1).

## История «Открытого корпуса вепсского и карельского языков» и популяризация проекта

Проект «Открытый корпус вепсского и карельского языков» (ВепКар) [Зайцева, 2012] ведет свою историю с 2009 года, когда в Институте языка, литературы и истории Карельского научного центра РАН (ИЯЛИ КарНЦ РАН) под руководством д. фил. н. Н. Г. Зайцевой началась работа по созданию «Корпуса вепсского языка»[8]. Корпус, включающий в себя электронный словарь и пять текстовых корпусов (диалектные тексты; фольклорные тексты с двумя подкорпусами: причитания и сказки; два младописьменных корпуса: переводы Нового Завета и тексты на младописьменном вепсском языке), содержит собственные системы поиска по различного рода характеристикам: по отдельным словам, диалектам, жанрам фольклора и младописьменных текстов и т. д. [Зайцева, 2012, с. 16]. «Корпус вепсского языка» оказал большую помощь в подготовке «Орфографического словаря вепсского языка» [Зайцева, Харитонова, 2012] языковедам, поскольку смог обеспечить возможность поиска в диалектном материале при отборе форм для создания правил орфографии вепсского языка [Крижанов-

[1] См. https://ruscorpora.ru
[2] См. http://www.natcorp.ox.ac.uk
[3] См. https://www.korpus.cz
[4] См. https://www.kielipankki.fi/aineistot/ftc
[5] См. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=en
[6] См. http://mnsz.nytud.hu/index_eng.html
[7] См. http://volgakama.web-corpora.net
[8] См. http://vepsian.krc.karelia.ru/about/

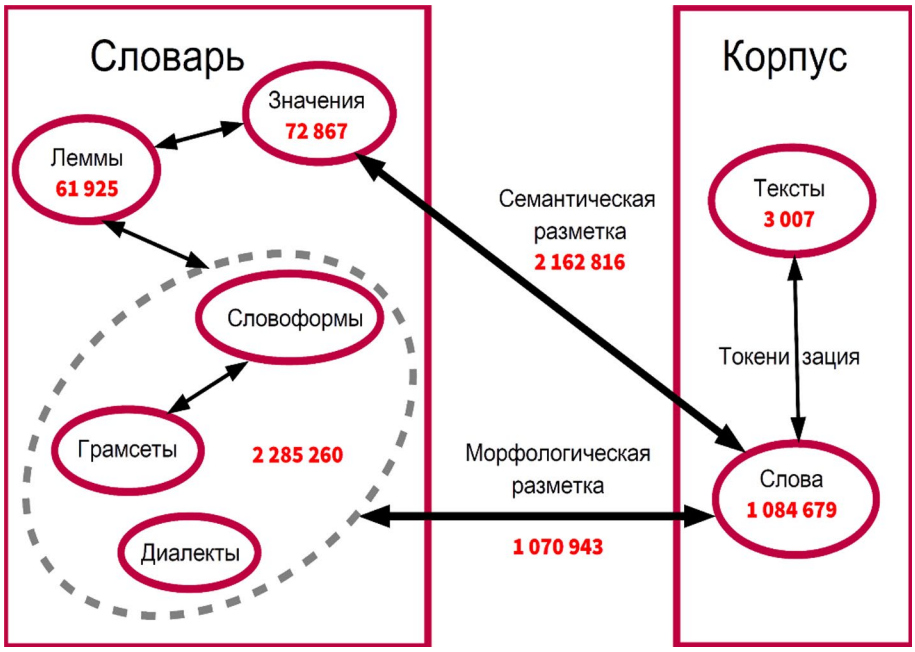



*Рис. 1*. Организация данных ВепКар, объем корпуса и словаря на февраль 2021 года

*Fig. 1*. VepKar data architecture, volume of the corpus and dictionary as of February 2021

ский, 2019, с. 20]. Авторы корпуса в 2017 году получили премию «Ключевое слово», учрежденную Федеральным агентством по делам национальностей, в номинации «За лучший научный проект»[1].

Представителями вепсского и карельского народов решаются идентичные проблемы по возрождению языков и культур, и все это требует больших электронных языковых ресурсов, которые позволят исследователям создавать и развивать, например, правила карельской орфографии, введут 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 как вепсского, так и карельского языка в 2016 году сотрудники ИЯЛИ и Института прикладных математических исследований КарНЦ РАН приступили к созданию многоязычного корпуса. В 2016 году на базе вепсского корпуса создан карельский корпус. Иными словами, объединенный корпус стал продолжением «Вепсского корпуса» и получил название «Открытый корпус вепсского и карельского языков» (ВепКар)[2]. Направление корпусной лингвистики было включено в план научно-исследовательской работы сектора языкознания ИЯЛИ КарНЦ РАН на 2021–2023 годы, что указывает на значительные результаты и еще большие надежды, возлагаемые на корпус [Зайцева, Крижановская, 2018].

Современные научные проекты решают задачи популяризации с помощью таких инструментов, как YouTube и Википедия. Редакторы ВепКар, они же сотрудники ИЯЛИ КарНЦ РАН, представили небольшие видеорассказы на исследуемых языках. В корпусе ВепКар появилась возможность добавлять видео к текстам. Такие тексты с видеозаписью информанта собраны в отдельный раздел сайта «Видео»[3]. Помимо сайта проекта ВепКар эти видеорепортажи теперь можно увидеть на канале YouTube[4] и в Русской Википедии (статьи «Ливвиковское наречие», «Вепсский язык», «Зайцева, Нина Григорьевна», «Тверской диалект карельского языка»), в Английской Википедии (статьи “Veps language”, “Livvi-Karelian language”, “Karelian language”). Видеофайлы доступны на Викискладе в категории VepKar[5].

[1] См. http://fadn.gov.ru/news/2017/10/03/3406-na-forume-yazykovaya-politika-organizovannom-fadn-rossii-nagradili-laureatov-vserossiyskoy-premii-klyuchevoe-slovo
[2] См. dictorpus.krc.karelia.ru
[3] См. http://dictorpus.krc.karelia.ru/ru/corpus/video
[4] См. https://www.youtube.com/playlist?list=PLURxZmJJsseOSJ8upo1sSWztSu0-ou_ys
[5] См. https://commons.wikimedia.org/wiki/Category: VepKar

Леммы и словоформы с наборами грамматических признаков вепсского и карельских словарей корпуса ВепКар экспортированы и включены в международную морфологическую базу данных UniMorph 3.0. Эти данные доступны на сайте GitHub в виде отдельных подпроектов проекта UniMorph. В качестве имен подпроектов выбраны языковые коды по стандарту ISO 639–3, а именно: *vep* для вепсского языка[1], *krl* – собственно карельское наречие[2], *olo* – ливвиковское наречие карельского языка[3], *lud* – людиковское наречие карельского языка[4].

Результаты исследований, связанные с корпусом ВепКар, были представлены на международных конференциях «Корпусная лингвистика», «Диалог» (Москва), «Бубриховские чтения» (Петрозаводск), «Международная конференция, посвященная памяти Микаэля Агриколы» (Нарва), «Международная филологическая конференция» секция «Уралистика» (Санкт-Петербург) и др.

**Архитектура и количественные характеристики корпуса**

Разрабатываемый корпус ВепКар обладает следующими характеристиками:

- является многоязычным корпусом, включает тексты на вепсском и карельском языках, вепсские и карельские словари имеют толкования на русском и частично английском языках;
- является полнотекстовым корпусом, то есть разметка текстов выполняется полностью и поиск осуществляется по всему массиву текстов;
- предоставляет пользователям доступ к полным текстам документов, то есть корпус ВепКар можно рассматривать как собрание текстов (электронную библиотеку). Ограничение возникает при включении в корпус копирайтных текстов. Материалами для ВепКар могут служить тексты только с открытой лицензией[5], поэтому такого ограничения нет;
- относится ко всему языку, то есть включает различные жанры и стили. В этом смысле, в соответствии с определением, предложенным в работе [Кибрик и др., 2019, с. 418], ВепКар является национальным корпусом вепсского и карельского языков;
- включает разметку:
  – метатекстовую (название текста, дата создания, автор, жанр, место записи и др.);
  – морфологическую (у слов в текстах указаны части речи и морфологические признаки);
  – семантическую (слова в текстах связаны со значениями словарных статей).

