

Архитектура корпусного менеджера и разметка текстов корпуса

ВепКар¹

Аннотация

Открытый корпус вепского и карельского языков (ВепКар) представляет интерес не только для лингвистов, но и для диалектологов, поскольку диалектные тексты даны с подробной паспортизацией (указано имя информанта, год и место его рождения, место записи, диалект и так далее). В статье рассмотрен наименее проработанный людиковский подкорпус. Кратко представлена архитектура разработанного корпусного менеджера *dictorpus*, обслуживающего корпус ВепКар. Описана разметка текстов, семантическая и грамматическая связь между словами в текстах корпуса и словарными статьями. В планы развития корпуса входит семантическая категоризация значений лемм на основе сопоставительно-ономасиологического словаря и разработка морфологического анализатора на основе конечных автоматов.

Ключевые слова: вепский язык, карельский язык, корпусная лингвистика, разрешение лексической многозначности, разметка текста

Abstract

The Open corpus of Veps and Karelian languages (VepKar) is useful both for linguists and dialectologists, since dialect texts are given with a detailed description: informant's name, year and place of his birth, place of recording, dialect, etc. The article discusses the features of the smallest Ludian subcorpus. The architecture of the developed corpus manager *dictorpus* (used in the VepKar) is presented in the article. Semantic and morphological tags links texts with VepKar dictionary entries. Semantic categorization of the meanings of lemmas is planned to be based on

¹ Статья подготовлена в рамках проекта РФФИ 18-012-00117 «Проблемы создания корпусов языков малочисленных народов России на примере Открытого корпуса вепского и карельского языков»

categories from the comparative onomasiological dictionary. As part of future work, the morphological analyzer will be developed on the basis of finite state machines.

Keywords: Vepsian language, Karelian language, corpus linguistics, word-sense disambiguation, text tagging

История создания

Проект «Открытый корпус вепсского и карельского языков» (ВепКар)² ведет свою историю с 2009 года, когда в Институте языка, литературы и истории Карельского научного центра РАН (ИЯЛИ КарНЦ РАН) под руководством д. ф. н. Нины Григорьевны Зайцевой началась работа по созданию Корпуса вепсского языка. В свое время «Корпус вепсского языка» помог в подготовке «Орфографического словаря вепсского языка» [ОСВЯ, 2012], поскольку обеспечил возможность поиска в диалектном материале при отборе форм для создания правил орфографии вепсского языка. В 2017 году проекту «Корпус Вепсского языка» была присуждена премия ФАДН «Ключевое слово» в номинации «За лучший научный проект».

Представителями вепсского и карельского народов решаются идентичные проблемы по возрождению языков и культур. Это требует значительных языковых ресурсов, которые позволят более надежно создавать и развивать правила также и карельской орфографии, введут в научный и общественный оборот значительное количество материалов, востребованных авторами учебников по языку и истории родного края, учителями, писателями.

В целях сохранения, развития и популяризации не только вепсского, но и карельского языка в 2016 году сотрудники ИЯЛИ и Института прикладных математических исследований КарНЦ РАН приступили к созданию многоязычного корпуса ВепКар на базе Вепсского корпуса. Карельский язык представлен в корпусе тремя наречиями (собственно карельским, ливвиковским и людиковским).

² <http://dictorpus.krc.karelia.ru/>

Архитектура корпуса и объем данных

Корпусным менеджером в проекте ВепКар является разрабатываемая нами компьютерная программа dictorpus. Это программа с открытым исходным кодом, доступна для разработчиков онлайн (<https://github.com/componavt/dictorpus>). Система поддержки корпуса разработана на платформе Laravel, язык программирования PHP. Данные хранятся в базе данных MySQL.

Архитектура корпусного менеджера dictorpus (рис. 1) включает корпус текстов и связанный с ним словарь. На ноябрь 2019 года в корпусе было 2600 текстов на 26 диалектах вепсского и карельского языков, свыше 900 тысяч словоупотреблений³. У некоторых текстов дан параллельный перевод на русский язык, десяток текстов сопровождается видеодорожкой. Словарь корпуса содержит свыше 30 тысяч словарных статей со словоформами на 18 диалектах. Толкования слов в основном даны на русском и английском языках, хотя есть возможность давать толкования на вепсском, наречиях карельского и финском языках.



