ЗАЙЦЕВА Нина Григорьевна / ZAITSEVA Nina
Институт языка, литературы и истории Карельского научного центра Российской академии наук / Institute of Language, Literature and History, Karelian Research Centre, Russian Academy of Sciences
Россия, Петрозаводск / Russia, Petrozavodsk
zng@ro.ru

КРИЖАНОВСКАЯ Наталья Борисовна / KRIZHANOVSKAIA Natalia
Институт прикладных математических исследований Карельского научного центра Российской академии наук / Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
Россия, Петрозаводск / Russia, Petrozavodsk
nataly.krizhanovsky@gmail.com

КОРПУСНАЯ ЛИНГВИСТИКА В ПРИБАЛТИЙСКО-ФИНСКОМ ИССЛЕДОВАТЕЛЬСКОМ ПРОСТРАНСТВЕ (НА МАТЕРИАЛЕ КОРПУСА ВЕПССКОГО ЯЗЫКА И ОТКРЫТОГО КОРПУСА ВЕПССКОГО И КАРЕЛЬСКОГО ЯЗЫКОВ)*

CORPUS LINGUISTICS IN THE BALTIC-FINNIC RESEARCH AREA (THE CORPUS OF THE VEPS LANGUAGE AND THE OPEN CORPUS OF THE VEPS AND KARELIAN LANGUAGES)

Abstract: The basics of preparing electronic language corpora and some related problems are described. The presented project is focused on a language of a small people (Veps) and another one of a larger kindred people (Karelian). Both are young written languages that require significant textual resources to create written literature, and the presented corpora can contribute to promoting it.

Ключевые слова / Keywords: Корпусная лингвистика, вепсский язык, карельский язык, диалекты, младописьменные языки / Corpus linguistics, Veps language, Karelian language, dialects, young written languages

Тексты на разных языках, которые в языкознании называются обычно «образцами речи», становятся всё более востребованными в исследовательской практике не только лингвистов, но и этнографов, фольклористов. Историки, введя понятие «история повседневности», также активно стали использовать разного рода тексты и магнитофонные записи как источник сведений по истории. В этой связи появление электронных ресурсов с размещёнными в них текстами исключительно востребовано. Что же касается малочисленных народов, языки которых используются в повседневной практике небольшим кругом населения и известны узкому количеству специалистов, то параллельные корпусы, включающие в себя

* Статья подготовлена в рамках программы фундаментальных исследований президиума РАН на 2018 год № 25 «Памятники материальной и духовной культуры в современной информационной среде», проект «Открытый корпус вепсского и карельского языков».

переводы текстов на другие языки (русский, английский) и корпусы с поморфермной нотацией особенно популярны. Из них черпаются различные знания не только по языку, но и по истории народа, его духовной и материальной культуре.

Как известно, набор традиционных методов получения и анализа языковых данных, куда входит сбор текстового материала, его расшифровка, опрос, анкетирование и т. д., в современном языкознании дополняется корпусным методом, а создание лингвистических корпусов — сбалансированных коллекций текстов со встроенным лингвистическим аппаратом (разметкой) — осознаётся как одна из актуальных задач лингвистики. Корпусные ресурсы позволяют повысить прозрачность исследований и достоверность результатов, так как до сих пор исследователь редкого языка часто был практически лишён возможности проверить те или иные данные, приводимые в научных работах.

В распоряжении лингвистов имеются довольно богатые материалы по прибалтийско-финским языкам Карелии и сопредельных областей — по вепсскому и карельскому. Они представлены прежде всего в книгах образцов вепсской речи, изданных в Финляндии и России, диалектных словарях, магнитофонных записях Фонограммархива Института языка, литературы и истории Карельского научного центра Российской академии наук (ИЯЛИ КарНЦ РАН), однако все эти издания уже давно являются раритетными. Кроме того, образцы вепсской речи, изданные в Финляндии, обладают переводами только на финский язык, и их содержание доступно узкому кругу исследователей. В свою очередь, материалы Фонограммархива ИЯЛИ КарНЦ РАН, представляющие собой уникальную коллекцию фонозаписей на вепсском и карельском языках (более 1000 единиц хранения), по большей части даже не расшифрованы и мало кому доступны. Лишь небольшая часть коллекции введена в научный оборот в виде образцов вепсской и карельской речи, а также использована в трудах специалистов института: языковедов, фольклористов, этнологов.

