ВЕСТНИК РОССИЙСКОЙ АКАДЕМИИ НАУК, 2024, том 94, № 9, с. 814–823

С КАФЕДРЫ ПРЕЗИДИУМА РАН

ОТКРЫТЫЙ КОРПУС ВЕПССКОГО И КАРЕЛЬСКОГО ЯЗЫКОВ

© 2024 г. И.И. Муллонен[a,*], И.П. Новак[a,]**

[a]*Институт языка, литературы и истории Карельского научного центра РАН, Петрозаводск, Россия*

**E-mail: irma.mullonen@hotmail.com*

***E-mail: bel.irina@rambler.ru*

Поступила в редакцию 15.04.2024 г.
После доработки 20.05.2024 г.
Принята к публикации 17.07.2024 г.

С целью сохранения и системного изучения вепсского и карельского языков сотрудники Института языка, литературы и истории КарНЦ РАН и Института прикладных математических исследований КарНЦ РАН создали и продолжают совершенствовать языковой корпус ВепКар. Проект нацелен на накопление письменных текстов на карельском и вепсском языках, фиксацию и сбережение устной речи, исследование прибалтийско-финских языков Карелии, редактирование норм новописьменных вариантов языков, создание обучающих приложений. Любой желающий может пользоваться ВепКаром как электронной библиотекой и полноценным электронным словарём, что делает этот ресурс весьма востребованным.

Ключевые слова: ВепКар, языковой корпус, карельский язык, вепсский язык, автоматическая разметка, мультимедийный словарь, электронная библиотека, речевой подкорпус.

DOI: 10.31857/S0869587324090045, **EDN:** FCFRPO

Открытый корпус вепсского и карельского языков (ВепКар) [1] – пример языкового корпуса так называемых малых языков России, который, помимо решения научных задач, призван сохранять эти языки в условиях глобализации. Работы над ним ведутся с 2009 г. совместными усилиями сотрудников Института языка, литературы и истории (ИЯЛИ КарНЦ РАН) и Института прикладных математических исследований Карельского научного центра РАН (ИПМИ КарНЦ РАН). Изначально возникнув как корпус вепсского языка [2, 3], он был впоследствии переформатирован в единый ресурс для двух близкородственных языков коренных народов Карелии – вепсского и карельского (рис. 1). Они входят в прибалтийско-финскую языковую семью и сформировались на рубеже I–II вв. близ Онежского и Ладожского озёр, образовав со временем широкую сеть диалектов и говоров. До XX в. языки оставались бесписьменными. Первая попытка создания письменности, предпринятая в 1930-е годы, не увенчалась успехом, главным образом по причинам идеологического и политического характера. Проблему удалось решить только в конце XX в.

МУЛЛОНЕН Ирма Ивановна – член-корреспондент РАН, главный научный сотрудник сектора языкознания ИЯЛИ КарНЦ РАН. НОВАК Ирина Петровна – кандидат филологических наук, директор ИЯЛИ КарНЦ РАН.

Несмотря на создание письменности, вепсский и карельский языки находятся под угрозой исчезновения. Согласно официальным данным Всероссийской переписи населения 2020–2021 гг., численность карелов составляет 32.4 тыс. человек (Республика Карелия – 25.9 тыс., Тверская область – 2.8 тыс., Санкт-Петербург – 727, Ленинградская область – 644, Мурманская область – 631), вепсов – 4.7 тыс. (Республика Карелия – 2.5 тыс., Ленинградская область – 925, Вологодская область – 509, Санкт-Петербург – 254) [4], носителей карельского языка – 13.9 тыс. (Республика Карелия – 11.1 тыс., Тверская

Рис. 1. Главная страница Открытого корпуса карельского и вепсского языков

область – 1.8 тыс., Ленинградская область – 193, Санкт-Петербург – 179, Мурманская область – 137), вепсского – 2.2 тыс. (Ленинградская область – 875, Республика Карелия – 754, Вологодская область – 341, Санкт-Петербург – 98) [5].

О тревожном положении карельского языка свидетельствует его включение в “Атлас языков мира, находящихся под угрозой исчезновения” ЮНЕСКО (статус “definitely endangered” – “под угрозой”). Вепсский язык отнесён к категории “под серьёзной угрозой” (“severely endangered”) [6, с. 36]. В разработанном в Институте языкознания РАН списке языков России вепсский язык входит в группу 2А – “прерванных” (“межпоколенческая передача прервана на всём ареале сообщества, регулярная коммуникация ограничена”), а карельский – в группу 2А+ – “прерывающихся” языков (“существуют такие действия по поддержке языка, которые при их продолжении в будущем и эффективности позволят отнести язык к последующей группе”) [3, с. 16]. В сложившейся ситуации целью корпуса ВепКар стали сохранение, системное изучение, развитие и популяризация языков прибалтийско-финских народов Карелии – вепсского и карельского.

