УЧЕНЫЕ ЗАПИСКИ ПЕТРОЗАВОДСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Proceedings of Petrozavodsk State University

Т. 45, № 5. С. 89–98 **2023**

Научная статья **Теоретическая, прикладная и сравнительно-сопоставительная лингвистика**
DOI: 10.15393/uchz.art.2023.924
EDN: IBHVUE
УДК 004.522

ИРИНА СЕРГЕЕВНА КИПЯТКОВА
кандидат технических наук, доцент, старший научный сотрудник лаборатории речевых и многомодальных интерфейсов
Санкт-Петербургский федеральный исследовательский центр Российской академии наук
(Санкт-Петербург, Российская Федерация)
ORCID 0000-0002-1264-4458; kipyatkova@iias.spb.su

АЛЕКСАНДРА ПАВЛОВНА РОДИОНОВА
кандидат филологических наук, научный сотрудник лаборатории речевых и многомодальных интерфейсов
Санкт-Петербургский федеральный исследовательский центр Российской академии наук
(Санкт-Петербург, Российская Федерация)
ORCID 0000-0001-5645-9441; santrar@krc.karelia.ru

ИЛЬДАР АМИРОВИЧ КАГИРОВ
научный сотрудник лаборатории речевых и многомодальных интерфейсов
Санкт-Петербургский федеральный исследовательский центр Российской академии наук
(Санкт-Петербург, Российская Федерация)
ORCID 0000-0003-1196-1117; kagirov@iias.spb.su

АНДРЕЙ АНАТОЛЬЕВИЧ КРИЖАНОВСКИЙ
кандидат технических наук, научный сотрудник лаборатории речевых и многомодальных интерфейсов
Санкт-Петербургский федеральный исследовательский центр Российской академии наук
(Санкт-Петербург, Российская Федерация)
ORCID 0000-0003-3717-2079; andrew.krizhanovsky@gmail.com

ПОДГОТОВКА РЕЧЕВЫХ И ТЕКСТОВЫХ ДАННЫХ ДЛЯ СОЗДАНИЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ КАРЕЛЬСКОЙ РЕЧИ

Аннотация. Описывается процесс сбора и подготовки языковых материалов по ливвиковскому наречию карельского языка, необходимых для обучения системы автоматического преобразования карельской речи в текстовую форму. Актуальность создания подобных технологий для карельского языка обусловлена его статусом малоресурсного языка, что является серьезным препятствием для изучения и сохранения. Основной задачей на текущем этапе исследования является первичный сбор и аннотация речевого и текстового корпусов, а также создание словаря транскрипций. В состав речевого корпуса вошли аудиозаписи 15 дикторов (6 мужчин и 9 женщин). Аудиозаписи расшифрованы и сегментированы на отдельные фразы. Объем речевого корпуса после удаления не подходящих для использования фрагментов составил 3,5 часа. Объем текстового корпуса после обработки и удаления повторяющихся предложений составил более 5 миллионов словоупотреблений. На базе собранного текстового корпуса был сформирован словарь для системы распознавания карельской речи. Для всех слов, вошедших в словарь, были автоматически созданы фонематические транскрипции. В дальнейшей работе собранные текстовые и речевые данные будут использоваться для обучения и тестирования системы автоматического распознавания речи на ливвиковском наречии карельского языка.

Ключевые слова: карельский язык, ливвиковское наречие, автоматическая обработка естественного языка, обучение системы распознавания речи, наборы данных, корпусная лингвистика

Благодарности. Исследование выполнено за счет гранта Российского научного фонда № 22-21-00843 «Автоматическое распознавание речи для малоресурсных языков России (на примере карельского языка)».

Для цитирования: Кипяткова И. С., Родионова А. П., Кагиров И. А., Крижановский А. А. Подготовка речевых и текстовых данных для создания системы автоматического распознавания карельской речи // Ученые записки Петрозаводского государственного университета. 2023. Т. 45, № 5. С. 89–98. DOI: 10.15393/uchz.art.2023.924

ВВЕДЕНИЕ

На протяжении последних нескольких десятков лет наблюдается рост числа исследований, посвященных автоматической обработке малоресурсных языков [16]. Факт отсутствия развитых речевых технологий для подобных языков неоднократно становился предметом обсуждения в научном сообществе [11], [21], однако в силу объективных причин [5] работа по приложению современных технологий обработки естественного языка к материалу малоресурсных языков еще далека от завершения.

С 2022 года сотрудниками СПб ФИЦ РАН (г. Санкт-Петербург, Россия) ведутся исследования по разработке системы автоматического распознавания речи для малоресурсных языков России на примере карельского языка. Карельский язык относится к прибалтийско-финской группе финно-угорских языков. На 2020 год карельским языком в России в какой-либо степени владело около 14 тысяч человек[1], и на сегодняшний день от 11 до 20 тысяч человек в Финляндии говорят по-карельски или понимают этот язык [23]. Типологически карельский язык относится к агглютинативным языкам, современная карельская письменность основана на латинице. Принято выделять три основных наречия карельского языка: собственно-карельское, ливвиковское и людиковское [3: 19], [23]. Наиболее близки к карельскому в генетическом плане ижорский и вепсский языки, а также восточные диалекты финского языка; существует точка зрения, согласно которой ливвиковское и людиковское наречия сформировались под сильным влиянием вепсского языка [1], [2: 44], [15]. На территории России карельский язык распространен в Республике Карелия, а также в Тверской, Ленинградской и Мурманской областях. История художественной литературы на карельском языке насчитывает более века, сегодня существуют СМИ на карельском языке (например, газета «Oma Mua»). Поскольку на данный момент ливвиковское наречие является самым распространенным [20: 20], широко представленным в большей части современных изданий на карельском и СМИ, авторами настоящей статьи было принято решение создать систему автоматического преобразования речи в текстовую форму именно для ливвиковского наречия.