Таким образом, ВепКар – это многоязычный полнотекстовый лингвистический корпус, содержащий морфологическую, семантическую и метатекстовую разметку.

Корпусным менеджером в проекте ВепКар является разрабатываемый на платформе Laravel и языке программирования PHP комплекс программ с открытым исходным кодом dictorpus[6]. Данные хранятся в базе данных MySQL.

Проект ВепКар включает корпус текстов и связанный с ним словарь (рис. 1). На февраль 2021 года в корпусе свыше 3000 текстов на 26 диалектах вепсского и карельского языков, свыше миллиона словоупотреблений. У некоторых текстов есть параллельный перевод на русский язык. Словарь корпуса содержит свыше 60 тысяч словарных статей со словоформами на 18 диалектах. Толкования слов в основном приводятся на русском и английском языках, хотя есть возможность давать толкования на вепсском, наречиях карельского и финском языках.

Тексты корпуса автоматически размечены на 73 %, то есть для 790 тысяч слов в текстах система нашла соответствие в словаре: известны начальная форма (лемма), часть речи и другие морфологические признаки. Следующим этапом является работа экспертов по проверке разметки и снятию семантической (выбор значения) и морфологической (выбор грамматических признаков) омонимии. Например, на рис. 2 для карельского слова “missä” было найдено три значения местоимения “mi”[7] (*что*; *сколько*, *что*; *какой*), одно значение наречия “missä”[8] (*где*) и один набор грамсетов, то есть грамматических признаков (*инессив*, *ед. ч.*; *наречие* – неизменяемое слово, поэтому без грамсетов). Для слова “muissa” было найдено одно значение глагола “muistua”[9], одно

[1] См. https://github.com/unimorph/vep
[2] См. https://github.com/unimorph/krl
[3] См. https://github.com/unimorph/olo
[4] См. https://github.com/unimorph/lud
[5] См. разрешения от вепсских и карельских писателей и поэтов на включение текстов их произведений в корпус ВепКар http://dictorpus.krc.karelia.ru/ru/page/permission
[6] См. https://github.com/componavt/dictorpus
[7] См. словарную статью “mi” http://dictorpus.krc.karelia.ru/ru/dict/lemma/24862
[8] См. словарную статью “missä” http://dictorpus.krc.karelia.ru/ru/dict/lemma/24881
[9] См. словарную статью “muistua” http://dictorpus.krc.karelia.ru/ru/dict/lemma/24929

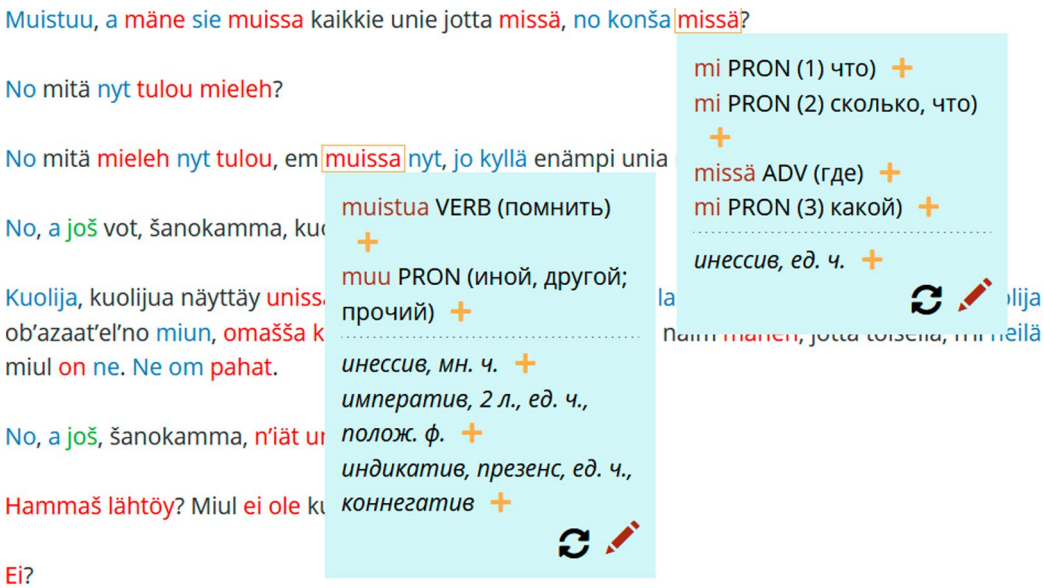



*Рис. 2*. Примеры семантической и морфологической омонимии в разметке текста корпуса ВепКар, редактор выбирает значение и морфологические признаки

*Fig. 2*. Examples of semantic and morphological homonymy in the text markup of the VepKar corpus, the editor chooses one of the word meanings and morphological features

значение местоимения "muu"[1] и три грамсета (одно для именной части речи и два для глагола). Эксперт, кликая на иконку "+", выбирает верное значение и подходящий грамсет.

При клике на неразмеченное слово в тексте (система не нашла соответствия в словаре) открывается дополнительное окно, в котором можно вручную выбрать или создать новую лемму (с полной парадигмой, если ввести шаблон), указать грамматические признаки (рис. 3). Здесь же можно воспользоваться прогнозом модуля-предсказателя и выбрать готовый ответ (начальная форма, часть речи и набор грамматических признаков). В результате будет не только размечено слово в тексте, но и словарь пополнится новыми данными. Модуль-предсказатель работает по алгоритму поиска несловарного слова по конечным буквосочетаниям [Krizhanovsky, 2020].

Подробнее о структуре и разметке корпуса рассказано в работе [Крижановский, 2019] и в докладе на конференции «Бубриховские чтения: задокументированное народное слово»[2].

## Многообразие ВепКар: подкорпусы и жанры

Корпус ВепКар состоит из подкорпусов (рис. 4), выделение которых базируется на двух параметрах: языковая и стилистическая принадлежность текста.

Объединение в рамках корпуса вепсского и карельского языков предполагает выделение соответствующих языковых подкорпусов. Характерной чертой карельского языка является его диалектная разобщенность, заключающаяся в наличии в нем трех наречий (собственно карельского, ливвиковского и людиковского), у которых имеются существенные отличия на всех языковых уровнях: в фонетике, морфологии, лексике. В связи с этим при включении карельского языка в корпус разработчиками создано три его подкорпуса в соответствии с наречиями [Крижановский, 2019, с. 289].

