

*Крижановский А.А., Крижановская Н.Б., Новак И.П.
A.A. Krizhanovsky, N.B. Krizhanovskiy, I.P. Novak*

**ПРЕДСТАВЛЕНИЕ ДИАЛЕКТОВ В ОТКРЫТОМ
КОРПУСЕ ВЕПСКОГО И КАРЕЛЬСКОГО ЯЗЫКОВ
(ВЕПКАР) ¹**

**DIALECTS IN OPEN CORPUS OF VEPS AND KARELIAN
LANGUAGES (VEPKAR)**

Аннотация. В статье описывается организация работы с диалектами карельского языка в Открытом корпусе вепсского и карельского языков (<http://dictorpus.krc.karelia.ru/>). Обсуждаются вопросы, связанные с необходимостью обеспечения технической поддержки словарной части корпуса по заполнению и указанию связей между фонетическими вариантами лемм разных диалектов карельского языка. Показаны различные способы пополнения словаря словоформами.

Ключевые слова. карельский язык, корпус, словарь

Abstract. The article describes the work with dialects of the Karelian language in the Open Corpus of Veps and Karelian languages (<http://dictorpus.krc.karelia.ru/>). The issues related to the variety of Karelian dialects are discussed. Several ways of the dictionary extending with word forms and the connection of phonetic variants of lemmas are shown.

Keywords. Karelian language, corpus, dictionary

¹ Работа выполнена при поддержке РФФИ (проект 18-012-00117).

1. Введение

Открытый корпус вепсского и карельского языков (ВепКар)² содержит словари и тексты прибалтийско-финских языков народов Карелии.

Проект ВепКар является продолжением работ по Корпусу вепсского языка (<http://vepsian.krc.karelia.ru/>), который разрабатывался в 2009-2016 годы сотрудниками Карельского научного центра РАН. В 2016 году проект расширился тремя наречиями карельского языка и пополнился новым функционалом [Зайцева и др. 2017]. На 2019 год в корпусе более 21 тысячи словарных статей и почти 2 тысячи текстов на 27 диалектах вепсского и карельского языков. В проект включены еще 18 диалектов карельского языка, но они пока не отражены в корпусе.

Сотрудники Карельского научного центра РАН заполняют словарь и добавляют тексты в Корпус вепсского и карельского языков. Корпус карельского языка включает три подкорпуса в соответствии с делением языка на три наречия: собственно карельское и ливвиковское, обладающие собственными младописьменными формами, а также людиковское, над нормированным вариантом которого ведется работа в настоящее время (рис. 1).

Для отображения многообразия диалектов и говоров карельского языка появилась возможность при редактировании леммы указывать словоформы по отдельному диалекту. Для установления связей между одинаковыми лексемами в разных диалектах одного наречия был предложен новый вид отношений между леммами — фонетический вариант.

² См. <http://dictorpus.krc.karelia.ru/>

2. О диалектах карельского языка

В корпусе ВепКар три карельских наречия даны отдельно, а не представлены единым языком по ряду причин. Формирование этих наречий произошло в процессе разновременных волн переселений карелов, носителей древнекарельских диалектов, на новые обширные территории и контактов с их автохтонным населением. Таким образом, разделение языка на наречия явилось результатом влияния со стороны как родственных вепсского и финского языков, так и иноязычной системы русского языка, которого невозможно было избежать в условиях многовекового экономического и культурного взаимодействия карельского населения с русским. Основному воздействию оказались подвержены лексическая и фонетическая системы карельских наречий. Развившиеся в них существенные различия диктуют необходимость использования в корпусе ВепКар трех разных словарей карельского языка (данные на февраль 2019 года):

- ливвиковское наречие (17 тысяч лемм),
- людиковское наречие (450 лемм),
- собственно карельское наречие (130 лемм).

Трудности разработки корпуса ВепКар вызваны необходимостью поддерживать многообразие диалектов и говоров карельского языка (рис. 1), сохранить в корпусе и представить читателю наличие взаимосвязей между словами разных диалектов. В частности, в корпусе ВепКар реализованы следующие возможности:

- поиск в словаре по языкам и по наречиям;
- добавление словоформы для отдельного диалекта в процессе редактирования леммы;
- использование «фонетического варианта» для связи одной и той же леммы, имеющей в диалектах разное написание.