Полученные в ходе проекта решения могут иметь определенную значимость для создания автоматических систем транскрибирования не только для карельского, но и для других малоресурсных языков, в том числе систем автоматического распознавания речи и машинного перевода. Кроме того, разработка технологий автоматической обработки языка может способствовать исследованиям карельского языка, предоставляя исследователям эффективный инструмент для записи и обработки карельского языкового материала.

УСТРОЙСТВО СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Прежде чем перейти к описанию собственно языкового материала, собранного для обучающего корпуса, следует дать краткое описание того, что представляют собой системы автоматического распознавания речи и какие данные нужны для их обучения. Можно сказать, что автоматическое распознавание речи – это представление непрерывного речевого сигнала, поступающего от диктора через микрофон или из предварительно записанной базы данных в виде последовательности слов, которая ему соответствует. На рис. 1 представлена упрощенная схема стандартной системы распознавания речи.

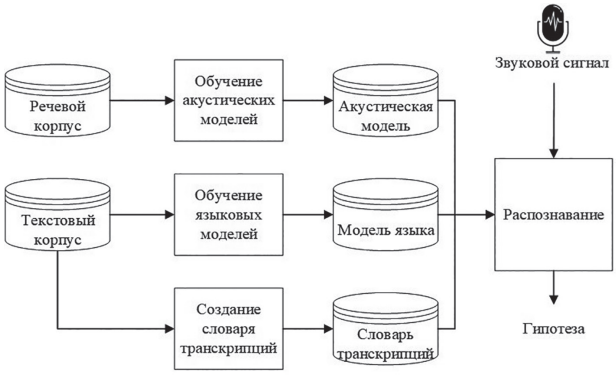

Рис. 1. Общая архитектура стандартной системы распознавания речи

Figure 1. General architecture of a standard speech recognition system

Обычно такая система состоит из акустической модели, устанавливающей взаимосвязь между акустической информацией и аллофонами конкретного языка [13], языковой модели, необходимой для построения компьютером грамматически и лексически правильных гипотез распознанной фразы, а также из словаря слов с транскрипциями, то есть с передачей на письме тем или иным набором графем (фонетическим алфавитом) элементов звучащей речи (звуков). Для обучения акустических моделей используется обучающий речевой корпус, то есть структурированная совокупность речевых аудиоданных, а для обучения модели языка строится вероят-

ностная модель на основе доступных текстов на целевом языке.

Система распознавания речи работает в двух режимах: обучение и распознавание. В режиме обучения создаются акустическая и языковая модели, а также формируется словарь словоформ с транскрипциями, которые будут использоваться в процессе распознавания. В режиме распознавания входной речевой сигнал преобразуется в последовательность векторов признаков, затем производится поиск наиболее вероятной гипотезы с использованием предварительно обученных акустических и языковых моделей [4]. Таким образом, одним из важнейших условий создания системы автоматической обработки естественного языка является наличие обучающих наборов данных (аудио- и текстовых корпусов). Для карельского языка это условие является принципиальным в связи с тем, что он относится к так называемым малоресурсным языкам [12], [19], то есть к языкам с небольшим объемом электронных языковых ресурсов, пригодных для использования в сфере автоматической обработки естественных языков [9].

Перед создателями языкового корпуса обычно возникает ряд проблем, которые вслед за [7] можно разделить на технические, исполнительские, содержательные и структурные. И если последние две проблемы относятся собственно к языковому материалу (выбор дикторов, тип текстового материала, стилистическая и фонетическая сбалансированность текстов и т. п.), то первые две – к организации процесса записи материала и его репрезентации (технические условия записи, формат и качество аудиодорожки, стандарт транскрипции и т. п.). К сожалению, на текущем этапе проекта на первый план вышли именно технические и исполнительские проблемы: участники проекта, располагая достаточно ограниченными организационными, людскими и финансовыми ресурсами, приняли решение сосредоточиться на сборе и обработке, в первую очередь, находящихся в свободном доступе аудиозаписей и текстов, соответствующих определенным техническим критериям (аудиозаписи студийного качества и печатные тексты) и современным стандартам литературного карельского языка (см. [20]).

ТЕКСТОВЫЙ КОРПУС

Текстовый корпус был составлен по материалам печатных изданий, предоставленных издательствами «Периодика» и «Версо», периодики на ливвиковском наречии (в первую очередь газета «Oma Mua», номера 2000–2022 годов), текстов на ливвиковском наречии из открытого корпуса вепсского и карельского языков ВепКар[2] [8], а также ряда других открытых источников. Дополнительно в корпус были включены расшифровки аудиозаписей обучающей части речевого корпуса. Таким образом, в текстовый корпус вошли тексты различного стиля: художественного, публицистического, а также – в небольшом объеме – разговорного. Часть текстов изначально была в нередактируемом формате (сканированные тексты в формате .pdf), поэтому для них выполнялось полуавтоматическое распознавание текста. Все тексты были приведены в формат .txt.

В рамках подготовки текстовых данных была произведена обработка и нормализация текстового материала, которая включала в себя разбиение текста на отдельные предложения, при этом предложения, содержащие прямую и косвенную речь, разбивались на отдельные высказывания, например:

Vie Aleksandr Puškin sanoi: «Pantaloni, frak, žilet – vseh etih slov na russkom net. Šiškov prosti, ne znaju kak perevesti».

<s> vie aleksandr puškin sanoi <s/>

<s> pantaloni frak žilet vseh etih slov na russkom net <s/>

<s> šiškov prosti ne znaju kak perevesti <s/>

'Ещё Александр Пушкин говорил: «Панталоны, фрак, жилет – всех этих слов на русском нет. Шишков, прости, не знаю, как перевести»'.

Tyttölöin nagro havaškoitti händy: «Aiga töhlö olen!» – pyhkijen higie očaspäi, čakkualdi iččiedäh.