Таким образом, вепсский и карельский языки представлены во многих источниках, однако нельзя утверждать, что они легкодоступны для всех интересующихся. В связи с этим работа по подготовке «Корпуса вепсского языка» (начата в 2009 г.), а затем и «Открытого корпуса вепсского и карельского языков» (ВепКар) была ожидаема, поскольку их материалы могут послужить более глубокому исследованию теоретических основ названных языков, их лексики и семантики, грамматических особенностей, что особенно востребовано и практикой жизни в период возрождения родной словесности вепсским и карельским языками.

Тема «Корпусная лингвистика» активно поддерживается в последние годы программами Отделения филологических наук РАН и Президиума РАН. Поддержку получают такие направления российской корпусной лингвистики, как:

1) создание и развитие корпусных ресурсов по современному русскому языку;
2) создание и развитие корпусных ресурсов по истории русского языка;
3) создание и развитие корпусных ресурсов по языкам народов России.

Среди проектов данного направления поддержку получил и «Корпус вепсского языка», который на момент начала работы являлся практически единственным языком из группы родственных прибалтийско-финских языков, попавших в поле корпусной лингвистики.

Как показывает поиск, среди иных финно-угорских языков России имеется лишь пилотная версия «Корпуса удмуртского языка», представляющая язык прессы 2007–2014 гг. и некоторое количество нехудожественных текстов.

Науке и пользователям известны прежде всего электронные ресурсы наиболее крупных финно-угорских языков, таких как:

- Языковой банк Финляндии;
- Справочный корпус эстонского языка;
- Фонетический корпус спонтанной эстонской речи;
- Венгерский национальный корпус.

На сайте Хельсинкского университета размещены небольшие электронные ресурсы уральских языков, материалы которых широкому кругу пользователей не вполне доступны.

В сети интернет имеется также электронный ресурс вепсских материалов известного финского лингвиста, исследователя вепсского языка Л. Кеттунена[1], подготовленный его учениками и последователями. В своё время он совершил несколько продолжительных поездок в места расселения вепсов, подготовив и опубликовав работы, которые до сих пор являются настольными книгами у вепсологов и специалистов по прибалтийско-финским языкам[2]. Данный ресурс с двумя системами поиска по населённым пунктам и словам особенно активно используется исследователями вепсского языка, поскольку содержит точные и исключительно важные для диалектолога паспортизованные сведения о записи материала в определённых населённых пунктах вепсской территории.

Совсем недавно появился электронный ресурс, который его создатели назвали *«Raja-Karjalan korpus»*[3] («Корпус карельских диалектов Приграничной Карелии»).

[1] Vepsän verkkosanasto // Kotimaisten kielten keskus [Электронный ресурс]. URL: http://kaino.kotus.fi/sanat/vepsa/ (27.12.2018).

[2] *Kettunen L.* Lõunavepsa häälikajalugu.. Tartu, 1922. Vol. I. Konsonantid; Vol. II. Vokaalid; *Idem.* Vepsän murteiden lauseopillinen tutkimus. Helsinki, 1943 и др.

[3] Raja Karjalan korpus // Kielipankki. URL: https://korp.csc.fi/download/finka/ (27.12.2018); The Corpus of Border Karelia // Metashare [Электронный ресурс]. URL: http://urn.fi/urn:nbn:fi:lb-2014073033 (27.12.2018).

Данный корпус содержит 119 часов расшифрованных интервью и их фонозаписи с представителями пяти приграничных диалектов карельского языка. Работа по его созданию поддерживалась Академией наук Финляндии. Это довольно крупный электронный ресурс современных бесед со знатоками карельского языка, который пока не содержит системы поиска и может быть скачан желающими на собственные носители. Он может показать, в каком состоянии находятся диалекты в настоящее время, каковы их особенности и перспективы развития.