Рис. 1. Архитектура системы ВепКар, объем корпуса и словаря на ноябрь 2019 года

³ См. статистику корпуса: <http://dictorpus.krc.karelia.ru/ru/stats>

Вместе с текстом хранятся метаданные: информация о записи (информанты, собиратели, год и место записи), информация о месте хранения материалы (обычно в Научном архиве КарНЦ РАН) и выходные данные публикации в печати.

Совместными усилиями идет разметка текста. Тексты разбиты на предложения. Если у текста есть параллельный перевод, то выполнено выравнивание текстов по предложениям (рис. 2). Предложения разбиты на слова (токенизация). Устанавливаются семантические связи (между токенами-словами в тексте и значениями лемм) и грамматические (между словами текста и тройками: (1) словоформа, (2) набор грамматических признаков, (3) диалект).

Система ВепКар при разметке текста автоматически ищет слова из текста среди словоформ из словаря и устанавливает связи между ними. Задача специалиста — проверить эти связи и подтвердить их корректность, либо выбрать верную связь из нескольких возможных, либо добавить новую связь.

Новую связь требуется добавлять в том случае, когда нет подходящей словоформы, либо нет словарной статьи (леммы), либо нет нужного значения в словаре корпуса ВепКар. В этом случае редактор создаёт соответственно новую словоформу, новую словарную статью или добавляет отсутствующее значение леммы.

Akal oli maks	
Корпус: подкорпус вепсских сказок	
Северновепсский	
Информант(ы): Арестова Клавдия Алексеевна, 1907, Матвеева Сельга (Matvejan selg), Прионежский р-н, Республика Карелия	
г. записи: 1957	
записали: Лонин Рюрик Петрович	
Источник: Вепсские народные сказки, (1996), с. 103-105 НА КарНЦ, кол.58, ед. хр. 68	
Akal oli maks (вепсский)	Неверная жена (русский)
Ende eletihe ukk akanke.	Жили-были муж с женой.
Lapsid hiil ii olnu.	Детей у них не было.
Ukk käveskel' kaikušti mecha radole haugod čarpaha, a akk oli kodiš, emagjičihe kodiradol.	Старик всегда ходил в лес дрова рубить, а старуха была дома, занималась домашней работой.
A mužikad ičeze ii g	Мужа своего она не любила.
Oli maks.	Был у нее любовник.
Kut, nägettjk, mužik hānen edes ozuteli	Поскольку мужа своего она не любила, то всегда прикидывалась перед ним больной.
Kerdan läks' mužik	Однажды пошел старик в лес рубить дрова.
Tuli maks besedale,	Пришел к ней любовник на беседу, и стали
paniškat'he.	

Рис. 2. Пример размеченного текста на вепсском языке. Желтым цветом выделено параллельное предложение на русском языке. В вепсском тексте синим цветом выделены слова, для которых система нашла одно значение, красным — более одного значения, зеленым — слово проверено специалистом. По клику на красное слово *kodiš* открывалась панель со всевозможными значениями леммы. Специалист должен выбрать правильное значение.

Для автоматизации процесса пополнения словаря ВепКар словоформами были построены таблицы словоизменения для вепсского и собственно карельского наречия, то есть сформулированы флективные правила, согласно которым по заданной лемме и набору окончаний формируются все возможные основы. Затем по списку основ строятся словоформы. Вепсский язык

отличается большим разнообразием диалектных форм и этому уделялось особое внимание.

Таким образом, вепсская и собственно карельская части корпуса ВепКар существенно выросли, пополнившись громадным количеством словоформ. Например, в 2019 году в словарь корпуса ВепКар было добавлено вручную более 5000 вепских лемм, с ними автоматически были созданы системой ВепКар около 320 тысяч вепских словоформ по флективным правилам. Для карельского языка автоматически (с помощью разработанного парсера словаря) было добавлено 3000 собственно карельских лемм, и соответственно около 130 тысяч собственно карельских словоформ.