<s> tyttölöin nagro havaškoitti händy <s/>

<s> aiga töhlö olen <s/>

<s> pyhkijen higie očaspäi čakkualdi iččiedäh <s/>

'Смех девчонок вызвал у него раздражение: «Какой же я дурак!» – вытирая пот со лба, поругал себя'.

Кроме того, производилось удаление всех текстов, заключенных в скобки, замена заглавных букв на строчные и удаление знаков препинания, замена графемы *ӥ* на *у* (для старых изданий). Дополнительно проводилась проверка на наличие повторяющихся предложений в связи с тем, что тексты брались из разных источников (изданий), в которых могло возникать дублирование материала. Итоговый объем корпуса составил около 5 млн словоупотреблений.

СЛОВАРЬ ТРАНСКРИПЦИЙ

Одним из обязательных условий создания системы автоматического распознавания речи является наличие словаря фонематических транскрипций для слов, которые будут использовать-

ся системой. Для карельского языка был создан инвентарь из 82 фонем, из них 26 гласных и 56 согласных. При этом были выделены ударные и безударные варианты гласных и долгие гласные; как самостоятельная фонема интерпретировался заднерядный аллофон фонемы /i/. Для согласных были выделены твердые и мягкие варианты, а также удвоенные согласные. В таблице представлен используемый фонематический алфавит и приведены примеры транскрипций, созданных автоматически. В скобках дана транскрипция фонем в международном фонетическом алфавите (МФА). В разработанном варианте алфавита знак /!/ используется для обозначения ударения в слове, знак /'/ – для обозначения мягкости согласных, знак /:/ – для обозначения долгих звуков. Знак /ˈ/ до гласной означает ударность в алфавите МФА.

Фонематический алфавит ливвиковского наречия карельского языка
Phonemic alphabet for the Livvi-Karelian dialect

Фонема	Слово	Транскрипция	Фонема	Слово	Транскрипция
/**a!**/ (/ˈ**a**/)	*k**a**la*	/k **a!** l a/	/**k**/ (/**k**/)	***k**ana*	/**k** a! n a/
/**a**/ (/**a**/)	*zirk**a**lo*	/z' i! r k **a** l o/	/**k:**/ (/**k:**/)	*ku**kk**o*	/k u! **k:** o/
/**o**!/ (/ˈ**ɔ**/)	*k**o**tku*	/k **o!** t k u/	/**k'**/ (/**k'**/)	***k**äzi*	/**k'** ae! z' i/
/**o**/ (/**ɔ**/)	*buab**o***	/b **u!** a b o/	/**k':**/ (/**k':**/)	*hä**kk**i*	/h' ae! **k':** i/
/**u!**/ (/ˈ**u**/)	***l**umi*	/l **u!** m' i/	/**l**/ (/**l**/)	***l**ammas*	/**l** a! m: a s/
/**u**/ (/**u**/)	*ikk**u**n*	/i! k: **u** n/	/**l:**/ (/**l:**/)	*ke**ll**o*	/k' e! **l:** o/
/**u:!**/ (/ˈ**u:**/)	*s**uu**ri*	/s **u:!** r' i/	/**l'**/ (/**l'**/)	***l**öyly*	/**l'** oe! y l' y/
/**u:**/ (/**u:**/)	*vil**uu***	/v' i! l **u:**/	/**l':**/ (/**l':**/)	*ve**ll**i*	/v' e! **l':** i/
/**i!**/ (/ˈ**i**/)	*v**i**hmu*	/v' **i!** h m u/	/**m**/ (/**m**/)	***m**uam**o***	/**m** u! a **m** o/
/**i**/ (/**i**/)	*käz**i***	/k' ae! z' **i**/	/**m'**/ (/**m'**/)	***m**issä*	/**m'** i! s': ae/
/**i:!**/ (/ˈ**i:**/)	*h**ii**ri*	/h' **i:!** r' i/	/**m:**/ (/**m:**/)	*ha**mm**as*	/h a! **m:** a s/
/**i:**/ (/**i:**/)	*kyl**ii***	/k' y! l' **i:**/	/**m':**/ (/**m':**/)	*lä**mm**in*	/l' ae! **m':** i n/
/**i^!**/ (/ˈ**i**/)	*c**i**nku*	/ts **i^!** n k u/	/**n**/ (/**n**/)	***n**ogi*	/**n** o! g' i/
/**i^**/ (/**i**/)	*virš**i***	/v' i! r sh **i^**/	/**n'**/ (/**n'**/)	***nen**ä*	/**n'** e! **n'** ae/
/**i^!**/ (/ˈ**i:**/)	*š**ii**loi*	/sh **i^!** l o j/	/**n:**/ (/**n:**/)	*pa**nn**a*	/p a! **n:** a/
/**i^**/ (/**i:**/)	*veššš**ii***	/v' e! sh: **i^**/	/**n':**/ (/**n':**/)	*nän**n**i*	/n' ae! **n':** i/
/**e!**/ (/ˈ**ɛ**/)	*n**e**nä*	/n' **e!** n' ae/	/**p**/ (/**p**/)	***p**ala*	/**p** a! l a/
/**e**/ (/**ɛ**/)	*leh**e**t*	/l' e! h' **e** t/	/**p'**/ (/**p'**/)	***p**äivy*	/**p'** ae! j v' y/
/**ae!**/ (/ˈ**æ**/)	*t**ä**mä*	/t' **ae!** m' ae/	/**p:**/ (/**p:**/)	*lo**pp**u*	/l o! **p:** u/
/**ae**/ (/**æ**/)	*nähtä*	/n' ae! h' t' **ae**/	/**p':**/ (/**p':**/)	*lo**pp**i*	/l o! **p':** i/
/**oe!**/ (/ˈ**œ**/)	*šl**ö**pöi*	/sh l' **oe!** p' oe j/	/**r**/ (/**r**/)	***r**ukku*	/**r** u! k: u/
/**oe**/ (/**œ**/)	*töhl**ö***	/t' oe! h' l' **oe**/	/**r'**/ (/**r'**/)	***r**iähky*	/**r'** i! ae h' k' y/
/**y!**/ (/ˈ**y**/)	*k**y**lä*	/k' **y!** l' ae/	/**r:**/ (/**r:**/)	*ke**rr**an*	/k' e! **r:** a n/
/**y**/ (/**y**/)	*vävy*	/v' ae! v' **y**/	/**r':**/ (/**r':**/)	*py**rr**itys*	/p' y! **r':** i t' y s/
/**y:!**/ (/ˈ**y:**/)	*t**yy**ni*	/t' **y:!** n' i/	/**s**/ (/**s**/)	***s**ana*	/**s** a! n a/
/**y:**/ (/**y:**/)	*väz**yy***	/v' ae! z' **y:**/	/**s'**/ (/**s'**/)	***s**iä*	/**s'** i! ae/
/**b**/ (/**b**/)	*n'a**b**a*	/n' a! **b** a/	/**s:**/ (/**s:**/)	*kos**s**u*	/k o! **s:** u/
/**b**/ (/**b'**/)	*lei**b**ä*	/l' e! j **b'** ae/	/**s':**/ (/**s':**/)	*mis**s**ä*	/m' i! **s':** ae/
/**ts**/ (/**ts**/)	***c**inku*	/**ts** i^! n k u/	/**sh**/ (/**ʃ**/)	*školu*	/**sh** k o! l u/
/**t'ch**/ (/**t'ʃ**/)	*veičel*	/v' e! j **t'ch** e l/	/**sh:**/ (/**ʃ:**/)	*dovariššu*	/d o! v a r' i **sh:** u/
/**t'ch:**/ (/**t'ʃ:**/)	*mečču*	/m' e! **t'ch:** u/	/**z**/ (/**z**/)	*oza*	/o! **z** a/
/**d**/ (/**d**/)	*ka**d**ai*	/k a! **d** a j/	/**z'**/ (/**z'**/)	*vez**i***	/v' e! **z'** i/
/**d'**/ (/**d'**/)	***d**iedoi*	/**d'** i! e d o j/	/**zh**/ (/**ʒ**/)	*kaži*	/k a! **zh** i/
/**d':**/ (/**d':**/)	*la**d'd'**ata*	/l a! **d':** a t a/	/**t**/ (/**t**/)	***t**ukku*	/**t** u! k: u/
/**g**/ (/**g**/)	***g**arbalo*	/**g** a! r b a l o/	/**t'**/ (/**t'**/)	***t**ämä*	/**t'** ae! m' ae/