Что касается «Корпуса вепсского языка», то его подготовка была начата в 2009 г. при поддержке программы Президиума РАН. Его цель первоначально была исключительно скромной — создание корпуса оригинальных вепсских устных и письменных текстов небольшим объёмом и размещение их в открытом доступе в сети интернет.

Достижение этой цели предполагало последовательное решение ряда *задач*.

1. Определение круга оригинальных устных и письменных текстов на вепсском языке и составление их репрезентативной коллекции, что предполагает дополнительный сбор необходимых образцов аутентичных вепсских текстов.

2. Урегулирование вопросов авторского права. Поскольку сбалансированный корпус предполагает использование некоторого числа художественных текстов, достижение соглашения с правообладателями необходимо.

3. Компьютерная подготовка текстов (набор и редактирование). Сейчас значительная часть вепсских текстов, которой располагает ИЯЛИ КарНЦ РАН, хранится не в электронном виде. Многие аудиозаписи не расшифрованы, поэтому предстояла серьезная работа по расшифровке, сканированию, оцифровке, набору и редактированию текстов.

4. Лемматизация (отождествление всех словоформ одной лексемы). Эта в целом легко автоматизируемая задача для вепсского языка не имела простого решения: весь материал корпуса следовало размечать вручную.

5. Разработка схемы метаразметки и детальной паспортизации текстов в соответствии с этой разметкой. Каждый текст описывается по строгой схеме, учитывающей пол и дату рождения автора текста, место и время записи, принадлежность диалекту, жанр и некоторые другие признаки.

6. Создание интернет-сайта на русском языке (а по возможности и английском и/или финском языках) с поисковым механизмом.

На основе имеющегося опыта по корпусной лингвистике, который уже известен науке, были решены проблемы метаразметки, паспортизации текстов и т. д. Проблема лемматизации оказалась наиболее сложной. В качестве леммы было избрано слово вепсского письменного языка, к которому привязывались все прочие диалектные формы. Оказалось, что для изучающих вепсский язык студентов, которые были привлечены к работе в корпусе, эта задача была непосильной, поскольку они не могли спроецировать и привязать диалектную форму к лемме.

Трудно было решить, например, что форма адессива вепсского младописьменного языка *sild* ‘мост’ — *sildal* ‘на мосту’ (язык северных вепсов, младописьменный вепсский язык) может в восточных говорах вепсского языка звучать как *süuduu*. Избрать же в качестве леммы какой-либо диалект было ещё сложнее, поскольку только представители местного говора смогли бы правильно образовать все формы, которые выступают у именных частей речи и глаголов в качестве заглавных, например, в словарях. Таким образом, оказалось, что для этой работы нужны специалисты с хорошим или абсолютным знанием языка.

Небольшим коллективом разработчиков цель проекта «Корпус вепсского языка», которая заключалась в создании, а затем в дальнейшем пополнении и развитии компьютерной онлайн-системы, в основном достигнута. И научный мир, и все заинтересованные пользователи обладают в настоящее время впервые созданным в отечественной науке доступным вепсским электронным ресурсом[4]. Этот ресурс содержит тексты и словарь. Тексты расположены в пяти подкорпусах:

1) подкорпус диалектных текстов;

2) два подкорпуса фольклорных текстов: а) подкорпус вепсских причитаний, б) подкорпус вепсских народных сказок;

3) тексты переводов Библии;

4) младописьменный подкорпус публицистических и художественных текстов на языке вепсов.

Подкорпусы обладают рядом собственных характеристик:

1. В рубрике «Диалектные тексты» можно обнаружить тексты на всех трёх диалектах вепсского языка.

2. В рубрике «Вепсские причитания», редкого, но исключительно мало введённого в научный оборот жанра вепсского устного народного творчества, размещены причитания двух жанров — свадебные и похоронно-поминальные на любом диалекте. Они впервые были введены в научный оборот в столь значительном объеме — 88 текстов причитаний. В большинстве своём тексты в этой рубрике были расшифрованы специально для создаваемого корпуса[5].