Практически для всех наречий к настоящему моменту разработаны нормированные варианты языка (для собственно карельского даже два – севернокарельский и тверской). Для людиковского наречия нормированный вариант языка находится в стадии разработки. Для таких вариантов, в отличие от диалектов, определены и закреплены в грамматиках и словарях орфоэпические, орфографические, морфологические, словообразовательные правила. Эти языки используются в процессах обучения и обслуживания культурных потребностей

[1] См. словарную статью "muu" http://dictorpus.krc.karelia.ru/ru/dict/lemma/24949

[2] Крижановская Н. Б., Крижановский А. А. Архитектура корпуса ВепКар и диалектные особенности // Бубриховские чтения: задокументированное народное слово. Петрозаводск, 2020. YouTube: https://youtu.be/cUpqM97LXGs

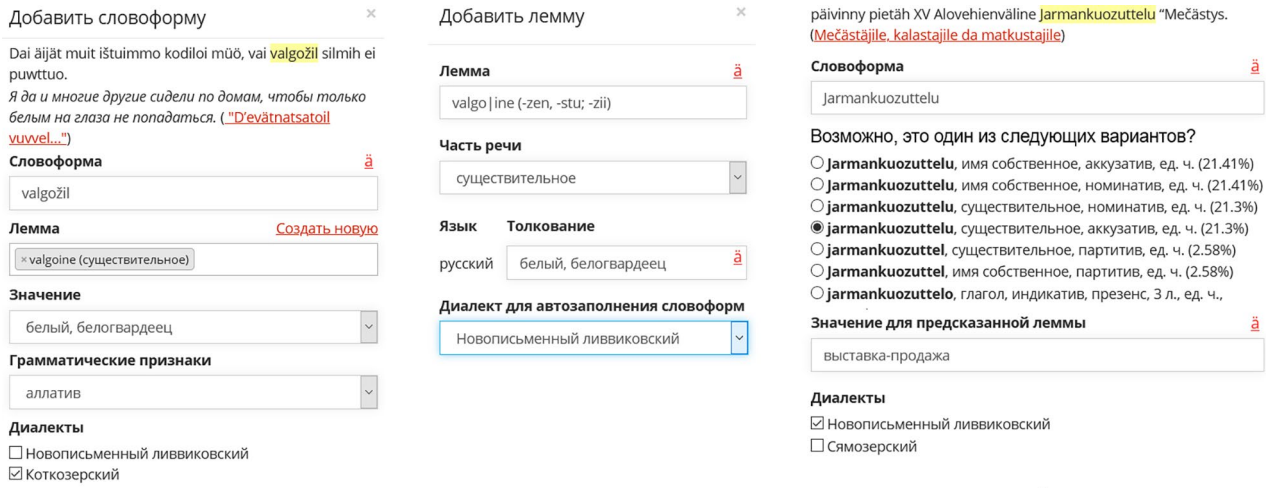



*Рис. 3*. Примеры ручной разметки слов, для которых не были найдены соответствия в словаре:

а) добавление новой словоформы 'valgožil' для найденной в словаре леммы 'valgoine'; б) добавление новой леммы 'valgoine' с одновременной генерацией словоформ по заданному шаблону; в) выбор варианта, предложенного модуль-предсказателем

*Fig. 3*. Examples of manual marking of words with missing entries in the dictionary:

a) adding a new word form *valgožil* for the lemma *valgoine* found in the dictionary; б) adding a new lemma *valgoine* with simultaneous generation of word forms according to the given template; в) choosing the option proposed by the predictor module

народа, в средствах массовой информации, в художественной литературе. Наличие нормированных вариантов позволяет в ходе работы над корпусом, с одной стороны, отталкиваться от них в процессе обработки большого числа текстов (например, при разметке диалектных текстов), а с другой, перепроверять существующие нормы и, в случае выявления несоответствий, корректировать их.

По историческим причинам, по доступному материалу и в попытке разделить тексты по стилю в корпусе ВепКар сложились следующие подкорпусы (рис. 4): младописьменный подкорпус, диалектные тексты, переводные

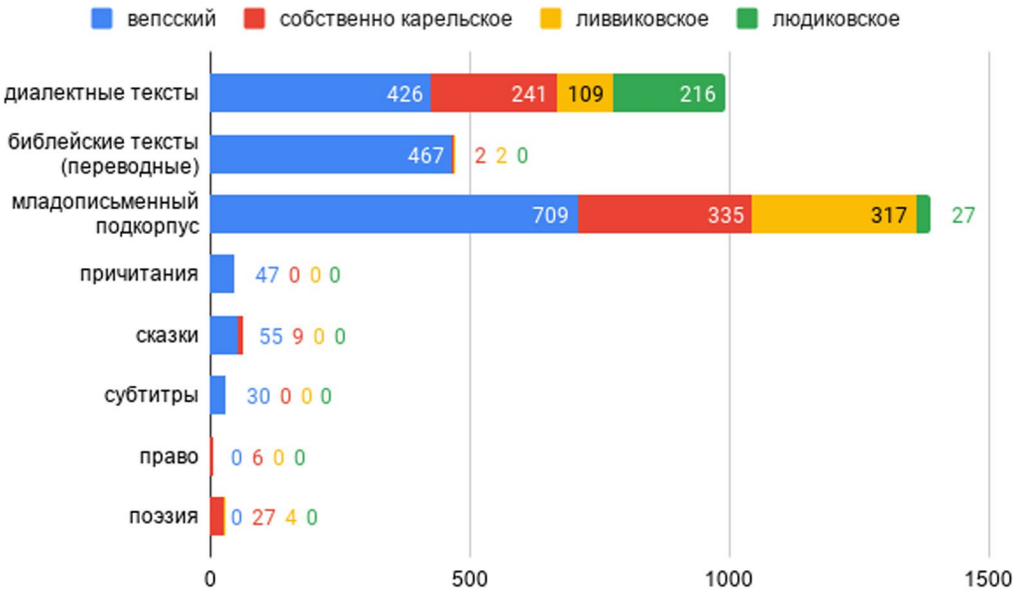



*Рис. 4*. Распределение текстов ВепКар по подкорпусам

*Fig. 4*. The number of texts in different subcorpora of VepKar

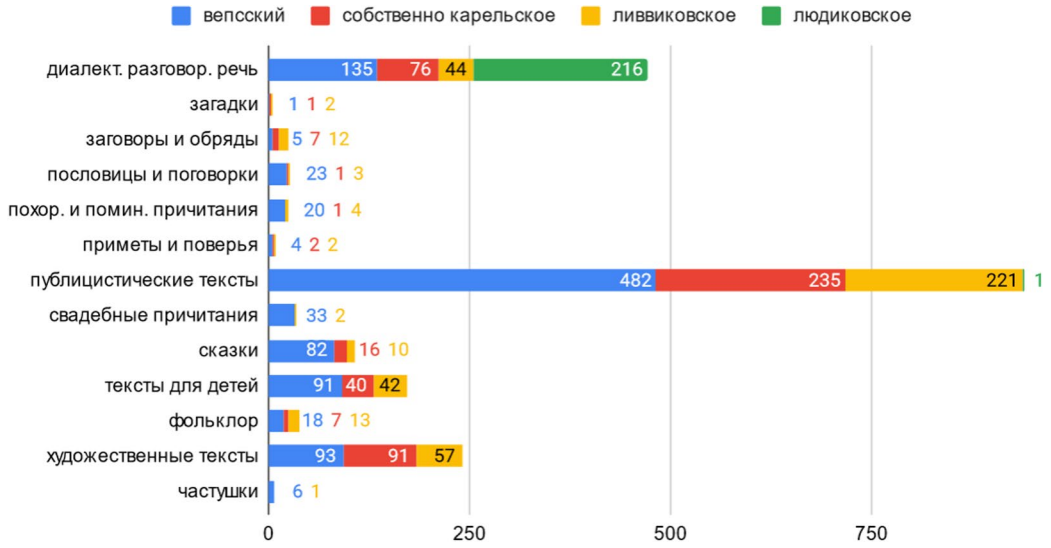



*Рис. 5*. Распределение текстов ВепКар по жанрам

*Fig. 5*. The number of the VepKar texts of different genres

библейские тексты, сказки, причитания, поэзия, право и субтитры. ВепКар пополняется материалами по соглашению с издательствами и авторами текстов. Неисчерпаемым ресурсом является также Научный архив КарНЦ РАН.

Распределение текстов ВепКар по жанрам (рис. 5) достаточно неравномерное. Это главным образом объясняется характером доступных источников (газеты, сборники диалектных текстов). В то же время потребуется внимательная работа специалиста по определению жанровой принадлежности ряда текстов.