Накопление и сохранение письменных текстов. Корпус представляет собой своеобразную библиотеку, аккумулирующую всё многообразие текстов на карельском и вепсском языках. Сегодня в нём представлено более 6 тыс. записей (около 2 млн словоупотреблений). Источниками для пополнения ресурса служат художественные произведения, сборники диалектных и фольклорных текстов, периодическая печать, учебные пособия. Самый ранний текст – предание о первом жителе пос. Кестеньга – датирован 1871 г. В ВепКар непрерывно заносятся материалы из свежих номеров газет на карельском и вепсском языках, а также расшифровки ежегодных экспедиционных записей. Все тексты размещаются целиком, с согласия авторов и издательств. Открытость корпуса подразумевает свободный доступ ко всем данным и отсутствие ограничений на их использование.

Сформировано четыре подкорпуса по языковой принадлежности: один вепсский и три карельских, в соответствии с количеством новописьменных нормированных вариантов языка. В зависимости от стилистики в ВепКаре сложились публицистический, художественный, диалектный, фольклорный и другие подкорпусы [7, 8]. В ближайшие годы планируется разработка подкорпусов памятников письменности, переключения кодов и учебного. Наличие жанровой системы позволяет составлять тематические коллекции текстов, в частности, вепсских причитаний, карельских рун и карельских топонимических преданий. Следующий этап – наполнение коллекций по отдельным поэтам и писателям, а также тематических коллекций этнографического содержания (рис. 2).

Планомерное обогащение ВепКара текстами уже сейчас позволяет использовать его в качестве основного источника информации для исследований, а также базы для лингвистических приложений. В будущем он поможет в разработке автоматизированных систем машинного перевода и построении компьютерных моделей карельского и вепсского языков.

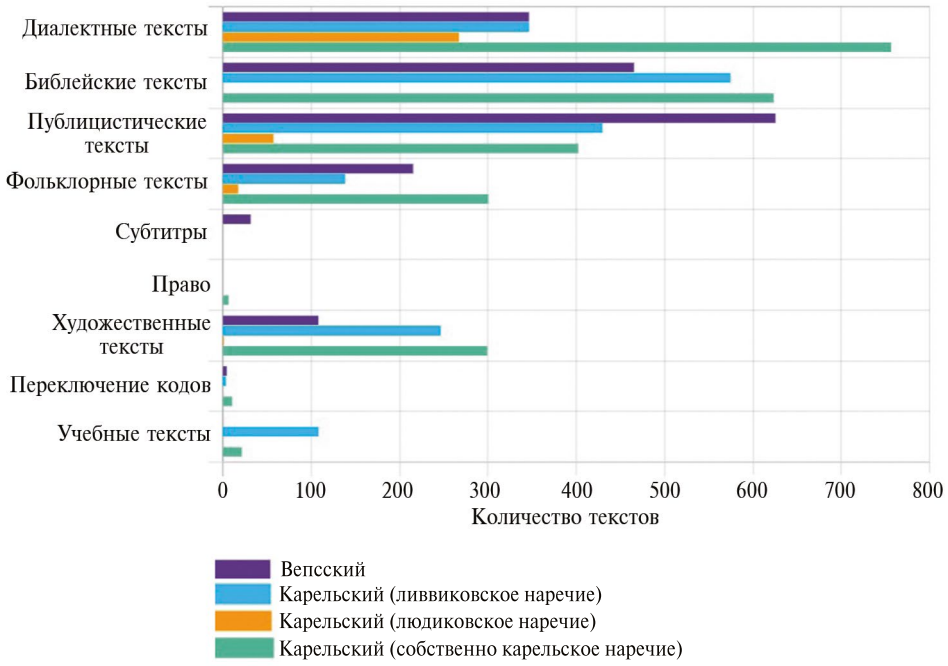

Рис. 2. Статистика текстов по подкорпусам ВепКара

Фиксации и хранение устной речи. С 2022 г. ведётся работа по формированию речевого подкорпуса, содержащего аудиозаписи устной речи, их аннотированные транскрипции и перевод на русский язык [9]. Аудиофайлы сопровождают диалектные, учебные и художественные тексты. Это крайне важно не только для исследования фонетических систем карельского и вепсского языков, но и для создания приложений по распознаванию и синтезу речи.