3. Впервые всему заинтересованному сообществу были представлены также тексты на *младописьменном* языке вепсов: это переводные тексты Нового Завета, которые оформлены как отдельный подкорпус, а также младописьменные публицистические тексты, художественные тексты и тексты для детей. Названный подкорпус был создан также и в связи с участившимися обращениями к специалистам-вепсологам по проблемам функционирования младописьменного языка как входящего в «Красную книгу языков народов России»[6]. В мировой науке данная проблема вызывает особый интерес, проводятся разного типа опросы. Одной

[4] Корпус вепсского языка [Электронный ресурс]. URL: http://vepsian.krc.karelia.ru/about/ (27.12.2018).

[5] Опубликованный книжный вариант причитаний был осуществлён значительно позднее, в 2012 г.

[6] Красная книга языков народов России. М., 1994. С. 21–22

из последних подобных работ, в которой представлен также и младописьменный вепсский язык, следует назвать международный мультидисциплинарный проект *ELDIA* (*European Language Diversity for All*)[7].

Подкорпусы диалектной речи, вепсских сказок, вепсских причитаний выполнены как параллельные, содержат удобно расположенные переводы на русский язык. Система поиска позволяет найти любой текст на нужном диалекте и/или слово электронного словаря. На рис. 1 проиллюстрирован одна страница поиска форм и словоупотреблений лексемы *mec* ‘лес’ в диалектах вепсского языка, где наглядно обнаруживается и паспортизация каждой иллюстрации.

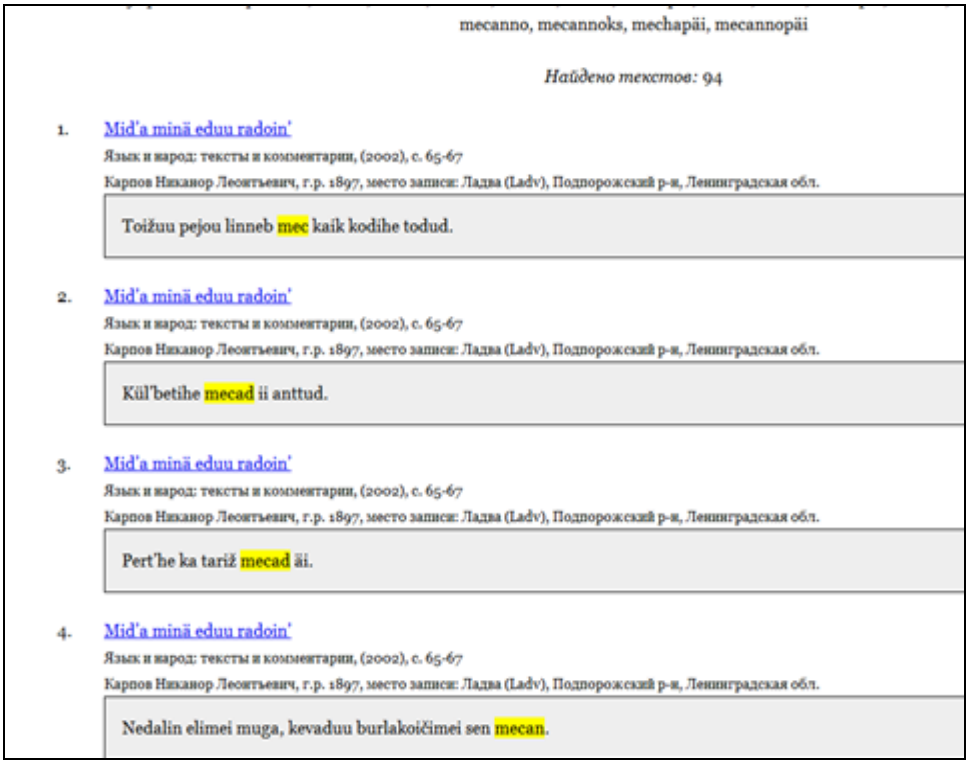

Рис. 1

Поиск форм лексемы *mec* ‘лес’ в «Корпусе вепсского языка»

Важно, что в корпусе с 2014 г. предусмотрен поиск не только по вепсскому, но и по русскому слову, поскольку большая часть пользователей не владеет в полной мере вепсским языком.