Окончание таблицы

Фонема	Слово	Транскрипция	Фонема	Слово	Транскрипция
/**g’**/ (/**g’**/)	***g**iidu*	/**g’** i:! d u/	/**t:**/ (/**t:**/)	*ai**tt**u*	/a! j **t:** u/
/**h**/ (/**h**/)	***h**anhi*	/**h** a! n’ h’ i/	/**t’:**/ (/**t’:**/)	*elä**tt**i*	/e! l’ ae **t’:** i/
/**h’**/ (/**h’**/)	***h**iili*	/**h’** i:! l’ i/	/**v**/ (/**v**/)	***v**ačču*	/**v** a! t’ch: u/
/**h:**/ (/**h:**/)	*u**hh**arskoi*	/u! **h:** a r s k o j/	/**v’**/ (/**v’**/)	***v**ibu*	/**v’** i! b u/
/**h’:**/ (/**h’:**/)	*ravahta**hh**äi*	/r a! v a h t a **h’:** ae j/	/**v:**/ (/**v:**/)	*a**vv**oi*	/a! **v:** o j/
/**j**/ (/**j**/)	***j**ogi*	/**j** o! g’ i/	/**v’:**/ (/**v’:**/)	*li**vv**in*	/l’ i! **v’:** i n/

Все транскрипции для словаря создавались автоматически с помощью специально разработанного программного модуля, выполняющего преобразование «графема-фонема» для ливвиковского наречия карельского языка. Поскольку автоматическое распознавание печатных текстов на карельском языке не исключает наличия ошибок, слова, встретившиеся только один раз, чаще всего оказывались неверно распознанными. Поэтому все они были удалены из словаря транскрипций. Общее количество различных слов, встретившихся более одного раза, составило более 135 тысяч.

Для карельского языка создание автоматических транскрипций является относительно простой задачей по сравнению, например, с русским языком, потому что в карельском ударение фиксированное, всегда падающее на первый слог, а гласные чаще всего не подвержены редукции. Таким образом, процедура автоматического транскрибирования сводится к локализации ударения, идентификации сдвоенных графем как репрезентации долгих фонем, а также определению палатализованных согласных (предшествование гласным переднего ряда).

РЕЧЕВОЙ КОРПУС

При работе с малоресурсными языками распространенной практикой является привлечение дикторов, которые читают заранее подготовленные фразы или связный текст. Другим способом сбора материала является использование свободно доступных речевых данных. Речевые данные для проекта, описанного в настоящей статье, были собраны на материале радиопередач «Родной берег» ГТРК «Карелия». Всего были использованы записи 10 передач. Каждая строится в формате интервью, что подразумевает наличие как минимум двух дикторов: интервьюера и интервьюируемого. Следует отметить, что в двух передачах присутствовало больше двух дикторов, а интервьюеры в некоторых повторялись. Интервьюируемые были всегда разные. Таким образом, в записанном речевом корпусе представлено 15 дикторов (6 мужчин и 9 женщин).