Все диалектные материалы, снабжённые звуковой дорожкой, для наглядности выводятся на Аудиокарту прибалтийско-финских языков Карелии и сопредельных областей [10]. Сейчас на карте представлено более 100 образцов диалектной речи (рис. 3). Основной их источник – коллекция звукозаписей из экспедиционных материалов, хранящаяся в фонограммархиве ИЯЛИ КарНЦ РАН. Она начала формироваться в послевоенные годы и включает записи самых разных говоров, в том числе уже утраченных с диалектной карты обоих языков. В 2024 г. коллектив корпуса получил грант Русского географического общества на проведение экспедиционных исследований, который позволит наполнить диалектный речевой подкорпус образцами говоров, отсутствующих в более ранних записях.

Исследование прибалтийско-финских языков Карелии. Для удобства работы в корпусе ВепКар представлено три вида разметки:

• *метатекстовая*: язык, диалект, жанр (для фольклорных произведений – циклы, сюжеты, мотивы, темы), название текста, автор, дата создания, автор перевода, информация о публикации (автор, название, год, страницы), данные об информанте (ФИО, год и место рождения), год записи, место записи, информация о собирателе, место хранения (архивные сведения), комментарии к источнику и тексту;

• *морфологическая*: у каждого слова в тексте указаны части речи и морфологические признаки;

• *семантическая*: слова в текстах связаны с определениями словарных статей.

Морфологическая и семантическая разметка возможна благодаря связи корпуса текстов со словарём, который насчитывает 67.5 тыс. словарных статей и более 2.5 млн словоформ (табл. 1). Для всех имён и глаголов в словаре имеются полные словоизменительные парадигмы (40 словоформ для имён, 150 – для глаголов). Агглютинативная структура карельского и вепсского языков позволила разработать генераторы словоформ [11]. При

Рис. 3. Пример текста из речевого диалектного подкорпуса

Таблица 1. Статистика по словарю корпуса ВепКар (по состоянию на 24.07.2024 г.)

Язык	Леммы	Словоформы
Вепсский	19 005	835 714
Карельский (ливвиковское наречие)	27 697	1 275 286
Карельский (людиковское наречие)	6 332	96 103
Карельский (собственно карельское наречие)	15 345	589 899
Всего	68 425	2 797 002

внесении в словарь нового слова редактору теперь нет необходимости вручную вводить длинные словоизменительные ряды, достаточно указать основы слова (слабую гласную для одноосновных, гласную и согласную для двуосновных имён и глаголов), чтобы программа автоматически сгенерировала все его возможные грамматические формы. Одновременно для каждой словоформы осуществляется поиск совпадений по всему корпусу текстов, что позволяет постепенно увеличивать долю автоматической разметки (рис. 4).

Все загружаемые тексты размечаются автоматически в среднем на 78% (табл. 2). Достигнуть такого показателя удалось за счёт внесения в словарь корпуса всех изданных ранее словарей нормированных вариантов карельского и вепсского языков с полными словоизменительными парадигмами. Редакторы проверяют автоматическую разметку, снимают омонимию, а также производят ручную разметку слов, не распознанных программой. Кликая на иконку “+”, можно выбирать верное значение и соответствующие грамматические признаки. Слово может остаться нераспознанным по двум причинам: в нём допущена орфографическая ошибка или оно отсутствует в словаре (рис. 5).

Зелёным цветом отмечены проверенные редактором слова, синим (отсутствие омонимии) и красным (наличие омонимии) – результат автоматической разметки, требующий проверки экспертом

Для продолжения работы по повышению доли автоматической разметки запланирован важный