Таким образом, прибалтийско-финская лингвистика Карелии вместе с этой работой получила большой и необходимый опыт по созданию электронных языковых ресурсов. В 2016 г. назрела необходимость создания и «Корпуса карельского языка». Было принято решение использовать теоретические

[7] *Puura U., Karjalainen H., Zajceva N., Grünthal R.* The Veps Language in Russia: ELDIA Case-Specific Report. Mainz; Wien; Helsinki; Tartu; Mariehamn; Oulu; Maribor, 2013. URL: https://phaidra.univie.ac.at/detail_object/o:315545 (27.12.2018).

и практические наработки в создании «Корпуса вепсского языка» и с его привлечением создать объединенный «Открытый корпус вепсского и карельского языков» (ВепКар), который в настоящее время уже размещен в сети интернет[8].

Новая система ВепКар унаследовала от Корпуса вепсского языка (1) идею объединения в одно целое текстов и словаря, (2) проработанную структуру метаданных текста (паспорта текстов). При этом система ВепКар является совершенно другой системой, поскольку она:

1) основана на новой платформе *Laravel*. Старый «Корпус вепсского языка» был разработан с нуля без использования каких-либо современных сред. Новая платформа существенно упростила решение ряда задач. Например, благодаря использованию платформы *Laravel* в системе ВепКар кроме поиска по тексту достаточно быстро и без существенных затрат была реализована возможность поиска по разным метаданным текстов;

2) использование платформы *Laravel* привело к тому, что внешне (интерфейс пользователя) и внутренне (структура базы данных и структура программного кода) ВепКар значительно отличается от компьютерной программы «Корпуса Вепсского языка»;

3) словарь в системе ВепКар является многофункциональным;

4) в системе ВепКар добавлена возможность разметки корпуса. На рис. 2 хорошо видна интеграция словаря и корпуса текстов, а также возможности и особенности разметки текстов. Для каждого из значений можно выбрать подходящий пример, то есть предложение с заглавным словом словарной статьи. Предложения можно классифицировать по качеству иллюстрирования значения слова;

5) в словаре у словоформ появилась привязка к диалекту;

6) в рамках словаря создана инфраструктура для формирования многоязычного тезауруса;

7) в новой версии корпуса предложено автоматическое связывание слов текста со значениями лемм в словаре, что исключительно приветствуется пользователями, плохо или не владеющими вепсским и карельским языками;

8) в Корпус добавлена возможность работы с многозначными словами, а в компьютерную программу — функционал для создания и работы с толковым словарём и тезаурусом.

[8] Открытый корпус вепсского и карельского языков VepKar [Электронный ресурс]. URL: http://dictorpus.krc.karelia.ru/ru (27.12.2018).