Из рис. 6 видно, что тексты добавлялись в корпус с 2013 года по настоящее время (столбцы желтого цвета). Значительное увеличение текстовой базы в 2018–2020 годах объясняется получением корпусом финансовой поддержки в виде гранта РФФИ. Активная запись информантов (голубой цвет) велась до 1990-х годов. Год публикации (красный цвет) имеет пики, поскольку одна книга может содержать много текстов. Например, в 1969 году вышла в свет книга «Образцы карельской речи. Говоры ливвиковского диалекта карельского языка» [Макаров, Рягоев, 2019], а в 2006-м – перевод Библии на вепсский язык [Uz' Zavet, 2006].

В последнее время ВепКар стал активнее пополняться произведениями карелоязычных

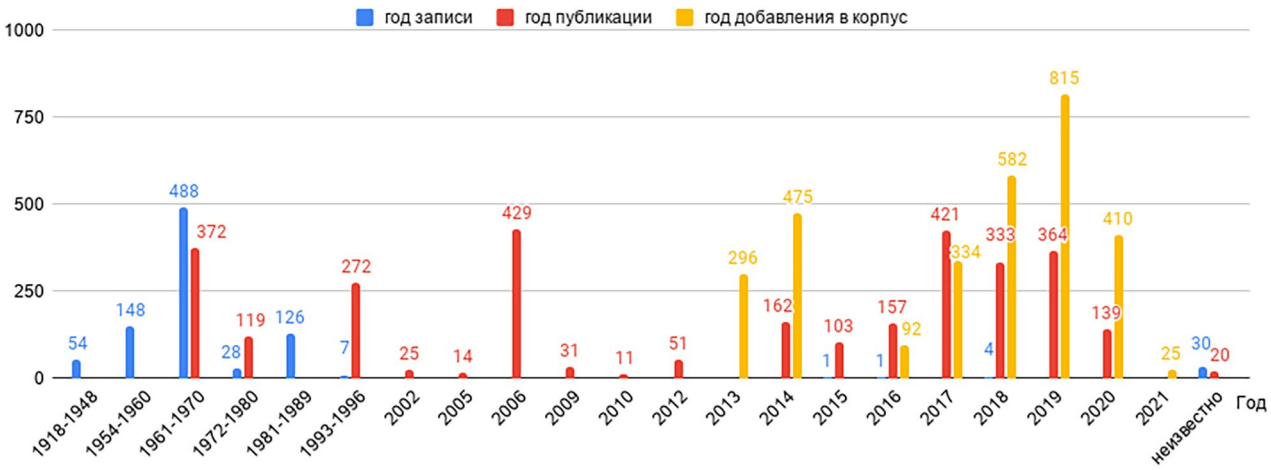



*Рис. 6*. Распределение числа текстов ВепКар по годам записи (информантов), дате публикации и добавления в корпус

*Fig. 6*. Number of the VepKar texts by year of recording (informants), date of publication, and date of adding the text to the corpus

писателей и поэтов. Постоянное комплектование корпуса подобными материалами способствует популяризации карельского и вепсского языков, а также решению целого ряда просветительских, образовательных и исследовательских задач (не только в области литературоведения, но и в сфере культурологии, лингвистики и др.).

Увеличение числа художественных текстов в VepKar, равно как и продолжающаяся работа по введению в корпус новых лемм и его разметке, открывает широкие перспективы для исследования языка карелоязычной прозы и поэзии. Так, в дальнейшем на платформе корпуса станет возможным, например, изучение произведений отдельно взятого автора (как вариант, на основе статистики использования им определенных частей речи, деминутивов, фразеологизмов и т. д.). Развитие подкорпуса художественных текстов может способствовать разработке программ для решения целого спектра прикладных задач, например, реконструкции утраченных слов и частей текстов литературного произведения, выбора вариантов текста из черновиков автора при подготовке не изданных им при жизни текстов и др. Одной из масштабных перспектив подобных наработок, в свою очередь, может стать создание и развитие морфоанализатора в VepKar. Это означает, что программа, обученная на массивах текстов, в том числе художественных, в дальнейшем, вероятно, сможет прогнозировать нужные формы имен при глаголах (глагольное управление), порядок слов в предложении, распознавать основу предложения и определять окончания именных и глагольных словоформ.

## Возможности и приложения корпуса

Корпус VepKar является многофункциональным: он содержит большое количество инструментов, позволяющих языковедам успешно использовать его в своих исследованиях.

В корпусе организована удобная система поиска (рис. 7–9). Возможности расширенного поиска (рис. 7) помогают отфильтровать тексты не только по языковой или стилистической, но и по диалектной принадлежности, или, например, по информанту, собирателю или автору, году записи или году публикации, поиск лемм (рис. 9) возможен по диалектам, частям речи, грамматическим признакам и даже по лексико-семантическим категориям.

Поиск по лексико-семантическим категориям оказался возможным благодаря большой работе по интеграции «Сопоставительно-ономасиологического словаря диалектов карельского, вепсского, саамского языков» (СОСД) в словарную часть VepKar. Этот монументальный словарь подготовлен коллективом авторов (А. П. Баранцев, П. М. Зайков, М. И. Зайцева, Н. Г. Зайцева, Л. Ф. Маркианова, В. Д. Рягоев, В. П. Федотова) и представляет диалектные именования полутора тысяч понятий на 24 карельских, 6 вепсских и 5 саамских диалектах и говорах, находящихся между собой в генетическом родстве [СОСД…, 2007]. Словарь иллюстрирует лексическое многообразие названий понятий и их фонетических вариантов по диалектам и говорам, а также показывает сходства и различия сопоставляемых языков и диалектов. Благодаря данным сопоставительно-ономасиологического словаря в словарь VepKar:

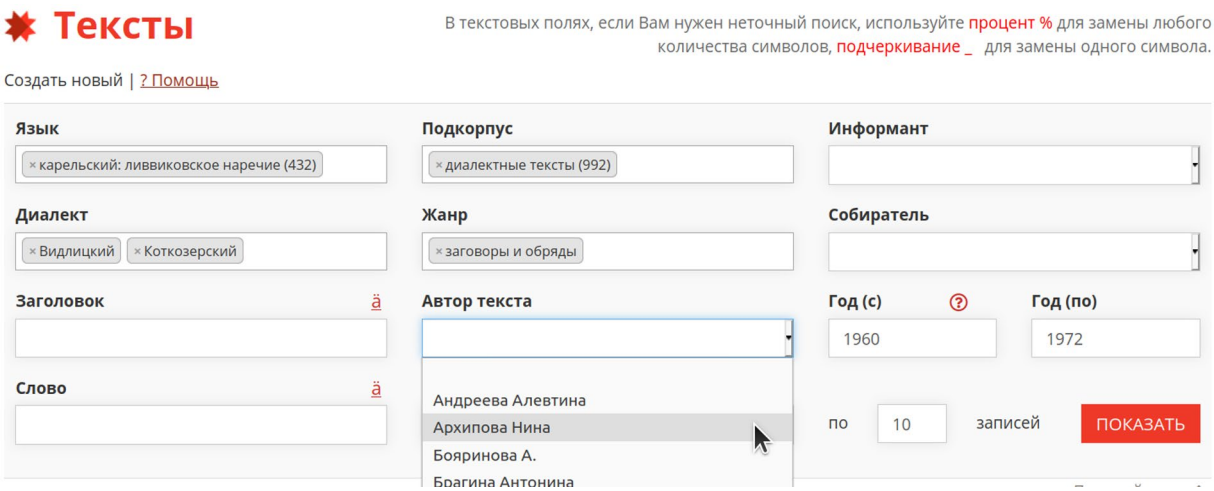
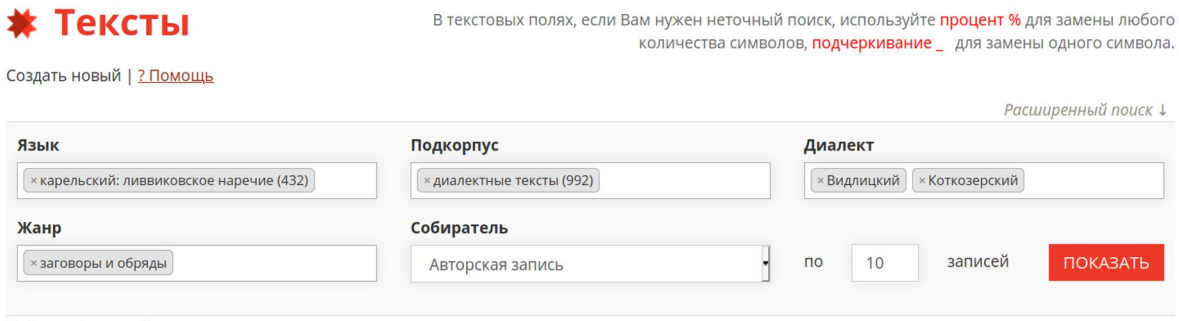