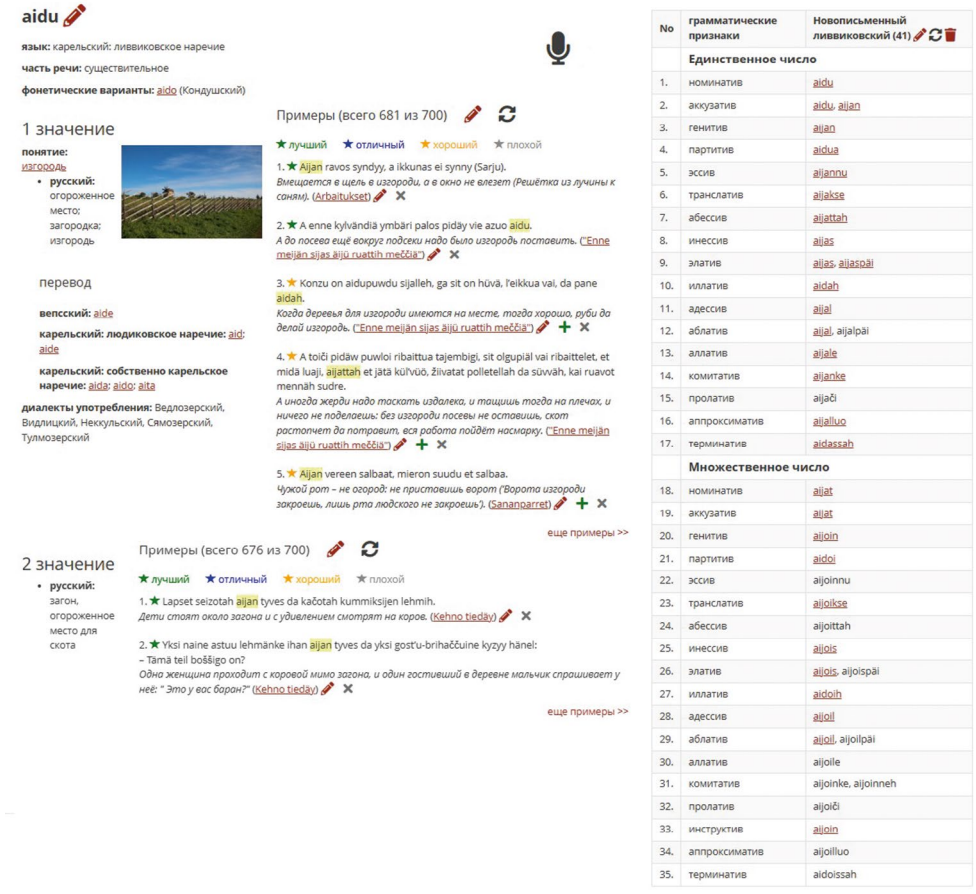

Рис. 4. Пример словарной статьи из Словаря лемм корпуса ВепКар

Таблица 2. Статистика автоматической разметки по подкорпусам ВепКара (по состоянию на 06.06.2024 г.)

Язык	Количество слов в текстах	Количество размеченных слов	Доля размеченных слов, %
Вепсский	530 898	452 368	85.2
Карельский (ливвиковское наречие)	603 213	504 916	83.7
Карельский (людиковское наречие)	104 761	66 636	63.6
Карельский (собственно карельское наречие)	750 407	545 990	72.8
Всего	1 989 279	1 569 910	78.9

этап – внесение в корпус данных диалектных словарей. Кроме того, перед коллективом поставлена задача завершить создание “золотого стандарта”, то есть массива текстов с проверенной редактором разметкой, который в дальнейшем будет использоваться в различных экспериментах, нацеленных на разработку программы для автоматического снятия грамматической омонимии.

Сейчас можно с уверенностью заявить, что главная цель, ради которой 15 лет назад был создан ВепКар, а именно исследование карельского и вепсского языков, достигнута. Материалы корпуса в сочетании с программами обработки, поиска и представления данных позволяют решать научные задачи в области лексики и грамматики карельского и вепсского языков. На базе корпуса ведётся изучение сочетаемости слов, управления, словообразовательных моделей и пр. На основе частотных словарей проводятся статистические исследования, к процессу определения словоизменительных типов имён и глаголов привлекаются обратные словари. Удобная система лексико-грамматического поиска позволяет выбирать из массива текстов сложные грамматические конструкции, которые представляют собой заданную последовательность словоформ, обладающих определённым набором признаков (рис. 6).