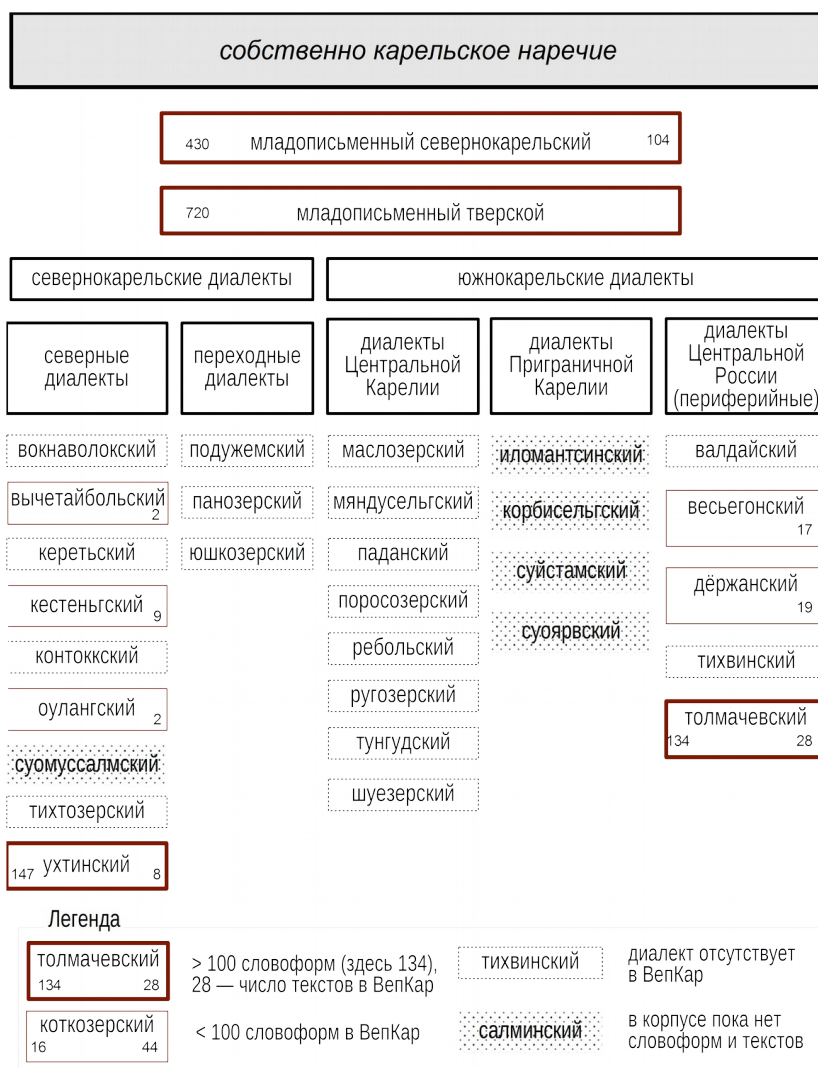


Рис. 1.а. Диалекты собственно карельского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар

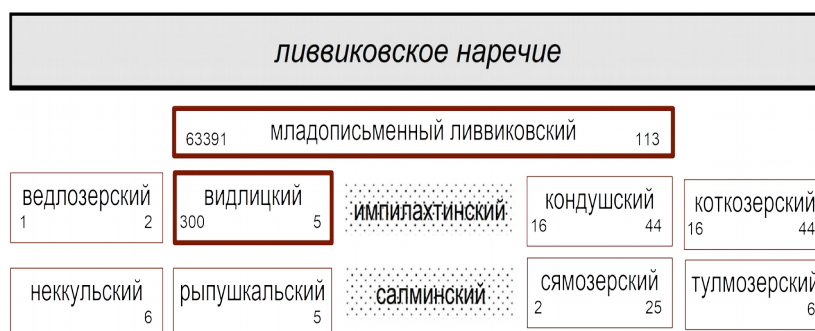


Рис. 1б. Диалекты ливвиковского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар

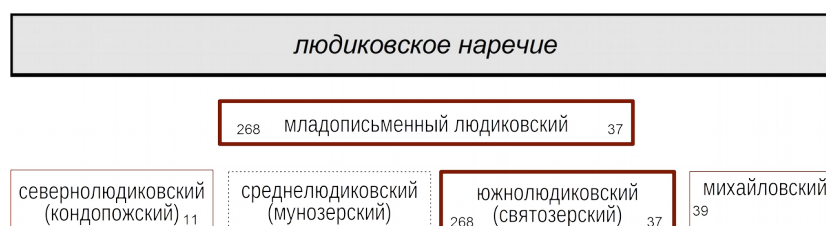


Рис. 1в. Диалекты людиковского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар

Технически «фонетический вариант» реализован в корпусе следующим образом. Корпус ВепКар имеет тезаурус (<http://dictorpus.krc.karelia.ru/ru/dict/lemma/relation>). К восьми семантическим отношениям был добавлен «фонетический вариант» для связывания лексем, имеющих отличия в звучании в разных наречиях и диалектах карельского языка.