*Рис. 7*. Расширенный интерфейс поиска по текстам VepKar, раскрыто выпадающее меню для выбора автора текста

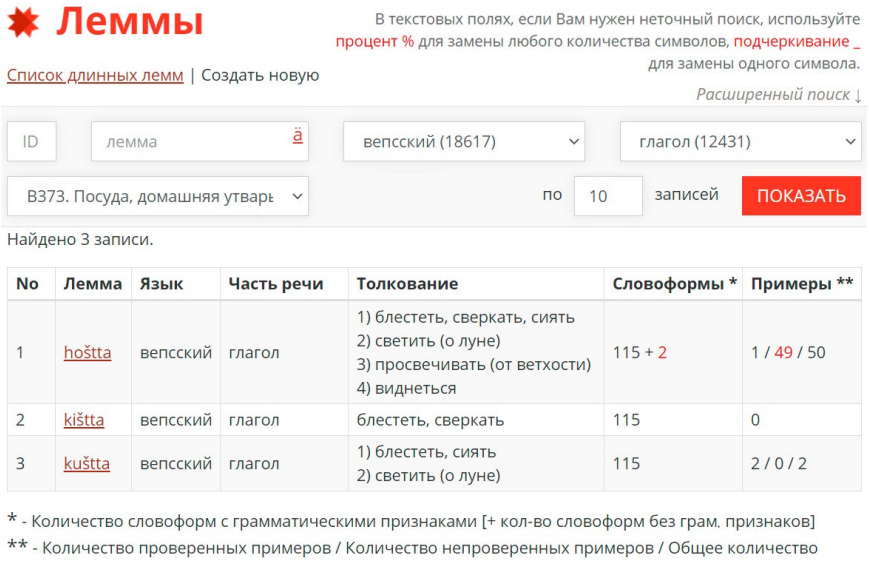
*Fig. 7*. Advanced search interface for the VepKar texts, a drop-down menu for selecting the author of the text is expanded




| No | Язык | Диалект | Подкорпус | Заголовок | Перевод |
|---|---|---|---|---|---|
| 1 | карельский: ливвиковское наречие | Коткозерский | диалектные тексты | Kui enne svuad'bua piättih | Как раньше свадьбу играли |

*Рис. 8*. Компактный интерфейс поиска по текстам и результаты поиска текста в жанре «заговоры и обряды», написанного на коткозерском диалекте ливвиковского наречия в диалектном подкорпусе ВепКар, найден один текст

*Fig. 8*. Compact interface for text search and found texts in the genre of "zagovory and rituals" written in the Kotkozero dialect of the Livvi dialect in the dialect VepKar subcorpus, one text was found




| No | Лемма | Язык | Часть речи | Толкование | Словоформы * | Примеры ** |
|---|---|---|---|---|---|---|
| 1 | hoštta | вепсский | глагол | 1) блестеть, сверкать, сиять<br>2) светить (о луне)<br>3) просвечивать (от ветхости)<br>4) виднеться | 115 + 2 | 1 / 49 / 50 |
| 2 | kištta | вепсский | глагол | блестеть, сверкать | 115 | 0 |
| 3 | kuštta | вепсский | глагол | 1) блестеть, сиять<br>2) светить (о луне) | 115 | 2 / 0 / 2 |

* - Количество словоформ с грамматическими признаками [+ кол-во словоформ без грам. признаков]
** - Количество проверенных примеров / Количество непроверенных примеров / Общее количество

*Рис. 9*. Интерфейс поиска по леммам и результаты поиска ливвиковских глаголов, связанных с понятием «Посуда, домашняя утварь», в словаре ВепКар

*Fig. 9*. Interface for the search for lemmas and the search results for Livvi verbs related to the concept of "Dishes, household utensils" in the VepKar dictionary

1) добавлено 1425 понятий;
2) связаны с этими понятиями значения 20 тыс. вепсских и карельских лемм;
3) добавлено 130 тыс. переводов (связей между значениями лемм из разных языков, наречий).

В процессе кодирования сопоставительно-ономасиологического словаря для добавления его в ВепКар выработана система, позволяющая отразить не только особенности лексического состава исследуемых языков, но и фонетические особенности каждого диалекта.

Специальные модули, разрабатываемые и добавляемые в корпусный менеджер ВепКар, помогают редакторам в решении их повседневных задач, например, связанных с разметкой или пополнением словарей и корпуса. Так, в 2019–2020 годы были сформулированы правила словоизменения[1] [Крижановская и др., 2021а, б] и реализованы алгоритмы генерации словоформ для вепсского языка [Krizhanovskaya, Krizhanovsky, 2019], ливвиковского

[1] Доступны онлайн разработанные и формализованные правила генерации глагольных словоформ для северно-карельского варианта карельского языка [Крижановская и др., 2021а] и правила генерации именных словоформ для собственно карельского и ливвиковского наречий [Крижановская и др., 2021б].

и собственно карельского наречий. Это значительно ускорило наполнение словаря и увеличило разметку корпуса [Новак и др., 2020]. В ходе проверки работы генераторов выявлено отсутствие унификации в образовании отдельных грамматических форм именного и глагольного словоизменения в новописьменных вариантах карельского языка. Откорректированные правила найдут применение в практике преподавания. Работа над модулем позволила существенно приблизиться к созданию приложения для проверки правописания карельского языка.

Введение разметки и работа по снятию омонимии в рамках материалов словаря фразеологизмов призвана помочь улучшить качество автоматической разметки текстов. Кроме того, на стадии перехода к работам над автоматическим переводом такой словарь позволит тестировать новые алгоритмы.

Частотные словари (токенов, размеченных лемм, символов) позволяют редакторам правильно расставить приоритеты в своей работе. Например, чтобы сразу покрыть и разметить большее количество текстов, следует в первую очередь вводить в словарь самые распространенные лексемы, а не единичные употребления (гапаксы). Данные частотных словарей могут выступить в качестве базы для перспективных в последние годы статистических и психолингвистических исследований.

Обратный словарь, в котором леммы отсортированы в алфавитном порядке не по начальным буквам, как в традиционных словарях, а по конечным, призван помочь редакторам, например, разобраться в сложных именных и глагольных системах карельского и вепсского языка, каждая из которых содержит около десяти словоизменительных типов. Кроме того, модуль может быть успешно использован начинающими поэтами в качестве помощника в процессе поиска подходящей рифмы.

### Мобильные приложения как наступившее лингвистическое будущее

Несмотря на то что на 2021 год на рынке приложений для телефона Google Play доступно 3,5 млн программ, задача проектирования мобильного лексикографического обеспечения практически не исследована [Caruso et al., 2019]. Успех разработки таких компьютерных программ обеспечивается двумя вещами: (1) уникальностью словарного материала и (2) представлением этого материала настолько удобным способом, чтобы превосходить все другие подобные приложения [Caruso et al., 2019]. Вепсских словарей на платформе Android ранее не было, а русско-карельских и карельско-русских для телефона было по одному, при этом созданных по материалам наших коллег [Бойко, 2016; Бойко, Маркианова, 2016]. Таким образом, разработка электронных словарей для этих языков – это свободная ниша на рынке мобильных приложений.