Развитие новописьменных вариантов языков. Карельский и вепсский языки имеют статус новописьменных, то есть история развития их письменности насчитывает не более 30 лет. Именно нормированные новописьменные варианты карельского (ливвиковский, севернокарельский, тверской) и вепсского языков положены в основу словарей лемм и словоформ корпуса. Наличие норм – важный систематизирующий фактор, позволяющий при наполнении ресурса сводить воедино отличающийся многообразием языковой материал из разных источников. При

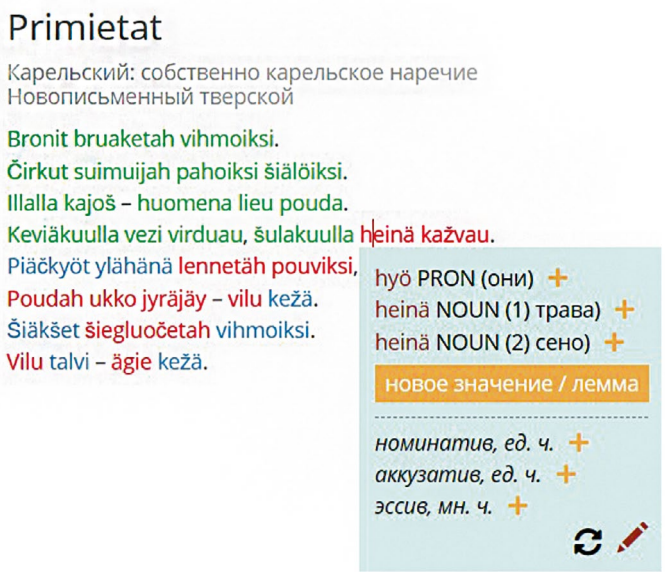

Рис. 5. Пример работы редактора по снятию омонимии

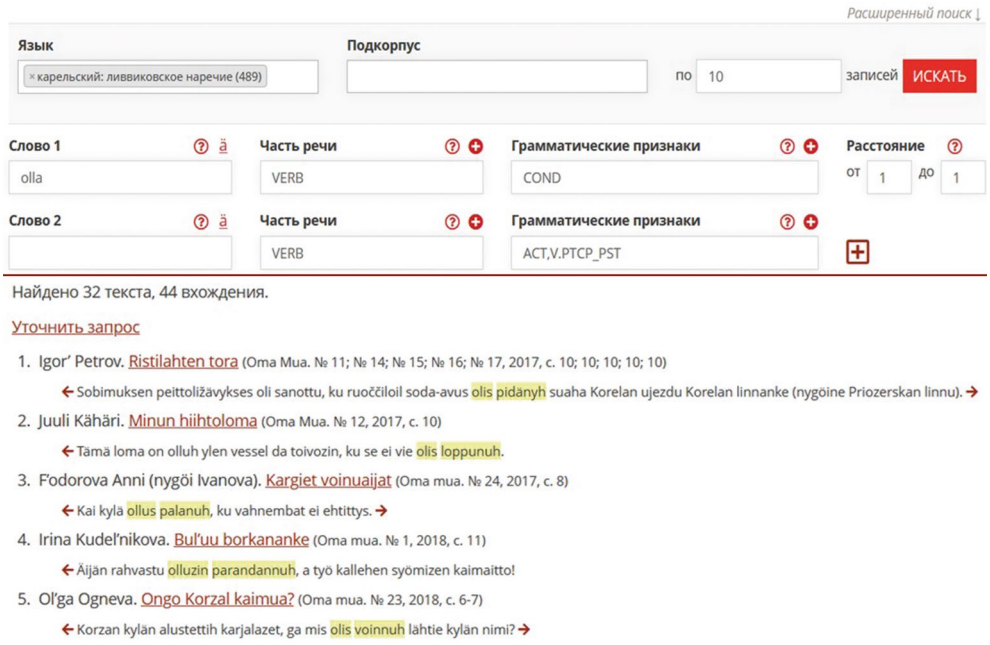

Рис. 6. Пример подбора ливвиковских глаголов в форме перфекта кондиционала с помощью системы лексико-грамматического поиска ВепКара

этом с помощью проверки автоматической разметки эксперт выявляет изъяны в существующих нормах и выдвигает предложения по их корректировке. Это естественная ситуация для младописьменных языков, нормы которых ещё не устоялись и не всегда учитывают все необходимые позиции. Практически все редакторы корпуса являются членами Республиканской термино-орфографической комиссии по карельскому и вепсскому языкам при главе Республики Карелия, один – член Комиссии по использованию письменной формы языка тверских карел в публичной сфере.

Рис. 7. Пример работы приложения по проверке орфографии
Слова, подчёркнутые красным, написаны с ошибкой или отсутствуют в базе корпуса

Материалы корпуса внесли большой вклад в подготовку “Орфографического словаря вепсского языка” [12] и “Грамматико-орфографического словаря карельского языка” [13]. Кроме того, создано приложение по проверке орфографии обоих языков (рис. 7).