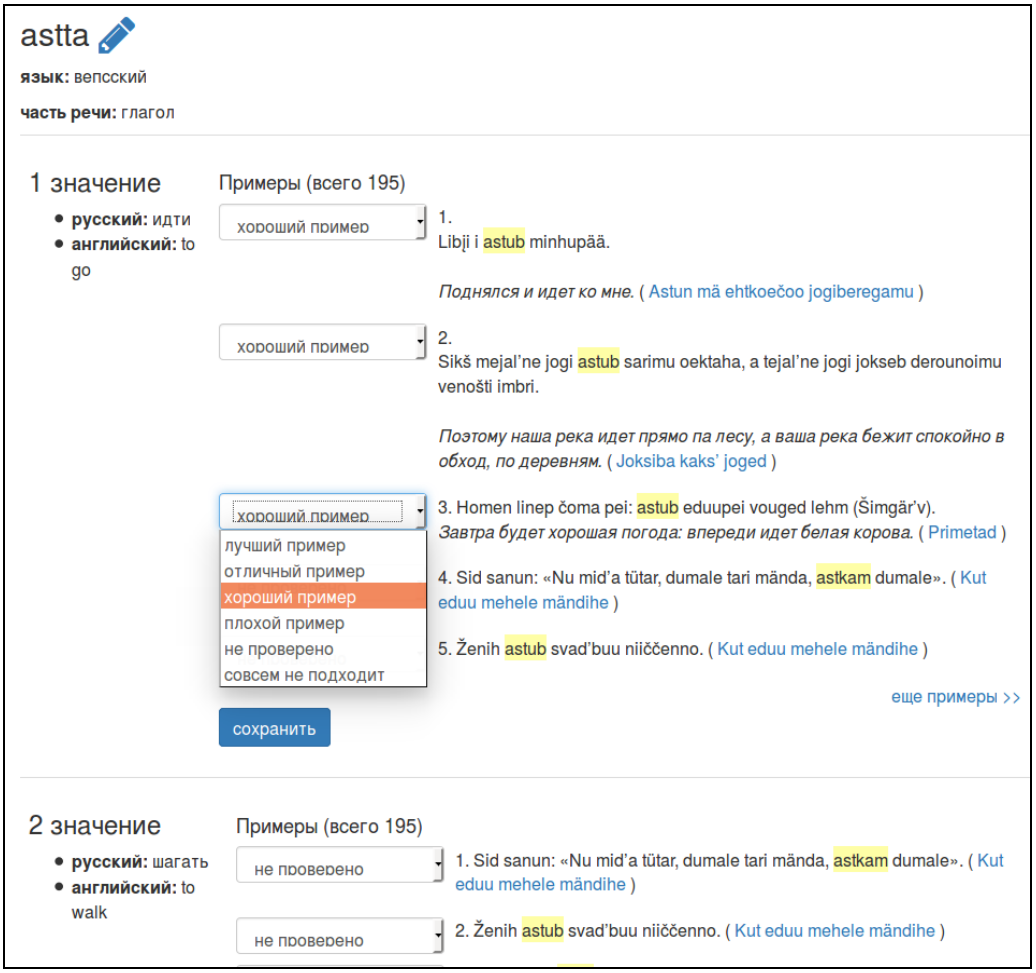

Рис. 2

Словарная статья многозначного вепсского слова *astta* в «Открытом корпусе вепсского и карельского языков» (ВепКар)

Проблемы подготовки названных корпусов и их научная и практическая значимость излагались отчасти и в более ранних публикациях[9]. Они заключаются прежде всего в следующем:

[9] См.: *Зайцева Н. Г.* Вепсские причитания в фокусе корпусной лингвистики и лингвофольклористики // Материалы XLI Международной филологической конференции. 26–31 марта 2012 г. Секция «Уралистика». СПб., 2012. С. 16–26; *Её же.* Корпус вепсского языка: формирование и развитие электронного ресурса // Классический университет в пространстве трансграничности на Севере Европы: стратегия инновационного развития: Материалы международного форума. Петрозаводск, 2014. С. 151–152; *Зайцева Н. Г., Крижановский А. А., Филатова М. М., Шибанова Н. А.* // Корпусная лингвистика-2015: Труды международной конференции. СПб., 2015. С. 2002–2012; *Крижановский А. А., Кириллов А. Н.* Модель геометрической структуры синсета // Труды Карельского научного центра РАН. 2016. № 8. С. 44–54. URL: http://journals.krc.karelia.ru/index.php/mathem/article/view/394 (27.12.2018); *Зайцева Н. Г., Крижановский А. А., Крижановская Н. Б., Пеллинен Н. А., Родионова А. П.* // Труды международной конференции «Корпусная лингвистика-2017». СПб., 2017. С. 172–177. URL: http://scipeople.com/publication/124988/ (27.12.2018).

