



DOI 10.22363/2618-897X-2021-18-3-245-254

Научная статья

Разработка программы ASR для распознавания вариантов русского языка

И.И. Валуйцева, И.Е. Филатов

Московский государственный областной университет
Российская Федерация, 141014, Московская обл., г. Мытищи, ул. Веры Волошиной, д. 24

Аннотация. В статье впервые затрагивается проблема распознавания вариантов русского языка на уровне речи. С развитием и ростом популярности технологий автоматического распознавания речи (ASR) в настоящее время все больше внимания уделяется проблемам, связанным с непригодностью современных приложений к работе с нестандартными разновидностями языка. Этот вопрос особенно актуален для русского, который, вопреки устоявшемуся положению о его однородности, в действительности представлен множеством отличных от стандарта форм, имеющих в большинстве случаев широкое распространение в различных регионах России и мира. Изучение различных аспектов взаимодействия алгоритмов ASR с нестандартными разновидностями русского языка, а также существующих подходов к созданию ASR-продукта, умеющего обрабатывать такие идиомы, сегодня представляется актуальным направлением лингвистических исследований. Цель данной работы заключается в подробном анализе методов разработки ASR-систем, способной выполнять задачи по распознаванию и обработке образцов речи носителей, отличных от нормы вариантов русского, что может способствовать дальнейшему исследованию данной проблемы. В качестве материала исследования используется программный интерфейс приложения для автоматического распознавания речи SOVA ASR, а также выборка аудиозаписей речи носителей центральноазиатского и украинского вариантов русского языка, соответствующие им тексты транскрипций. Применяются такие методы исследования, как анализ и систематизация специальной литературы, сбор данных, качественный и количественный анализ, эксперимент и анализ полученных результатов.

Ключевые слова: автоматическое распознавание речи, обработка естественного языка, варианты русского языка, региолект, классификация вариантов русского языка для ASR

История статьи:

Дата поступления в редакцию: 14.05.2021

Дата принятия к печати: 10.07.2021

Модератор: О.А. Валикова

© Валуйцева И.И., Филатов И.Е., 2021



This work is licensed under a Creative Commons Attribution 4.0 International License
<https://creativecommons.org/licenses/by/4.0/>

Конфликт интересов: отсутствует

Для цитирования:

Валуйцева И.И., Филатов И.Е. Разработка программы ASR для распознавания вариантов русского языка // *Полилингвильность и транскультурные практики*. 2021. Т. 18. № 3. С. 245—254. DOI 10.22363/2618-897X-2021-18-3-245-254

Research Article

Design of ASR Software for Recognition of the Russian Language Variants

I.I. Valuitseva, I.Y. Filatov

Moscow Region State University
24, Very Voloshinoy St., Mytishi, 141014, Russian Federation

Abstract. The article first touches on the problem of speech recognition of Russian language variants. With the development and growing (ASR) popularity of the automatic speech recognition (ASR) technology, more and more attention is now being paid to the problems related to the incompatibility of modern applications to work with non-standard language varieties. This question is especially relevant for Russian, as it is, contrary to the conservative statement about its homogeneity, represented by many forms that differ from the standard one, which generally have a wide distribution in various regions of Russia and throughout the world. The study of various aspects of the interaction of ACER algorithms with non-standard varieties of the Russian language, as well as existing approaches to creating an ASR product that can process such idioms, today seems to be an urgent direction. The aim of the work is to analyze in detail the methods for developing ASR systems capable of performing the task of recognizing and processing speech samples of speakers of Russian language forms different from the standard, which may contribute to further research on this topic. The research material is based on the software interface of the SOVA ASR application for automatic speech recognition, as well as a selection of audio recordings of speech of native speakers of the Central Asian and Ukrainian versions of Russian, and the corresponding transcription texts. The research methods such as the study and analysis of specialized literature, data collection for subsequent software processing, qualitative and quantitative analysis, and experimental data are used.

Key words: automatic speech recognition, natural language processing, Russian language variants, regiolect, classification of Russian language variants for ASR

Article history:

Received: 14.15.2021

Accepted: 10.07.2021

Moderator: O.A. Valikova

Conflict of interests: none

For citation:

Valuitseva I.I., and I.Y. Filatov. 2021. “Design of ASR Software for Recognition of the Russian Language Variants”. *Polylinguality and Transcultural Practices*, 18 (3), 245–254. DOI 10.22363/2618-897X-2021-18-3-245-254

Введение

Русский язык, являясь одним из самых распространенных языков мира, сегодня используется миллионами людей в повседневном общении. Традиционно принято считать, что русский — однородный язык, однако даже у него существует определенное количество разновидностей: язык носителей различается по фонетике, лексике и по другим аспектам. Подобные различия зачастую становятся теми самыми факторами, которые могут создавать сложности для русскоговорящих пользователей при работе с различными ИТ-средствами. Так, например, быстро развивающаяся и популярная в настоящее время технология автоматического распознавания речи (ASR) является одной из тех областей индустрии, где можно наблюдать проблемы, возникающие при обработке локальных вариантов языка: местных говоров, койне, региональных диалектов, акцентов. Актуальность данной проблемы растет вместе с количеством пользователей различных сервисов ASR по всему миру.