Популяризация карельского и вепсского языков. Многофункциональность корпуса ВепКар заключается в том, что, кроме исследовательских задач, он может активно применяться рядовыми пользователями в качестве электронной библиотеки, полноценного электронного словаря или универсального обучающего ресурса. О востребованности корпуса свидетельствует статистика посещений (рис. 8, 9). Например, за первый квартал 2024 г. зафиксировано более 2.5 тыс. визитов на сайт (без учёта роботов), из которых около 1 тыс. – уникальные посетители. Основную аудиторию ресурса ожидаемо составили пользователи из России и Финляндии – стран, где проживает карелоязычное население.

На основе данных корпуса разрабатываются приложения с интуитивно понятным интерфейсом для широкого круга пользователей, интересующихся прибалтийско-финскими языками Карелии, например, мультимедийный словарь карельского языка ливвиковского наречия LiPaS (от карел. “lipas” – сундучок, шкатулка) (рис. 10). Его целевая аудитория – школьники, студенты, слушатели курсов карельского языка и преподаватели. Словник формируется автоматически из текстов корпуса и постоянно пополняется по мере загрузки новых данных. Программа также предлагает иллюстративный материал (контекстные примеры) для каждого слова. Задача редактора – проверить правильность заполнения страницы словарной статьи, отобрать наиболее удачные примеры и снабдить их переводом на русский язык. Работа по наполнению мультимедийного словаря LiPaS и созданию новых словарей (для других нормированных вариантов) продолжается. Подобные приложения позволяют компенсировать ограниченность информации традиционных бумажных словарей, что особенно актуально при изучении языка.

Успешное достижение всех поставленных перед коллективом ВепКара задач упирается в одну серьёзную проблему – развитие кадрового потенциала. Костяк сотрудников, регулярно задействованных в работе над корпусом, составляет восемь

Статистика посещений январь–март 2024	Просмотров
Словарь	23 400
Тексты	6 700
Стартовая страница	1 300
Словоформы	1000
Речевой подкорпус	550
Лексико-грамматический поиск	300

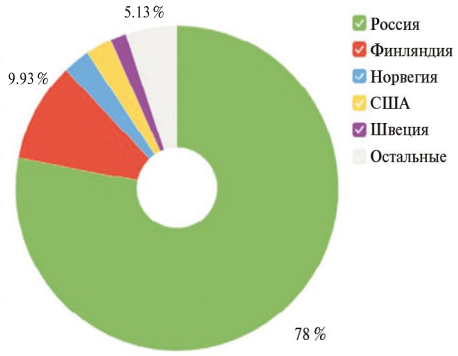

Рис. 8. Статистика посещений корпуса за январь–март 2024 г.

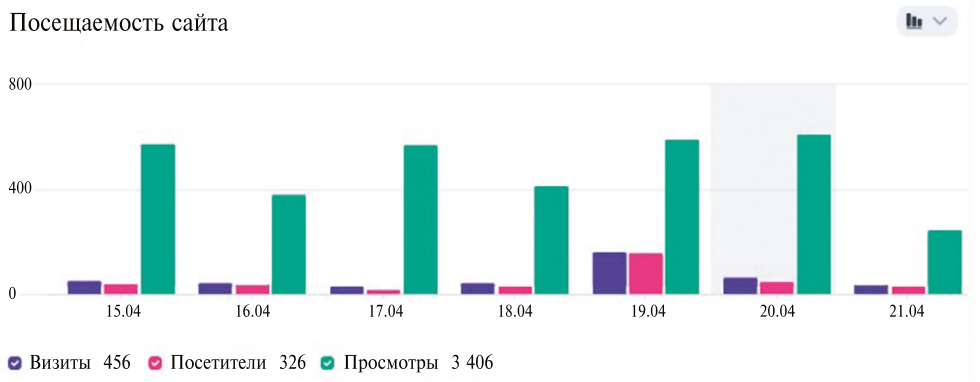

Рис. 9. Посещаемость корпуса за неделю 15.04.2024 г. – 21.04.2024 г.
Резкое увеличение числа визитов и посетителей 19.04.2024 г. связано с проведением всероссийской акции “Диктант на карельском и вепсском языках”