Объем речевого корпуса после удаления не подходящих для использования фрагментов составил 3,5 часа, общее число фраз – 3819. Записи были разбиты на отдельные фразы, каждая из которых сохранялась в отдельном wav-файле. На рис. 2 демонстрируется распределение числа фраз по дикторам. В основном на каждого диктора получилось более 100 фраз, кроме дикторов под номером 12, 13 и 14: указанные дикторы принимали участие в тех передачах, где было больше одного интервьюируемого. Корпус разбит на обучающую и тестовую части. В обучающую часть вошли 90 % фраз, в тестовую – 10 %.

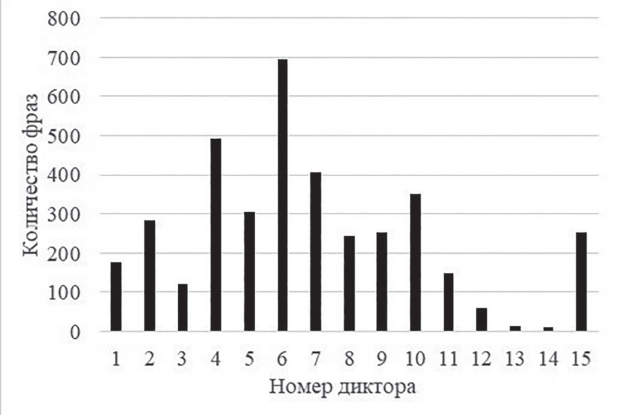

Рис. 2. Распределение количества фраз по дикторам

Figure 2. Distribution of utterances by speakers

Дополнительным инструментом расширения речевого материала была аугментация данных (англ. data augmentation), то есть увеличение выборки данных для обучения через модификацию существующих данных [14], [18], [22]. Аугментация выполняется только для обучающей части речевого корпуса с помощью инструментария Sox[3], посредством которого изменялся темп речи и высота голоса диктора. Темп изменялся с коэффициентом, полученным случайным образом для каждой записи из равномерного распределе-

ния в интервале от 0,7 до 1,3. Высота голоса изменялась на количество полутонов, полученное случайным образом из равномерного распределения от -2 до 2. Кроме того, было выполнено совместное изменение темпа речи и высоты голоса для каждой фразы. Полученные таким образом речевые данные были добавлены к реальным обучающим данным. Общий объем обучающих данных после процедуры аугментации возрос с 3 ч. 8 мин. до 12 ч. 49 мин.

Записанный речевой материал был расшифрован и сегментирован (разбит на отдельные высказывания) специалистами по карельскому языку из состава участников проекта. Примеры аннотации речевого сигнала представлены на рис. 3 (скриншот из программы WaveSurfer[4]).

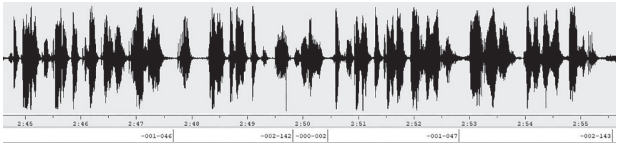

Рис. 3. Пример аннотации речевого сигнала (в строке под осциллограммой первая цифра аннотации обозначает номер диктора, вторая – номер фразы, фразы, помеченные цифрой 000, непригодны для использования)

Figure 3. Example of speech signal annotation (first number in the lower line indicates the speaker, second number indicates the utterance, “000” indicates corrupted utterances)

ОСОБЕННОСТИ СОБРАННЫХ ДАННЫХ

Собранный речевой материал обнаружил ряд особенностей, усложнивших работу над корпусом и заставивших изъять часть данных из итогового корпуса. Одной из основных проблем, выявленных в ходе обработки аудиозаписей, была одновременная речь нескольких дикторов, перебивавших друг друга или говорящих одновременно.

001-053[5]: Äijygo partu pidäy yhteh kylyh ***kuorittua****?*
002-155: ***Yheksä*** *venčia korgevuttu, kymmenendel salbuat.*
001-053: ‘Сколько бревен нужно очистить от коры для строительства одной бани?’
002-155: ‘Высотой в девять венцов, десятым венцом сруб перекрываешь’.

В приведенном примере слова, выделенные жирным шрифтом, были произнесены обоими дикторами одновременно. В эту же категорию попадают «звуковые вставки» и «вокальные жесты» [10] со стороны интервьюера:

002-162: ***sit*** *äski zavodittih* ***vestiä****.*
Диктор 2: ***uhu****….* ***uhu***
002-162: ‘затем начинали тесать бревна’.
Диктор 2: ‘угу... угу...’

Наложения речи трудно поддаются обработке, а удаление подобных фрагментов является нетривиальной задачей, поэтому фразы, содержащие одновременную речь двух дикторов, не были включены в корпус.

Другим фактором, усложнившим сбор аудиокорпуса, были фоновые шумы. Несмотря на то что для создания корпуса использовались только записи студийного качества, в ряде случаев фоновый шум был представлен музыкой, звуком перелистывания страниц, а в одной передаче – уличным шумом. Абсолютно все записи высказываний, в которых присутствовал фоновый шум, были также удалены из базы данных.

Еще одной особенностью, характерной для современного карельского языка, было переключение кодов (англ. code-switching) [6]. Обыкновенно под переключением кодов в лингвистике понимается спонтанный переход говорящего с одного идиома (языка / диалекта) на другой. Действительно, в настоящее время среди носителей карельского языка в России распространено карельско-русское двуязычие (с доминированием русского языка), поэтому переход на русский язык и обратно вполне естественен. Однако следует учитывать то, что в карельском языке (и это особенно справедливо для разговорной речи) присутствует мощный пласт заимствований из русского, фактически ставший частью языка. Примеры хорошо известны [17: 132]. В речевом материале, который использовался для создания корпуса, это были в первую очередь вводные слова и местоимения:

Minä en tiijä, kai on, ***naverno****, allun algu pandu.*
‘Я не знаю, всё, наверно, изначально установлено’.