Теперь редактор, введя одну строку текста, получает таблицу со 115-117 словоформами для глагола, и до 46 словоформ для именных частей речи. Для ливвиковского и людиковского наречия такая работа ещё предстоит, будут построены таблицы словоизменения для этих наречий с учетом диалектных особенностей.

Любые исследователи, заинтересованные в данных ВепКар, могут обратиться к разработчикам системы и получить данные корпуса в различных форматах. Например, тексты корпуса ВепКар в формате CONLL [Klyachko, 2019] участвовали в организации дорожек для соревнования по морфологическому анализу малоресурсных языков России. Словоформы с грамматической атрибуцией были предоставлены для международного проекта UniMorph (вепский⁴, собственно карельское наречие⁵, ливвиковский⁶ и людиковский⁷).

Людиковские тексты в ВепКаре

В карельском языке выделяются 3 крупных наречия, два из которых (ливвиковское и людиковское) возникли под влиянием языка вепсов, что сделало языки еще более близкими. В нашей статье рассмотрим людиковский подкорпус, поскольку по сравнению с остальными он наименее проработан.

4 См. <https://github.com/unimorph/vep>

5 См. <https://github.com/unimorph/krl>

6 См. <https://github.com/unimorph/olo>

7 См. <https://github.com/unimorph/lud>

Выбирая за основу тот или иной говор, исследователь помимо прочего анализирует его представленность и частотность в размещаемых материалах. Абсолютно все говоры, представленные в классификации П.М. Зайкова [Зайков 2000, 27], отразить в корпусе невозможно, так как часть из них либо практически исчезла, либо находится на грани исчезновения. Речь идёт, прежде всего, о людиковских говорах, носителей которых осталось, по неофициальным данным, не более 300 [Родионова 2017, 23].

Людиковское наречие в корпусе ВепКар представлено святозерским и михайловским диалектами. Святозерский диалект обладает наибольшим количеством “людиковских черт” по сравнению с михайловским диалектом, на который в большей степени оказал влияние вепсский язык.

В корпусе сейчас 100 людиковских текстов. Для людиковского подкорпуса используются тексты из опубликованных образцов речи [ОЛР 1978; НКК 1994], подготовленных А. П. Баранцевым⁸. Все тексты представляют собой расшифровки магнитофонных записей и передают естественную речь информанта со всеми ее индивидуальными и общезыковыми особенностями. Исследователь представил идиолект, то есть продемонстрировал особенности речи отдельного лица – уроженки деревни Пелдожа Святозерского сельсовета Пряжинского района Карельской АССР. Материал собирался в течение 10 лет: с 1963 по 1973 годы.

А.П. Баранцев планировал издать серию образцов людиковской речи, но, в итоге, был издан только один том. В этот том вошли материалы с сюжетами на три темы:

- описание деревни Пелдожа и ее окрестностей,
- рассказы о жителях деревни Пелдожа,
- повествования о пережитых информантом событиях, свидетелем

которых она являлась.

А.П. Баранцев включил в этот том более ста текстов. Они будут размещены в корпусе.

⁸ На сайте ИЯЛИ КарНЦ РАН в разделе «Имена и даты» размещена краткая информация о научных трудах А.П. Баранцева <http://illhportal.krc.karelia.ru/section.php?id=2663>

Кроме образцов святозерского диалекта А.П. Баранцев собирал материал у северных людиков в деревне Тивдия в 1980-х годах и в селе Михайловское в 1983 году, материал также был расшифрован и опубликован в первом томе сборника «Образцов карельской речи» (Näytteitä karjalan kielestä 1). Все тексты, опубликованные в образцах речи, обладают достаточно тонко представленной транскрипцией, что позволяет глубже прочувствовать фонетическую структуру людиковской речи, в чем А. П. Баранцев был и знатоком, и глубоким специалистом.