Примеры: *kaži* (севернокарельский, ливвиковский, людиковский) / *kazi* (с.к.) / *kasi* (с.к.) 'кошка' [NOM.SG], *kodi* (с.к., ливв., люд.) / *koti* (с.к.) 'дом' [NOM.SG], *aiga* (с.к.) / *aika* (с.к.) /

aigu (ливв.) / *aige* (люд.) / *aig* (люд.) 'время' [NOM.SG], *andua* / *andoa* (с.к., ливв.) / *antua* (с.к.) / *andada* (люд.) / *antta* (люд.) 'давать' [INF], *jogi* / *d'ogi* (с.к., ливв.) / *joki* (с.к.) / *d'og* (люд.) 'река' [NOM.SG].

Схема наречий и диалектов карельского языка (рис. 1) построена по данным ВепКар, использована классификация от 2001 года, предложенная П. М. Зайковым [Зайков, 2001, с. 27], и подробная таблица 2019 года [Новак и др., 2019].

В российском прибалтийско-финском языкознании в карельском языке принято выделять три наречия, каждое из которых распадается на диалекты. Традиционная классификация диалектов карельского языка, основанная на административном принципе (волостное деление территории Карелии начала XX в.), вызывает целый ряд вопросов. В имеющихся вариантах классификации количество диалектов варьируется от 35 до 45. При этом между выделенными диалектами далеко не всегда удается обнаружить наличие каких-либо существенных отличий, а изоглоссы диалектных явлений могут не совпадать с волостными границами. В традиционной классификации, таким образом, речь идет скорее о группах говоров, а проблема диалектного членения оставлена нерешенной. Диалектные материалы корпуса ВепКар в совокупности с его техническими возможностями могут быть использованы в качестве базы для решения этой проблемы.

Корпус на настоящий момент представлен 29 собственно карельскими, 10 ливвиковскими и 4 людиковскими диалектами из традиционной классификации. Всего получается 43 диалекта карельского языка и четыре младописьменные формы (севернокарельский, тверской, ливвиковский и людиковский).

3. Запросы в поисковой системе корпуса ВепКар

При формировании запроса можно уточнить язык, диалект (выбрать вариант из списка), либо искать по всему словарю. При поиске по лемме, по словоформе, толкованию можно указать шаблон. Например, *ka%i*.

В диалектных текстах часто присутствует разнообразное смягчение речи с помощью символа 'i', поэтому в системе помимо словарного написания словоформы, хранится запись для поиска без смягчений. Это не только расширяет поиск по словарю через поисковую форму, но и упрощает автоматическую разметку текстов.

4. Словарь корпуса ВепКар и генерация словоформ

По числу словоформ в словаре корпуса ВепКар лидируют младописьменный вепский язык (28874 словоформы) и младописьменный ливвиковский (63391)³.

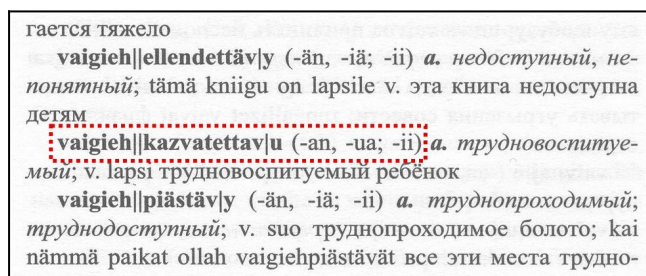


Рис. 2. Фрагмент страницы книги «Большой карельско-русский словарь (ливвиковское наречие)», используемый для генерации словоформ в корпусе ВепКар, см. словарную статью для сложного прилагательного *vaigiehkazvatettavu* ‘трудновоспитуемый’, содержащую автоматически сгенерированные словоформы

³ http://dictorpus.krc.karelia.ru/ru/dict/dialect?limit_num=45

vaigiehkazvatettavu

язык: карельский: ливвиковское наречие

часть речи: прилагательное

1 значение

- русский: трудновоспитуемый

словоформы

No	грамматические признаки	Младописьменный ливвиковский
Единственное число		
1.	номинатив	vaigiehkazvatettavu
2.	генитив	vaigiehkazvatettavan
3.	партитив	vaigiehkazvatettavua
Множественное число		
4.	партитив	vaigiehkazvatettavii

Рис. 3. Словарная статья для сложного прилагательного vaigiehkazvatettavu⁴ в корпусе ВепКар содержит список словоформ (выделен пунктиром), сгенерированных по окончаниям -an, -ua, -ii, представленным в «Большом карельско-русском словаре (ливвиковское наречие)»

Лучшая проработка вепсской части словаря и вепсских текстов объясняется тем, что корпус ВепКар, разрабатываемый с 2016 года, включил в себя и стал продолжением проекта “Корпус вепсского языка”, начало которому было положено в 2009 году [Зайцева, 2017]. Младописьменный ливвиковский занимает сейчас лидирующее положение в корпусе по числу лемм и словоформ, поскольку в 2018 году в словарь корпуса почти в полном объёме был добавлен лексический материал издания Бойко Т.П. (Большой карельско-русский словарь (ливвиковское

⁴ См. <http://dictorpus.krc.karelia.ru/ru/dict/lemma/21943>

наречие). Петрозаводск, 2016), что составило около 17 тысяч лемм и 63 тысячи словоформ (рис. 2).