Итак, в рамках проекта «Открытый корпус вепсского и карельского языков» (ВепКар) на языке программирования Kotlin создана первая версия мобильного приложения Sanahelmi с открытым исходным кодом[1] для платформы Android[2]. Эта программа представляет собой упрощенный вариант словаря корпуса ВепКар (например, здесь нет привязки из словарных статей к примерам словоупотребления в текстах, не приводится полная парадигма[3]). Мобильное приложение отличается от вебсайта корпуса более простым интерфейсом (рис. 10) и возможностью работы без подключения к сети Интернет.

База данных SQLite мобильного приложения содержит все вепсские и карельские слова (леммы и словоформы), грамматические признаки и толкования на русском языке из словаря ВепКар. Сейчас пользователь может выбрать для поиска вепсский язык или наречие карельского языка.

В приложении поиск осуществляется одновременно по леммам и словоформам, а результат обработки запроса оформляется в виде списка. Выводимые по запросу записи имеют разную структуру в зависимости от того, что было найдено: лемма или словоформа (рис. 10). Если найдена лемма, то на экран выводятся сама лемма, язык, часть речи и толкование. Если найдена словоформа, то на экране отображается словоформа, язык, часть речи, грамматический признак и начальная форма (лемма и ее толкование).

При вводе слова можно использовать специальный символ %, который соответствует любому числу символов.

Одной из трудностей разработки мобильного приложения является ограничение на объем памяти. Пока не все данные словарей удалось включить в мобильный словарь. Сейчас в программу Sanahelmi внесено только 60 тыс. лемм и 600 тыс. словоформ вместо 2 млн словоформ, объем всего проекта в среде разработки Android Studio составляет 240 Мб, объем самой

[1] См. https://github.com/componavt/sanahelmi

[2] См. https://play.google.com/store/apps/details?id=vepkar.test

[3] В дальнейшем можно было бы расширить возможности мобильного приложения и приводить полную словоизменительную парадигму слова.


*Рис. 10*. Пример найденных лемм и словоформ по запросу пользователя в мобильном приложении Sanahelmi

*Fig. 10*. An example of a user request and a list of lemmas and word forms found in the Sanahelmi mobile application

программы Sanahelmi – 18 Мб. Возможны следующие способы преодоления ограничения на объем программы:

1. Разработать отдельные приложения для каждого языка (сейчас словарь Sanahelmi содержит и вепсские, и карельские слова вместе).
2. С помощью частотного словаря корпуса ВепКар определить наиболее употребимые слова (лексическое ядро) и только их включить в словарь. Другой вариант – включить все леммы, а вот словоформы выбрать только самые употребимые.
3. С помощью текстов на карельском языке в сети Интернет (например, в социальных сетях) определить наиболее частотные карельские слова для включения в словарь.
4. Оптимизировать структуру словарной статьи с целью экономии места, а именно: хранить не все словоформы целиком, а только несколько аффиксов для нескольких падежей. В этом случае потребуется реализовать более замысловатый поиск по словарю.

Несмотря на трудности, следует продолжать разработку, поскольку, согласно результатам исследования [Rahimi, Miri, 2014], студенты, использующие мобильные словари, учатся лучше студентов, обремененных их печатными аналогами. При дальнейшей разработке конкурентоспособного мобильного приложения стоит постараться учесть разнообразные потребности пользователей словарей [Caruso et al., 2019].

## Заключение

За последние пять лет корпус ВепКар значительно увеличился в объеме и в функционале и сейчас включает частично размеченные тексты и многофункциональные словари (толковые, переводные, обратные, тезаурус и др.). Система содержит инструменты, позволяющие лингвисту выполнять импорт и экспорт данных в различные форматы (например, Universal Dependencies, CONLL), производить тестирование и проверку данных, получать выборки для анализа данных (поиск дублей, группы слов с одинаковыми флексиями, частота встречаемости аффиксов в словаре и т. д.). В дальнейшем планируется добавить в проект этимологический словарь, аудиофайлы со звучанием слов в словарных статьях, географическую привязку. В связи с непрерывно расширяющимися возможностями системы в ближайшем будущем можно будет говорить о переходе от корпуса к лингвистической платформе.

Материалы корпуса ВепКар востребованы на международном уровне. Во-первых, размеченные данные корпуса ВепКар в формате CONLL использовались в соревновании «Оценка методов обработки малоресурсных языков», проведенном в феврале–марте 2019 года в рамках международной конференции «Диалог»[1] [Klyachko, 2019]. Во-вторых, с 2019 года данные корпуса внесены в международную морфологическую базу данных UniMorph, содержащую леммы, словоформы и морфологические признаки 110 языков[2] [McCarthy, 2020]. Включение данных ВепКар в научный оборот оказалось возможным в том числе благодаря открытой лицензии Creative Commons Attribution (CC-BY), по которой распространяются тексты и словари корпуса. Экс-

[1] Результаты этого соревнования и данные корпусов доступны онлайн. См. https://lowresource-lang-eval.github.io
[2] См. https://github.com/unimorph

порт данных в общепринятые форматы (CONLL, UniMorph) важен для привлечения к исследованию вепсского и карельского языков международного научного сообщества.

*Финансирование исследований осуществлялось из средств федерального бюджета на выполнение государственного задания КарНЦ РАН (Институт языка, литературы и истории КарНЦ РАН, Институт прикладных математических исследований КарНЦ РАН).*

**Литература**

*Бойко Т. П., Маркианова Л. Ф.* Большой русско-карельский словарь (ливвиковское наречие). 2-е изд., перераб. и доп. Петрозаводск: Периодика, 2016. 399 с.

*Бойко Т. П.* Большой карельско-русский словарь (ливвиковское наречие). Петрозаводск: Периодика, 2016. 352 с.

*Зайцева Н. Г.* Вепсские причитания в фокусе корпусной лингвистики и лингвофольклористики // Матер. XLI Междунар. филолог. конф. (Санкт-Петербург, 26–31 марта 2012 г.), секция «Уралистика». СПб.: СПбГУ, 2012. С. 16–26.

*Зайцева Н. Г., Харитонова Е. Е.,. Жукова О. Ю.* Орфографический словарь вепсского языка. Петрозаводск: КарНЦ РАН, 2012. 432 с.

*Зайцева Н. Г., Крижановская Н. Б.* Корпусная лингвистика в прибалтийско-финском исследовательском пространстве (на материале Корпуса вепсского языка и Открытого корпуса вепсского и карельского языков) // Альманах североевропейских и балтийских исследований. Вып. 3. 2018. С. 264–273. [Электронный ресурс]. URL: https://nbsr.petrsu.ru/journal/article.php?id=1062 (дата обращения: 02.03.2021).

*Иншакова Е. С., Иомдин Л. Л., Митюшин Л. Г., Сизов В. Г., Фролова Т. И., Цинман Л. Л.* СинТаг-Рус сегодня // Труды Института русского языка им. В. В. Виноградова. Вып. 21. Национальный корпус русского языка: исследования и разработки. М., 2019. С. 14–40. [Электронный ресурс]. URL: http://ruslang.ru/doc/trudy/vol21/1-inshakova.pdf (дата обращения: 05.03.2021).

*Кибрик А. Е.* Введение в науку о языке / Под ред. О. В. Федорова и С. Г. Татевосова. М.: Буки Веди, 2019. 672 с. [Электронный ресурс]. URL: http://tipl.philol.msu.ru/application/files/9215/8507/9636/AEK_et_al_corrected_2020.pdf (дата обращения: 01.03.2021).