Тексты содержатся в параллельных подкорпусах с переводом на русский язык, что удобно для тех, кто не владеет карельским языком (в данном случае людиковским наречием). Диалектные тексты даны с подробной паспортизацией, нужной для работы диалектологов. В паспорте обычно указаны:

- подкорпус, к которому отнесен текст,
- диалект,
- информанты (имя, год рождения, место рождения),
- год и место записи,
- научный сотрудник, записавший текст,
- опубликованный и/или архивный источник текста.

Будущие работы

Планируется пополнять корпус текстами как на младописьменных формах карельского языка, это – поэтические и прозаические произведения, в том числе духовного содержания, статьи из периодической печати и пр., так и диалектными материалами, а именно изданными образцами речи и лично собранными полевыми материалами. Размещаемые тексты будут сопровождаться переводами на русский.

Семантическая категоризация

В сопоставительно-ономасиологическом словаре [СОС 2007] предложена проработанная схема семантической категоризации диалектов карельского, вепсского, саамского языков.

Мы планируем использовать эту схему для создания семантических категорий в словаре ВепКар. При редактировании словарных статей для каждого из значений слов можно будет указать привязку к какой-либо из 1500 категорий из сопоставительно-ономасиологического словаря [СОС 2007].

В первую очередь в словарь ВепКар будут добавлены слова из этого словаря [СОС 2007], поскольку для них категория уже известна. Затем будет решаться задача соотнесения других слов семантическим категориям. Неизбежно схема семантической категоризации будет расширяться и пополняться.

Наличие такой категоризации позволит увидеть список слов, имеющих близкое значение, относящихся к одной категории. Схема категорий представляет собой дерево.

Но задача редакторов словаря усложнится, поскольку им нужно будет хорошо ориентироваться в схеме категорий, чтобы присваивать значениям новых слов верные категории, или создавать новые категории, если их не хватает для покрытия значений слов.

Морфологический анализ

Задача морфологического синтеза заключается в том, чтобы по лемме и морфологическим признакам сгенерировать словоформу. И обратная задача — вычисление по словоформе морфологических признаков и леммы — это задача морфологического анализа.

С работой морфологических анализаторов, лемматизаторов и стеммеров связана задача моделирования недетерминированных конечных автоматов по регулярному выражению. Один из способов анализа и генерации словоформ основан на работе конечных преобразователей, то есть разновидности конечных автоматов. Конечные преобразователи, разработанные несколько десятилетий назад, применяются в решении задач частеречной разметки, фонологического анализа и морфологического анализа.

Мы планируем использовать программное обеспечение для построения морфологических анализаторов (библиотека OpenFst, Helsinki Finite-State

Transducer toolkit, система SMOR) для кодирования морфологических преобразований для вепсского и карельского языков.

Литература

Зайков П.М. Глагол в карельском языке (грамматические категории лица-числа, времени и наклонения). Петрозаводск: Издательство Петрозаводского государственного университета, 2000. 294с.

[ОЛР, 1978] Баранцев А. П. Образцы людиковской речи. Петрозаводск: «Карелия», 1978. 287 с.

[ОСВЯ, 2012] Орфографический словарь вепсского языка = Vepsän kelen orfografine vajehnik / Н.Г. Зайцева, Е.Е. Харитоновна, О.Ю. Жукова. Петрозаводск: Периодика, 2012. 432 с.

Родионова А.П., Нагурная С.В., Чикина Н.В. Людики: вопросы сохранения языка и культуры. Исследования и материалы. Петрозаводск: Карельский научный центр РАН. 2017. 169 с.

[COC, 2007] Сопоставительно-ономасиологический словарь диалектов карельского, вепсского, саамского языков // Под общей редакцией Ю.С. Елисеева и Н.Г. Зайцевой. Петрозаводск, 2007. 346 с.

[Klyachko, 2019] 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) Issue 18 (25). 2019. URL: http://www.dialog-21.ru/media/4870/_-dialog2019scopusvolplus.pdf#page=52

[NKK, 1994] Näytteitä karjalan kielestä I (Образцы карельской речи). Joensuu-Петрозаводск, 1994. 455 с.