Для ускорения процесса ввода ливвиковских слов были добавлены новые модули автоматического разбора текста словарной статьи. Если раньше нужно было сначала заполнить отдельно поля формы создания леммы, сохранить, потом открыть форму редактирования словоформы, вручную внести четыре словоформы из словарной статьи, то теперь достаточно в поле леммы вставить фрагмент словарной статьи вместе с окончаниями (например, “**vaigieh||kazvatettav|u** (-an, -ua; -ii)”) (рис. 2) и система автоматически разбирает текст, выделяет лемму, убирает разделители |, затем по основе леммы и окончаниям формируются словоформы (рис. 3) и записываются в базу данных корпуса.

Заключение

В статье описана организация работы с диалектами в корпусе ВепКар, в том числе указание связей между словами разных диалектов карельского языка. Показаны способы полуавтоматического пополнения словаря словоформами. Представлена схема наречий и диалектов карельского языка, построенная на основе данных корпуса ВепКар и ряда источников. Традиционно в языке выделяют три наречия, каждое из которых включает ряд диалектов: 29 собственно карельских, 10 ливвиковских и 4 людиковских. Таким образом, карельский язык представлен 43 диалектами, а также четырьмя младописьменными формами (севернокарельской, тверской, ливвиковской и людиковской).

Проблема диалектного членения карельского языка до настоящего момента не решена. Современная классификация карельских диалектов основана на административном принципе,

а не на лингвистических критериях. И именно диалектные материалы корпуса и его поисковые возможности, значительно ускоряющие процесс работы с ними, обеспечат языковедов данными для численных и теоретических выводов о классификации диалектов карельского языка.

Литература

1. *Зайков П.М.* (2001), *Глагол в карельском языке*. Петрозаводск: ПетрГУ.

2. *Зайцева Н.Г., Крижановский А.А., Крижановская Н.Б., Пеллинен Н.А., Родионова А.П.* (2017), *Открытый корпус вепсского и карельского языков (VepKar): предварительный отбор материалов и словарная часть системы // Труды международной конференции «Корпусная лингвистика – 2017»*, СПб., с. 172-177.

3. *Новак И., Пенттонен М., Руусканен А., Сиилин Л.* (2019), *Карельский язык в грамматиках. Сравнительное исследование фонетической и морфологической систем*. Петрозаводск: КарНЦ РАН, с. 22.

References

1. *Zaikov P.M.* (2001), *Glagol v karel'skom yazyke [Verb in Karelian language]*. Petrozavodsk: PetrSU.

2. *Zaitseva N.G., Krizhanovsky A.A., Krizhanovskaya N.B., Pellinen N.A., Rodionova A.P.* (2017), *Otkrytyy korpus vepsskogo i karel'skogo yazykov (VepKar): predvaritel'nyy otbor materialov i slovarnaya chast' sistemy [Open corpus of Veps and Karelian languages (VepKar): preliminary selection of materials and dictionary of the system] // International scientific conference "Corpus linguistics"*. Saint Petersburg, pp. 172-177.

3. Novak I., Penttonen M., Ruuskanen A., Siilin L. (2019), Karel'skiy yazyk v grammatikakh. Sravnitel'noe issledovanie foneticheskoy i morfologicheskoy sistem – Petrozavodsk: KarRC RAS, p. 22.

Крижановский Андрей Анатольевич

Институт прикладных математических исследований
Карельского научного центра РАН.

Krizhanovsky Andrew

Institute of Applied Mathematical Research, Karelian Research
Centre, RAS

E-mail: andrew.krizhanovsky@gmail.com

Крижановская Наталья Борисовна

Институт прикладных математических исследований
Карельского научного центра РАН.

Krizhanovskaya Natalia

Institute of Applied Mathematical Research, Karelian Research
Centre, RAS

E-mail: nataly.krizhanovsky@gmail.com

Новак Ирина Петровна

Институт языка, литературы и истории Карельского
научного центра РАН.

Novak Irina

Institute of Linguistics, Literature and History, Karelian
Research Centre, RAS

E-mail: bel.irina@rambler.ru