*Крижановская Н. Б., Новак И. П., Пеллинен Н. А.* Правила генерации глагольных словоформ по минимизированному шаблону для новописьменного севернокарельского варианта карельского языка // figshare. 2021а. Препринт. [Электронный ресурс]. doi: 10.6084/m9.figshare.14237843.v6

*Крижановская Н. Б., Новак И. П., Пеллинен Н. А., Бойко Т. П.* Правила генерации именных словоформ по минимизированному шаблону для новописьменных вариантов собственно карельского и ливвиковского наречий // figshare. 2021б. Препринт. [Электронный ресурс]. doi: 10.6084/m9.figshare.14241833.v1

*Крижановский А. А., Крижановская Н. Б., Новак И. П.* Представление диалектов в Открытом корпусе вепсского и карельского языков (ВепКар) // Корпусная лингвистика – 2019: Тр. междунар. конф. СПб., 2019. С. 288–295.

*Крижановский А. А., Крижановская Н. Б., Родионова А. П.* Архитектура корпусного менеджера и разметка текстов корпуса ВепКар // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы: Матер. II Межд. науч. конф. (Уфа, 27–29 ноября 2019 г.). Уфа: Башк. энцикл., 2019. С. 19–23. URL: http://resources.krc.karelia.ru/math/doc/publ/vepkar_ufa_2019_preprint.pdf (дата обращения: 11.03.2021).

*Макаров Г. Н., Рягоев В. Д.* Образцы карельской речи. Говоры ливвиковского диалекта карельского языка. Л.: Наука, 1969. 283 с.

*Новак И. П., Крижановская Н. Б., Бойко Т. П., Пеллинен Н. А.* Разработка правил генерации именных словоформ для новописьменных вариантов карельского языка // Вестник угроведения. 2020. № 4. С. 679–691.

*Сопоставительно-ономасиологический* словарь диалектов карельского, вепсского, саамского языков / Под общ. ред. Ю. С. Елисеева и Н. Г. Зайцевой. Петрозаводск: КарНЦ РАН, 2007. 346 с. (В тексте – СОСД).

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

*Caruso V., Balbi B., Monti J., Presta R.* How can app design improve lexicographic outcomes? Examples from an Italian Idiom Dictionary // Electronic lexicography in the 21st century: Proceed. of the eLex 2019 conf. (1–3 October 2019, Sintra, Portugal). 2019. P. 374–396. URL: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_21.pdf (дата обращения: 14.03.2021).

*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” (Moscow, May 29 – June 1, 2019). 2019. Iss. 18(25). P. 46–62. URL: http://www.dialog-21.ru/media/5196/_-dialog2019vol-plus.pdf#page=54 (дата обращения: 14.03.2021).

*Krizhanovskaya N. B., Krizhanovsky A. A.* Semi-automatic methods for adding words to the dictionary of VepKar corpus based on inflectional rules extracted from Wiktionary // Corpora 2019 Int. Conf. (St. Petersburg, June 24–28, 2019). St. Petersburg, 2019. P. 211–217. URL: https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf#page=211 (дата обращения: 09.03.2021).

*Krizhanovsky A., Krizhanovskaya N., Novak I.* Part of speech and gramset tagging algorithms for unknown words based on morphological dictionaries of the Veps and Karelian languages // Data Analytics and Mana-

gement in Data Intensive Domains (Voronezh, October 13–16, 2020). Voronezh: Voronezh State Univ., 2020. (In press).

*McCarthy A. D., Kirov C., Grella M., Nidhi A., Xia P., Gorman K., Vylomova E., Mielke S. J., Nicolai G., Silfverberg M., Arkhangelskij T., Krizhanovsky N., Krizhanovsky A., Klyachko E., Sorokin A., Mansfield J., Ernštreits V., Pinter Y., Jacobs C. L., Cotterell R., Hulden M., Yarowsky D.* UniMorph 3.0: Universal Morphology // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, France, 11–16 May, 2020. P. 3915–3924. URL: https://www.aclweb.org/anthology/2020.lrec-1.483.pdf (дата обращения: 18.03.2021).

*Rahimi M., Miri S. S.* The impact of mobile dictionary use on language learning. Procedia – Social and Behavioral Sciences. 2014. Vol. 98. P. 1469–1474. [Электронный ресурс]. URL: https://core.ac.uk/download/pdf/82156394.pdf (дата обращения: 18.03.2021).

*Uz' Zavet* (Новый Завет на вепсском языке). Петрозаводск: Карелия, 2006. 622 с.

*Поступила в редакцию 21.03.2021*

## References

*Boiko T. P., Markianova L. F.* Bol'shoi russko-karel'skii slovar' [Large Russian-Karelian dictionary]. Petrozavodsk: Periodika, 2016. 399 p.

*Boiko T. P.* Bol'shoi karel'sko-russkii slovar' [Large Karelian-Russian dictionary]. Petrozavodsk: Periodika, 2016. 352 p.

*Inshakova E. S., Iomdin L. L., Mityushin L. G., Sizov V. G., Frolova T. I., Tsinman L. L.* SinTagRus segodnya [SinTagRus today]. *Trudy Inst. russ. yazyka im. V. V. Vinogradova* [Trans. V. V. Vinogradov Russ. Language Inst.]. Moscow, 2019. Vol. 21. P. 14–40. URL: http://ruslang.ru/doc/trudy/vol21/1-inshakova.pdf (accessed: 05.03.2021).

*Kibrik A. E.* Vvedenie v nauku o yazyke [Introduction to the science of language] Eds. O. V. Fedorova, S. G. Tatevosov. Moscow: Buki Vedi, 672 p. URL: http://tipl.philol.msu.ru/application/files/9215/8507/9636/AEK_et_al_corrected_2020.pdf (accessed: 01.03.2021).

*Krizhanovskaya N. B., Novak I. P., Pellinen N. A.* Pravila generatsii glagol'nykh slovoform po minimizirovannomu shablonu dlya novopis'mennogo severnokarel'skogo varianta karel'skogo yazyka [Rules for generating verb word forms using a minimized template for the newly written North Karelian version of the Karelian language]. *figshare*. 2021. Preprint. doi: 10.6084/m9.figshare.14237843.v6

*Krizhanovskaya N. B., Novak I. P., Pellinen N. A., Boyko T. P.* Pravila generatsii imennykh slovoform po minimizirovannomu shablonu dlya novopis'mennykh variantov sobstvenno karel'skogo i livvikovskogo narechii [Rules for generating nominal word forms from a minimized template for newly written variants of the Proper Karelian and Livvik dialects]. *figshare*. 2021. Preprint. doi: 10.6084/m9.figshare.14241833.v1

*Krizhanovsky A. A., Krizhanovskaya N. B., Novak I. P.* Predstavlenie dialektov v Otkrytom korpuse vepsskogo i karel'skogo yazykov (VepKar) [Representation of dialects in the Open Corpus of Veps and Karelian languages (VepKar)]. *Trudy mezhd. konf. "Korpusnaya lingvistika – 2019"* [Proceed. int. conf. *Corpus linguistics – 2019*]. St. Petersburg, 2019. P. 288–295.

*Krizhanovsky A. A., Krizhanovskaya N. B., Rodionova A. P.* Arkhitektura korpusnogo menedzhera i razmetka tekstov korpusa VepKar [The architecture of the corpus manager and the layout of the texts of the VepCar corpus]. *Elektronnaya pis'mennost' narodov Rossiiskoi Federatsii: opyt, problemy i perspektivy. Mater. II Mezhd. nauch. konf. (Ufa, 27–29 noyabrya 2019 g.)* [*Electronic writing of the peoples of the Russian Federation: experience, problems, and prospects*. Proceed. II int. sci. conf. (Ufa, Nov. 27–29, 2019)]. 2019. P. 19–23. URL: http://resources.krc.karelia.ru/math/doc/publ/vepkar_ufa_2019_preprint.pdf (accessed: 11.03.2021).

*Makarov G. N., Ryagoev V. D.* Obraztsy karel'skoi rechi. Govory livvikovskogo dialekta karel'skogo yazyka [Samples of Karelian speech. The dialects of the Livvik dialect of the Karelian]. Leningrad: Nauka, 1969. 283 p.