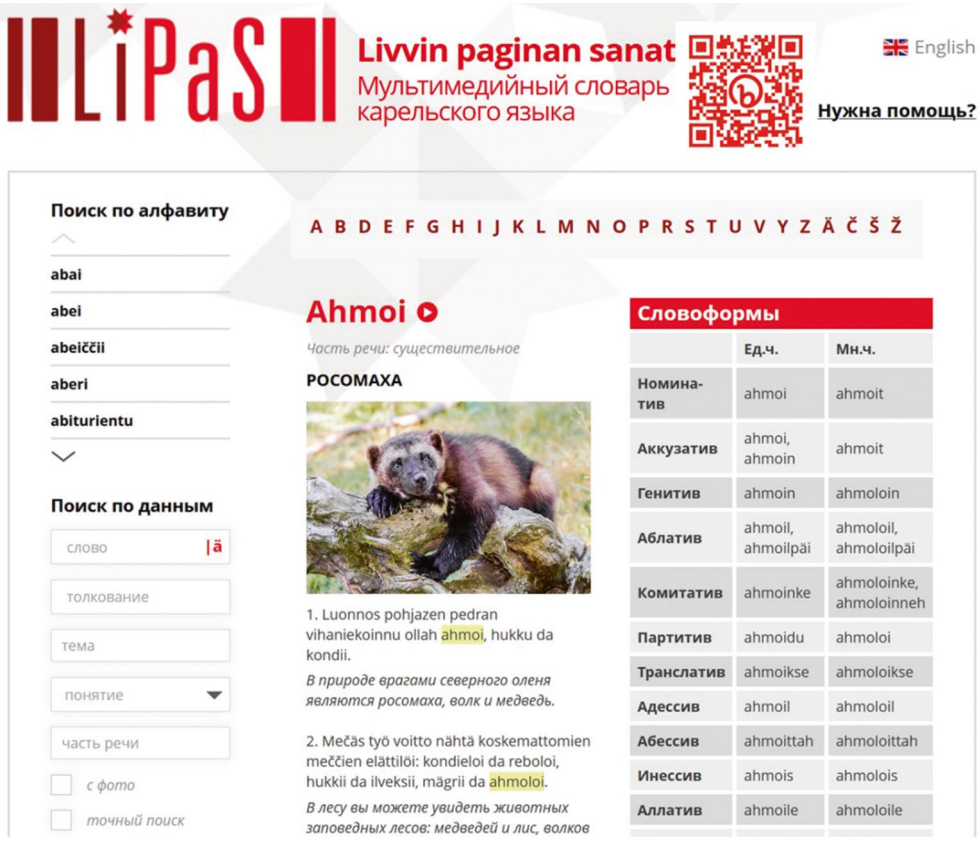

Рис. 10. Пример оформления словарной статьи в LiPaS

человек: пять языковедов, два математика-программиста и один инженер, пополняющий базу текстами. С 2021 г. направление корпусной лингвистики было включено в план научно-исследовательской работы сектора языкознания ИЯЛИ КарНЦ РАН и лаборатории информационных компьютерных технологий ИПМИ КарНЦ РАН, что позволило совершенствовать ресурс в рамках государственного задания. Решать кадровую проблему частично удаётся путём привлечения студентов-лингвистов и студентов-математиков Петрозаводского государственного университета, а также за счёт грантовой поддержки Российского научного фонда [14]. Однако формат конкурсов РНФ и требования к отчётам таковы, что в них сложно вписаться корпусной тематике. Помочь в сложившейся ситуации могла бы организация тематического конкурса проектов, направленных на создание и наполнение корпусных ресурсов, ведь они исключительно важны для сохранения малых языков России.

ЛИТЕРАТУРА

1. Открытый корпус вепсского и карельского языков. http://dictorpus.krc.karelia.ru/ru

   Open corpus of Vepsian and Karelian languages. (In Russ.)

2. Корпус вепсского языка. http://vepsian.krc.karelia.ru/about/

   Corpus of the Vepsian language. (In Russ.)

3. *Коряков Ю.Б., Давидюк Т.И., Харитонов В.С. и др.* Список языков России и статусы их витальности. Монография-препринт. М.: Институт языкознания РАН, 2022.

   Koryakov Yu.B., Davidyuk T.I., Kharitonov V.S. et al. List of languages of Russia and their vitality statuses. Monograph-preprint. Moscow: Institute of Linguistics RAS, 2022. (In Russ.)

4. Итоги Всероссийской переписи населения 2020 г. Т. 5. Табл. 1. Национальный состав населения. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Frosstat.gov.ru%2Fstorage%2Fmediabank%2FTom5_tab1_VPN-2020.xlsx&wdOrigin=BROWSELINK

   Results of the Russian Population Census 2020. Vol. 5. Table. 1. National composition of the population. (In Russ.)