Salbomestu nostamah ***konešno*** *ei tule*
‘Поднимать сруб, конечно, не придет’.

Иногда дикторы вставляли в речь целые словосочетания и клаузы на русском:

Sit se oli, häi pereimenovalis’ kolhozakse, ***kolhoz imeni Papanina*** *rodiiheze.*
‘Потом это было, переименовались в колхоз, стал (называться) колхоз имени Папанина’.

Kelle penziessäh pideli ruadua, net vahtoi myöte ruatah ***libo PMK-poka mamka kormit****.*
‘Кому до пенсии нужно было работать, те на вахтах работают, или на ПМК – «пока мамка кормит»’.

При этом письменные тексты, использованные для создания корпуса, были подвергнуты литературной обработке, вследствие чего в них оказалось относительно немного заимствований, поэтому в словарь системы вошло

очень мало русскоязычных слов. Обработка переключения кодов при распознавании речи требует специальных подходов, которые на первоначальном этапе разработки системы применять не планировалось. Поэтому все высказывания, в которых обнаруживалось переключение кодов, были удалены из речевого корпуса.

Особый случай составляют имена собственные: в подавляющем большинстве случаев они заимствуются из русского языка и произносятся в соответствии с нормами русской фонетики, в частности ударение в именах будет плавающим в соответствии с русским произношением. Эта проблема пока не решена, однако представляется, что самым рациональным решением будет составление отдельного словника имен собственных и их транскрибирование в соответствии с фонетикой русского языка. Ниже приведены примеры использования имен собственных в речевом корпусе:

003-173: ***Aleksandr Lukič****, työ oletto tozi ozavu ristikanzu, gu elittö moizen pitkän, moizen bohatan tapahtumil elaijan.*
003-173: 'Александр Лукич, Вы – по-настоящему счастливый человек, поскольку прожили такую длинную, богатую на события жизнь'.

010-107: ***Anna Vasiljevna*** *oli* ***Jakovleva****, ruavos täs oli* ***Korol'ov Večeslav Konstantinovič****, Karellesproman putin mužikku.*
010-107: 'Анна Васильевна Яковлева была, здесь работала, был Королёв Вячеслав Константинович, настоящий мужик из Кареллеспрома'.

ЗАКЛЮЧЕНИЕ

На текущем этапе исследования, посвященного разработке системы распознавания речи на карельском языке, его участниками были решены следующие задачи:

– подготовка речевых данных на карельском языке (ливвиковское наречие);

– подготовка текстовых данных на карельском языке (ливвиковское наречие);

– создание фонематических транскрипций для слов из текстового корпуса.

Тем не менее очевидно, что для повышения качества распознавания речи (а также для повышения значимости собранного корпуса для лингвистических исследований из более широкой области) необходимо продолжить работу над сбором и обработкой текстов на карельском языке. В первую очередь это касается речевых данных:

1. Относительно малый объем речевых данных;

2. Несбалансированность дикторов: в корпусе женщин больше, чем мужчин; кроме того, в корпусе представлены записи в основном дикторов среднего и старшего возраста. Эта ситуация связана, судя по всему, с социолингвистическими реалиями, в которых находится современный карельский язык: число носителей среди молодежи достаточно мало [17: 95];

3. Несбалансированность дикторов по количеству фраз. Для повышения дикторонезависимости системы следует, помимо увеличения собственно количества дикторов, добиваться их одинаковой представленности в записанном материале.

В дальнейшем участники настоящего проекта планируют увеличить объем речевых и текстовых данных для обучения системы, что позволило бы перейти от лабораторного прототипа к робастной системе, способной работать как в условиях фоновых шумов, так и с большим количеством дикторов.

ПРИМЕЧАНИЯ

[1] https://www.ethnologue.com/language/krl/
[2] http://dictorpus.krc.karelia.ru/ru
[3] http://sox.sourceforge.net/sox.html
[4] https://wavesurfer-js.org/
[5] В номерах, открывающих каждую строку, число до дефиса обозначает номер диктора, число после дефиса – номер высказывания.

СПИСОК ЛИТЕРАТУРЫ

1. Афанасьева А. А., Муллонен И. И. Карело-вепсский диалог на карте южной Карелии // Acta Linguistica Petropolitana. 2020. Т. 16, № 3. С. 9–28. DOI: 10.30842/alp2306573716301
2. Бубрих Д. В. Происхождение карельского народа. Повесть о союзнике и друге русского народа на Севере. Петрозаводск: Госиздат Карело-Финской ССР, 1947. 53 с.
3. Зайков П. М. Глагол в карельском языке. Петрозаводск: Петрозаводский гос. ун-т, 2000. 294 с.
4. Кипяткова И. С. Комплекс программных средств обработки и распознавания разговорной русской речи // Информационно-управляющие системы. 2011. Т. 53, № 4. С. 53–59.