*Novak I. P., Krizhanovskaya N. B., Boiko T. P., Pellinen N. A.* Razrabotka pravil generatsii imennykh slovoform dlya novopis'mennykh variantov karel'skogo yazyka [Development of rules of generation of nominal word forms for new-written variants of the Karelian language]. *Vestnik ugrovedenia* [Bull. Ugric Studies]. 2020. Vol. 10(4). P. 679–691.

*Sopostavitel'no-onomasiologicheskii slovar' dialektov karel'skogo, vepsskogo, saamskogo yazykov* [Comparative onomasiological dictionary of dialects of the Karelian, Veps, and Samic languages]. Eds. Yu. S. Eliseev, N. G. Zaitseva. Petrozavodsk, 2007. 346 p.

*Zaitseva N. G.* Vepsskie prichitaniya v fokuse korpusnoi lingvistiki i lingvofol'kloristiki [Vepsian lamentations in the focus of corpus linguistics and linguistic folkloristics]. *Mater. XLI Mezhdunar. filologich. konf. 26–31 marta 2012 g. Sektsiya "Uralistika"* [Proceed. XLI int. philol. conf., March 26–31, 2012. Section: Uralistics]. St. Petersburg: SPbGU, 2012. P. 16–26.

*Zaitseva N. G., Kharitonova E. E.,. Zhukova O. Yu.* Orfograficheskii slovar' vepsskogo yazyka [Spelling dictionary of the Vepsian language]. Petrozavodsk: KarRC RAS, 2012. 432 p.

*Zaitseva N. G., Krizhanovskaya N. B.* Korpusnaya lingvistika v pribaltiisko-finskom issledovatel'skom prostranstve (na materiale Korpusa vepsskogo yazyka i Otkrytogo korpusa vepsskogo i karel'skogo yazykov) [Corpus linguistics in the Baltic-Finnish research space (based on the Vepsian Language Corpus and the Veps and Karelian Open Corpus)]. *Al'manakh severoevropeiskikh i baltiiskikh issled.* [Almanac of Northern European and Baltic Studies]. Iss. 3. 2018. P. 264–273. URL: https://nbsr.petrsu.ru/journal/article.php?id=1062 (accessed: 2.3.2021).

*Arkhangelskiy T.* Web Corpora of Volga-Kama Uralic Languages. *Finno-Ugric Languages and Linguistics.* 2020. Vol. 9, no. 1–2. P. 58–66.

*Caruso V., Balbi B., Monti J., Presta R.* How can app design improve lexicographic outcomes? Examples from an Italian Idiom Dictionary. Electronic lexicography in the 21st century: Proceed. of the eLex 2019 conf. (1–3 October 2019, Sintra, Portugal). 2019. P. 374–396. URL: https://elex. link/elex2019/wp-content/uploads/2019/09/eLex_2019_21.pdf (accessed: 14.03.2021).

*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" (Moscow, May 29 – June 1, 2019). 2019. Iss. 18(25). P. 46–62. URL: http://www.dialog-21.ru/media/5196/_-dialog2019vol-plus.pdf#page=54 (accessed: 14.03.2021).

*Krizhanovskaya N. B., Krizhanovsky A. A.* Semi-automatic methods for adding words to the dictionary of VepKar corpus based on inflectional rules extracted from Wiktionary. Corpora 2019 Int. Conf. (St. Petersburg, June 24–28, 2019). St. Petersburg, 2019. P. 211–217. URL: https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf#page=211 (accessed: 09.03.2021).

*Krizhanovsky A., Krizhanovskaya N., Novak I.* Part of speech and gramset tagging algorithms for unknown words based on morphological dictionaries of the Veps and Karelian languages. *Data Analytics and Management in Data Intensive Domains* (Voronezh, October 13–16, 2020). Voronezh: Voronezh State Univ., 2020. (In press).

*McCarthy A. D., Kirov C., Grella M., Nidhi A., Xia P., Gorman K., Vylomova E., Mielke S. J., Nicolai G., Silfverberg M., Arkhangelskij T., Krizhanovsky N., Krizhanovsky A., Klyachko E., Sorokin A., Mansfield J., Ernštreits V., Pinter Y., Jacobs C. L., Cotterell R., Hulden M., Yarowsky D.* UniMorph 3.0: Universal Morphology // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, France, 11–16 May, 2020. P. 3915–3924. URL: https://www.aclweb.org/anthology/2020. lrec-1.483.pdf (accessed: 18.03.2021).

*Rahimi M., Miri S. S.* The impact of mobile dictionary use on language learning. *Procedia-Social and Behavioral Sciences.* 2014. Vol. 98. P. 1469–1474. URL: https://core.ac.uk/download/pdf/82156394.pdf (accessed: 18.03.2021).

*Uz' Zavet* (Novyi Zavet na vepsskom yazyke) [New Testament in Veps language]. Petrozavodsk: Kareliya, 2006. 622 p.

*Received March 21, 2021*

## СВЕДЕНИЯ ОБ АВТОРАХ:

**Бойко Татьяна Петровна**
научный сотрудник сектора языкознания
Институт языка, литературы и истории КарНЦ РАН,
Федеральный исследовательский центр
«Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: boiko@krc.karelia.ru

**Зайцева Нина Григорьевна**
ведущий научный сотрудник сектора языкознания,
д. фил. н.
Институт языка, литературы и истории КарНЦ РАН,
Федеральный исследовательский центр
«Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: zng@ro.ru

**Крижановская Наталья Борисовна**
ведущий инженер-исследователь, аспирант лаб. информационных компьютерных технологий
Институт прикладных математических исследований КарНЦ РАН,
Федеральный исследовательский центр
«Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: nataly@krc.karelia.ru

## CONTRIBUTORS:

**Boyko, Tatyana**
Institute of Linguistics, Literature and History,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: boiko@krc.karelia.ru

**Zaitseva, Nina**
Institute of Linguistics, Literature and History,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: zng@ro.ru

**Krizhanovskaya, Natalya**
Institute of Applied Mathematical Research,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: nataly@krc.karelia.ru

**Крижановский Андрей Анатольевич**
руководитель и ведущий научный сотрудник лаб. информационных компьютерных технологий, к. т. н.
Институт прикладных математических исследований КарНЦ РАН,
Федеральный исследовательский центр «Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: andrew.krizhanovsky@gmail.com

**Krizhanovsky, Andrey**
Institute of Applied Mathematical Research,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: andrew.krizhanovsky@gmail.com

**Новак Ирина Петровна**
научный сотрудник сектора языкознания, к. фил. н.
Институт языка, литературы и истории КарНЦ РАН,
Федеральный исследовательский центр «Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: novak@krc.karelia.ru

**Novak, Irina**
Institute of Linguistics, Literature and History,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: novak@krc.karelia.ru

**Пеллинен Наталия Александровна**
младший научный сотрудник сектора языкознания, к. фил. н.
Институт языка, литературы и истории КарНЦ РАН,
Федеральный исследовательский центр «Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: nataliapellinen@gmail.com

**Pellinen, Natalia**
Institute of Linguistics, Literature and History,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: nataliapellinen@gmail.com

**Родионова Александра Павловна**
научный сотрудник сектора языкознания, к. фил. н.
Институт языка, литературы и истории КарНЦ РАН,
Федеральный исследовательский центр «Карельский научный центр РАН»
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: santrar@krc.karelia.ru

**Rodionova, Alexandra**
Institute of Linguistics, Literature and History,
Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: santrar@krc.karelia.ru

**Трубина Елизавета Денисовна**
студентка
Институт математики и информационных технологий,
Петрозаводский государственный университет
пр. Ленина, 33, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: eliza.trubina@gmail.com

**Trubina, Elizaveta**
Institute of Mathematics and Information Technologies,
Petrozavodsk State University
33 Lenin Ave., 185910 Petrozavodsk, Karelia, Russia
e-mail: eliza.trubina@gmail.com