5. Итоги Всероссийской переписи населения 2020 г. Т. 5. Табл. 4. Владение языками и использование языков населением. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Frosstat.gov.ru%2Fstorage%2Fmediabank%2FTom5_tab4_VPN-2020.xlsx&wdOrigin=BROWSELINK

Results of the Russian Population Census 2020. Vol. 5. Table. 4. Language proficiency and language use by the population. (In Russ.)

6. Atlas of the world's languages in danger. Paris: Imprimerie Leclerc, 2010.
7. *Бойко Т.П., Зайцева Н.Г., Крижановская Н.Б. и др.* Лингвистический корпус ВепКар – "заповедник" прибалтийско-финских языков Карелии // Труды Карельского научного центра Российской академии наук. 2021. № 7. С. 100–115.

   Boyko T.P., Zaitseva N.G., Krizhanovskaya N.B. et al. The Linguistic Corpus VepKar is a Language Refuge for the Balticfinnish Languages of Karelia // Proceedings of the Karelian Research Centre of the Russian Academy of Sciences. 2021, no. 7, pp. 100–115. (In Russ.)
8. *Boyko T., Zaitseva N., Krizhanovskaya N. et al.* The Open corpus of the Veps and Karelian languages: overview and applications // KnE Social Sciences. 2022, no. 3, pp. 29–40.
9. *Родионова А.П., Крижановская Н.Б., Пеллинен Н.А.* Речевой корпус ВепКар как инструмент сохранения диалектной речи прибалтийско-финских народов Карелии // Ежегодник финно-угорских исследований. 2023. № 3. С. 343–351.

   Rodionova A.P., Krizhanovskaya N.B., Pellinen N.A. VepKar speech corpus as a tool to preserve the dialect speech of the Baltic-Finnish people of Karelia // Yearbook of Finno-Ugric Studies. 2023, no. 3, pp. 343–351. (In Russ.)
10. Аудиокарта прибалтийско-финских языков Карелии и сопредельных областей. http://dictorpus.krc.karelia.ru/ru/corpus/audiotext/map

    Audio map of the Baltic-Finnish languages of Karelia and adjacent regions. (In Russ.)
11. *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.

    Novak I.P., Krizhanovskaya N.B., Boiko T.P., Pellinen N.A. Development of rules of generation of nominal word forms for new-written variants of the Karelian language // Bulletin of Ugric Studies. 2020, no. 10 (4), pp. 679–691. (In Russ.)
12. *Зайцева Н.Г., Харитонова Е.Е., Жукова О.Ю.* Орфографический словарь вепсского языка. Петрозаводск: КарНЦ РАН, 2012.

    Zaitseva N.G., Kharitonova E.E., Zhukova O.Yu. Spelling dictionary of the Vepsian language. Petrozavodsk: Karelian Research Center RAS, 2012. (In Russ.)
13. *Бойко Т.П.* Грамматико-орфографический словарь карельского языка. Петрозаводск: Periodika, 2022.

    Boyko T.P. Grammar and spelling dictionary of the Karelian language. Petrozavodsk: Periodika, 2022. (In Russ.)
14. Проект РНФ "Создание речевого корпуса прибалтийско-финских языков Карелии". https://rscf.ru/project/22-28-20215/

    RSF project "Creation of a speech corpus of the Baltic-Finnish languages of Karelia." (In Russ.)

THE OPEN CORPUS OF THE VEPSIAN AND KARELIAN LANGUAGES

I.I. Mullonen[a,*], I.P. Novak[a,]**

[a] *Institute of Linguistics, Literature and History of the Karelian Research Centre of the Russian Academy of Sciences, Petrozavodsk, Russia*

**E-mail: irma.mullonen@hotmail.com*

***E-mail: bel.irina@rambler.ru*

In order to preserve and systematically study the Vepsian and Karelian languages, the staff of the Institute of Language, Literature and History and the Institute of Applied Mathematical Research of the KarSC RAS have created and continue to improve the VepKar language corpus. The project is aimed at preserving and accumulating written texts in Karelian and Vepsian languages, fixing and preserving oral speech, researching the Baltic-Finnish languages of Karelia, editing the norms of newly written versions of Karelian and Vepsian languages, and creating educational applications. Anyone can use VepKar as an electronic library and a full-fledged electronic dictionary, which makes this resource very popular.

Keywords: VepKar, language corpus, Karelian language, Vepsian language, automatic markup, multimedia dictionary, electronic library, speech subcorpus.