5. Кипяткова И. С., Кагиров И. А. Аналитический обзор методов решения проблемы малых наборов данных при создании систем автоматического распознавания речи для малоресурсных языков // Информатика и автоматизация. 2022. Вып. 21, Т. 4. С. 678–709. DOI: 10.15622/ia.21.4.2
6. Ковалева С. В., Родионова А. П. Традиционное и новое в лексике и грамматике карельского языка (по данным социолингвистического исследования). Петрозаводск: КарНЦ РАН, 2011. 138 с.
7. Кривнова О. Ф., Захаров Л. М., Строкин Г. С. Речевые корпусы (опыт разработки и использование) // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Т. 2. М., 2001. С. 230–236.
8. Крижановский А. А., Крижановская Н. Б., Новак И. П. Представление диалектов в Открытом корпусе вепсского и карельского языков (ВепКар) // Корпусная лингвистика-2019: Труды междунар. конф. СПб., 2019. С. 288–295.
9. Романенко А. Н. Робастное распознавание речи для низкоресурсных языков: Дис. … канд. техн. наук. Ульм, 2020 [Электронный ресурс]. Режим доступа: https://d-nb.info/1251880495/34 (дата обращения 20.12.2022).
10. Шаронов И. А. Междометия в языке, в тексте и в коммуникации: Дис. … д-ра филол. наук. М., 2009. 320 с.
11. Bender E. M. On achieving and evaluating language-independence in NLP // Linguistic Issues in Language Technology. 2011. Vol. 6, № 3. P. 1–26. DOI: https://doi.org/10.33011/lilt.v6i.1239
12. Berment V. Méthodes pour informatiser des langues et des groupes de langues «peu dotées»: Doct. Diss. Grenoble, 2004. Available at: https://theses.hal.science/tel-00006313/document (accessed 20.12.2022).
13. Bhatt Sh., Jain A., Dev A. Acoustic modeling in speech recognition: A systematic review // International Journal of Advanced Computer Science and Applications (IJACSA). 2020. Vol. 11, Issue 4. DOI: 10.14569/IJACSA.2020.0110455. Available at: https://thesai.org/Publications/ViewPaper?Volume=11&Issue=4&Code=IJACSA&SerialNo=55 (accessed 20.12.2022).
14. Hartmann W., Ng T., Hsiao R., Tsakalidis S., Schwartz R. Two-stage data augmentation for low-resourced speech recognition // Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech-2016). San-Francisco, 2016. P. 2378–2382.
15. Itkonen T. Aunuksen äänneopin erikoispiirteet ja aunukselaismurteiden synty // Virittäjä. 1971. № 2. P. 153–182. Available at: https://journal.fi/virittaja/article/view/35912 (accessed 20.12.2022).
16. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP world // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 6282–6293. DOI: 10.48550/arXiv.2004.09095
17. Karjalainen H., Ulriikka P., Riho G., Svetlana K. Karelian in Russia: ELDIA case-specific report, with contributions by Reetta Toivanen, Anneli Sarhimaa and Eva Kühhirt (Studies in European Language Diversity 26). Research consortium ELDIA, 2013.
18. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition // Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, 2015. P. 3586–3589.
19. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap // Proceedings of International workshop on speech and computer (SPECOM-2003). Moscow, 2003. P. 8–15.
20. Novak I., Penttonen M., Ruuskanen A., Siilin L. Karelian in grammars: A study of phonetic and morphological variation. Scientific electronic edition. Petrozavodsk: KarRC RAS, 2022. Available at: http://resources.krc.karelia.ru/illh/doc/knigi_stati/karelian_in_grammar.pdf (accessed 20.12.2022).
21. Ponti E. M., O'Horan H., Berzak Y., Vulic I., Reichart R., Poibeau T., Shutova E., Korhonen A. Modeling language variation and universals: A survey on typological linguistics for natural language processing // Computational Linguistics. 2019. Vol. 45, № 3. P. 559–601. DOI: https://doi.org/10.48550/arXiv.1807.00914
22. Rebai I., BenAyed Y., Mahdi W., Lorré J. P. Improving speech recognition using data augmentation and acoustic model fusion // Procedia Computer Science. 2017. Vol. 112. P. 316–322. DOI: https://doi.org/10.1016/j.procs.2017.08.003
23. Sarhimaa A. Karelian // Bakró-Nagy, Marianne, Johanna Laakso, and Elena Skribnik (Eds). The Oxford guide to the Uralic languages. Oxford: Oxford Academic, 2022. P. 269–290.

Поступила в редакцию 23.12.2022; принята к публикации 28.04.2023

Original article

Irina S. Kipyatkova, Cand. Sc. (Engineering), Associate Professor, St. Petersburg Federal Research Center of the Russian Academy of Sciences (St. Petersburg, Russian Federation)
ORCID 0000-0002-1264-4458; kipyatkova@iias.spb.su
Alexandra P. Rodionova, Cand. Sc. (Philology), Research Fellow, St. Petersburg Federal Research Center of the Russian Academy of Sciences (St. Petersburg, Russian Federation)
ORCID 0000-0001-5645-9441; santrar@krc.karelia.ru
Ildar A. Kagirov, Research Fellow, St. Petersburg Federal Research Center of the Russian Academy of Sciences (St. Petersburg, Russian Federation)
ORCID 0000-0003-1196-1117; kagirov@iias.spb.su
Andrey A. Krizhanovsky, Cand. Sc. (Engineering), Research Fellow, St. Petersburg Federal Research Center of the Russian Academy of Sciences (St. Petersburg, Russian Federation)
ORCID 0000-0003-3717-2079; andrew.krizhanovsky@gmail.com

SPEECH AND TEXT DATA PREPARATION FOR DEVELOPING OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM FOR THE KARELIAN LANGUAGE

A b s t r a c t . This paper addresses some aspects of collecting and preparing language data of the Livvi dialect of the Karelian language needed for training a system of automatic speech-to-text conversion. The importance of such technologies for the Karelian language derives from its status as a low-resource language, which is a serious obstacle to its study and preservation. The main tasks at the current stage of the research are to collect and annotate speech and text corpora, as well as to create a transcription dictionary. The speech corpus includes audio recordings of 15 speakers (6 men and 9 women). All the recordings were transcribed and segmented into single utterances. The volume of records after the removal of "junk" fragments was 3,5 hours. The volume of the text corpus after the removal of repeated sentences was over 5M word usages. Based on the collected text corpus, a dictionary was created, which will subsequently be used as a part of the Karelian speech recognition system. All the words included in the dictionary were automatically transcribed (phonemic transcription). In the further research collected text and speech data will be used for training and testing the Livvi-Karelian speech recognition system.

K e y w o r d s : Karelian language, Livvi-Karelian dialect, natural language automatic processing, speech recognition systems training, datasets, corpus linguistics

A c k n o w l e d g e m e n t s . This research was supported by the Russian Science Foundation (project No 22-21-00843 "Automatic speech recognitions tools for Russia's low-resource languages: the case of the Karelian language").