1) материалы находящихся в работе корпусов должны обеспечить частное и сопоставительное языкознание важным источником лингвистических данных по прибалтийско-финским языкам — вепсскому и карельскому, которые обладают богатой и своеобразной грамматикой, что повысит возможность независимой проверки примеров из названных языков и обогатит инструментарий лингвистов;

2) эти данные могут быть полезными в разработке модели документации языков малочисленных народов, таких как вепсы[10], которые уже в ближайшие десятилетия могут исчезнуть с лингвистической карты России. «Корпус вепсского языка», а также «Открытый корпус вепсского и карельского языков» могут стать своеобразным музеем вепсского и карельского языков с широким и доступным кругом экспонатов, на который можно было бы ориентироваться при описании других языков, в том числе и малочисленных народов, к каковым относится язык вепсов;

3) в настоящее время, в момент ревитализации вепсского и карельского языков, создание многих учебников и учебных пособий, корпусы нашли и ненаучную сферу применения. Они оказались востребованными широким кругом пользователей: редакторами, педагогами, студентами, школьниками, переводчиками — всеми, кто интересуется вепсским и карельским языками или работает с ними. Названные корпусы являются для них не исследовательским инструментом, а источником практических сведений о языках, способным быстро удовлетворить запрос, например, направленный на выявление значений слов, их употреблений, образования различных конструкций и т. п.

Список литературы

Зайцева, Н. Г. Вепсские причитания в фокусе корпусной лингвистики и лингвофольклористики / Н. Г. Зайцева // Материалы XLI Международной филологической конференции. 26–31 марта 2012 г. Секция «Уралистика». — Санкт-Петербург : Филологический факультет СПбГУ. 2012. — С. 16–26.

Зайцева, Н. Г. Корпус вепсского языка: формирование и развитие электронного ресурса / Н. Г. Зайцева // Классический университет в пространстве трансграничности на Севере Европы: стратегия инновационного развития : материалы международного форума. — Петрозаводск : Издательство ПетрГУ, 2014. — С. 151–152.

Зайцева, Н. Г. Корпус вепсского языка / Н. Г Зайцева, А. А. Крижановский, М. М. Филатова, Н. Л. Шибанова // Корпусная лингвистика-2015 : труды

[10] Красная книга языков народов России. С. 21–22.

международной конференции. — Санкт-Петербург : Изд-во СПбГУ, 2015. — С. 2002–2012.

Зайцева, Н. Г. Открытый корпус вепсского и карельского языков (ВепКар): предварительный отбор материалов и словарная часть системы / Н. Г. Зайцева, А. А. Крижановский, Н. Б. Крижановская, Н. А. Пеллинен, А. П. Родионова // Труды международной конференции «Корпусная лингвистика-2017». — Санкт-Петербург : Изд-во СПбГУ, 2017. — С. 172–177. — URL: http://scipeople.com/publication/124988/. — (27.12.2018).

Красная книга языков народов России : Энциклопедический словарь-справочник. — Москва : Academia, 1994. — 116 с.

Крижановский, А. А. Модель геометрической структуры синсета / А. А. Крижановский, А. Н. Кириллов // Труды Карельского научного центра РАН. — 2016. — № 8. — С. 44–54. — URL: http://journals.krc.karelia.ru/index.php/mathem/article/view/394. — (27.12.2018).

Kettunen, L. Lõunavepsa häälikajalugu / L. Kettunen. —Tartu : s. n., 1922. — Vol. I. Konsonantid. — 139 lk.; Vol. II. Vokaalid. — 129 lk.

Kettunen, L. Vepsän murteiden lauseopillinen tutkimus / L. Kettunen. — Helsinki : Suomalais-ugrilainen seura, 1943. — 576 s.

Puura, U. The Veps Language in Russia: ELDIA Case-Specific Report / U. Puura, H. Karjalainen, N. Zajceva, R. Grünthal. — Mainz; Wien; Helsinki; Tartu; Mariehamn; Oulu; Maribor : Research consortium, ELDIA, 2013. — 262 p. — (Staudies in European Language Diversity 25). — URL: https://phaidra.univie.ac.at/detail_object/o:315545. — (27.12.2018).