F o r c i t a t i o n : Kipyatkova, I. S., Rodionova, A. P., Kagirov, I. A., Krizhanovsky, A. A. Speech and text data preparation for developing an automatic speech recognition system for the Karelian language. *Proceedings of Petrozavodsk State University*. 2023;45(5):89–98. DOI: 10.15393/uchz.art.2023.924

REFERENCES

1. A f a n a s y e v a , A . A . , M u l l o n e n , I . I . A Karelian-Veps dialogue on the map of Southern Karelia. *Acta Linguistica Petropolitana*. 2020;16(3):9–28. DOI: 10.30842/alp2306573716301 (In Russ.)
2. B u b r i k h , D . V . Origins of the Karelian people: A Tale of the friend and ally of the Russian people in the north. Petrozavodsk, 1947. 53 p. (In Russ.)
3. Z a i k o v , P . M . Karelian verbs. Petrozavodsk, 2000. 294 p. (In Russ.)
4. K i p y a t k o v a , I . S . A software complex for conversational Russian speech processing and recognition. *Information and Control Systems*. 2011;53(4):53–59. (In Russ.)
5. K i p y a t k o v a , I . S . , K a g i r o v , I . A . Analytical review of methods for solving data scarcity issues regarding elaboration of automatic speech recognition systems for low-resource languages. *Informatics and Automation*. 2022;21(4):678–709. DOI: 10.15622/ia.21.4.2 (In Russ.)
6. K o v a l e v a , S . V . , R o d i o n o v a , A . P . The traditional and the innovative in the vocabulary and grammar of the Karelian language (based on a sociolinguistic research). Petrozavodsk, 2011. 138 p. (In Russ.)
7. K r i v n o v a , O . F . , Z a k h a r o v , L . M . , S t r o k i n , G . S . Speech corpora (experience in developing and application). *Proceedings of the international conference on computational linguistics and intellectual technologies "Dialog-2001"*. Vol. 2. Moscow, 2001. P. 230–236. (In Russ.)
8. K r i z h a n o v s k y , A . A . , K r i z h a n o v s k a y a , N . B . , N o v a k , I . P . Dialects in the Open Corpus of Veps and Karelian Languages (VepKar). *Proceedings of the international conference "Corpus Linguistics-2019"*. St. Petersburg, 2019. P. 288–295. (In Russ.)
9. R o m a n e n k o , A . N . Robust speech recognition for low-resource languages: Diss. Cand. Sc. (Engineering). Ulm, 2020. Available at: https://d-nb.info/1251880495/34 (accessed 20.12.2022). (In Russ.)

10. Sharonov, I. A. Interjections in language, text, and communication: Diss. Dr. Sc. (Philology). Moscow, 2009. 320 p. (In Russ.)
11. Bender, E. M. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*. 2011;6(3):1–26. DOI: https://doi.org/10.33011/lilt.v6i.1239
12. Berment, V. Méthodes pour informatiser des langues et des groupes de langues "peu dotes": Doctoral thesis. Grenoble, 2004. Available at: https://theses.hal.science/tel-00006313/document (accessed 20.12.2022).
13. Bhatt, Sh., Jain, A., Dev, A. Acoustic modeling in speech recognition: A systematic review. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 2020;11(4). DOI: 10.14569/IJACSA.2020.0110455. Available at: https://thesai.org/Publications/ViewPaper?Volume=11&Issue=4&Code=IJACSA&SerialNo=55 (accessed 20.12.2022).
14. Hartmann, W., Ng, T., Hsiao, R., Tsakalidis, S., Schwartz, R. Two-stage data augmentation for low-resourced speech recognition. *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech-2016).* San-Francisco, 2016. P. 2378–2382.
15. Itkonen, T. Aunuksen äänneopin erikoispiirteet ja aunukselaismurteiden synty. *Virittäjä*. 1971;2:53–182. Available at: https://journal.fi/virittaja/article/view/35912 (accessed 20.12.2022).
16. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. P. 6282–6293. DOI: https://doi.org/10.48550/arXiv.2004.09095
17. Karjalainen, H., Ulriikka, P., Riho, G., Svetlana, K. Karelian in Russia: ELDIA case-specific report, with contributions by Reetta Toivanen, Anneli Sarhimaa and Eva Kühhirt (Studies in European Language Diversity 26). Research consortium ELDIA, 2013.
18. Ko, T., Peddinti, V., Povey, D., Khudanpur, S. Audio augmentation for speech recognition. *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. Dresden, 2015. P. 3586–3589.
19. Krauwer, S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proceedings of International workshop on speech and computer (SPECOM-2003)*. Moscow, 2003. P. 8–15.
20. Novak, I., Penttonen, M., Ruuskanen, A., Siilin, L. Karelian in grammars: A study of phonetic and morphological variation. Scientific electronic edition. Petrozavodsk, 2022. Available at: http://resources.krc.karelia.ru/illh/doc/knigi_stati/karelian_in_grammar.pdf (accessed 20.12.2022).
21. Ponti, E. M., O'Horan, H., Berzak, Y., Vulic, I., Reichart, R., Poibeau, T., Shutova, E., Korhonen, A. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics.* 2019;45(3):559–601. DOI: https://doi.org/10.48550/arXiv.1807.00914
22. Rebai, I., BenAyed, Y., Mahdi, W., Lorré, J. P. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*. 2017;112:316–322. DOI: https://doi.org/10.1016/j.procs.2017.08.003
23. Sarhimaa, A. Karelian. *The Oxford Guide to the Uralic Languages*. (M. Bakró-Nagy, J. Laakso, E. Skribnik, Eds.). Oxford, 2022. P. 269–290.

Received: 23 December 2022; accepted: 28 April 2023