Разрешение же данного вопроса, однако, представляется довольно трудным. Языковая среда носителей русского языка действительно разнообразна, но как в академической сфере, так и в коммерческой отсутствует достаточное количество информации о характеристиках тех или иных форм языка, распространенных в определенных сообществах.

Тем не менее сегодня в руках исследователей имеются данные и инструменты, способные обеспечить подходящие условия для реализации такой системы, которая была бы способна распознавать речь носителей вариантов русского языка, отличающуюся от литературного стандарта.

Обсуждение

Варианты русского языка и принципы их выделения

Для русского, как и для любого живого языка, характерно динамическое развитие, определяемое социальной и политической ситуацией.

Сегодня вариативность и подвижность нормы современного русского языка нередко становятся объектами исследований лингвистов. Этот вопрос изучается специалистами всесторонне, однако зачастую акцент делается на таких положениях, как формирование вариантов национального языка, выделение отдельных разновидностей языка, их характеристики и отношения друг с другом, определение концепций койне и региолекта в парадигме идиомов русского языка. В конечном счете открытым остается вопрос о статусе русского языка как полинационального, что создает еще больше дискуссий не только в языкознании, но и в политике и социологии [1. С. 16–26].

Сегодня классическое деление русского языка на северорусское, южнорусское наречия и среднерусские говоры представляется неактуальным, если рассматривать его как классификацию, которую целесообразно применять в прикладных задачах: более или менее очевидные черты диалектов можно найти лишь в устной речи жителей сельской местности на некоторой части Европейской территории РФ. Фонетические и лексические различия диалектов зачастую незначительны для алгоритмов, используемых в сфере обработки естественного языка — по крайней мере, нельзя сказать, что у систем NLP возникают проблемы при работе с языком носителя того или иного наречия русского языка. Неразумно отрицать и тот факт, что не все формы русского языка одинаково хорошо обрабатываются компьютером.

Если говорить о прикладных типологиях вариантов русского языка для систем автоматического распознавания речи, то нужно отметить, что классификация должна быть более обширной и разнообразной, а характерные черты ее элементов должны определяться не только на лексическом уровне. Речь идет в первую очередь о локальных вариантах русского языка, разговорная речь носителей которых значительно отличается от стандартных наречий.

Из уже имеющихся в отечественной лингвистике понятий под подобное описание больше всего подходят региолекты. Несмотря на то, что данный термин является довольно спорным, ему трудно дать четкое определение, обычно языковые явления, описываемые с использованием этого понятия, представляют собой отдельные разновидности языка, характерные свойства которых очевидны для других носителей. Региолекты по своей сути являются территориальными койне, а использование отдельного термина для обозначения данного типа языка может быть мотивировано тем фактом, что ситуация с вариантами русского языка не похожа на таковую в других мировых языках [2. С. 55].

По ряду характеристик под описание региолектов русского языка подходят территориальные варианты русского языка в государствах постсоветского пространства. Эти идиомы не сформировали национальных языков в республиках бывшего СССР, но они функционируют в этих странах так же, как и региональные койне русского языка на территории РФ. К числу таких вариантов относят, к примеру, русский язык Украины, Казахстана и других стран ближнего зарубежья [3—5].

Стоит отметить, что с развитием региональной лингвистики появилось достаточно большое количество научных работ, посвященных региональному русскому языку, однако зачастую лингвистами описываются почти идентичные либо очень схожие формы языка, характерные для носителей, проживающих на разных территориях, — речь идет как о республиках-соседях и близких регионах в составе РФ, так и о независимых государствах с высокой долей русскоговорящего населения. Действительно, конкретизация идиомов обладает немалой академической ценностью, но порой локальность вариантов означает их невысокую распространенность, и знания о них не представляются полезными в практических задачах [6. С. 40—214].

В области ASR такая ситуация обусловлена двумя положениями. Во-первых, работа с местными вариантами русского языка, которыми пользуются довольно ограниченные группы людей, не несет никакой выгоды предприятиям, заинтере-

ресованным в создании более доступного продукта. Учитывая количество таких малораспространенных форм, можно утверждать, что их анализ и обработка — процесс, который занимает довольно много времени и труда, что вряд ли представляет интерес для компаний, так как подобные затраты не способны себя окупать — целевая аудитория пользователей не настолько широка. Во-вторых, говоря о разработке систем автоматического распознавания речи, мы прежде всего подразумеваем подходы, основанные на алгоритмах машинного обучения. Такой способ разработки требует привлечения огромного количества специально созданных данных, высокой продуктивности, мощности машин, обрабатывающих эти данные, а также довольно сложных алгоритмов.

Именно поэтому для работы с системами ASR разумнее применять данные по менее конкретным и более распространенным вариантам. Предполагается, что разделение форм языка в приложениях для автоматической обработки речи будет только вредить подобным программам; характерными признаками идиомов можно пренебречь, рассчитывая на то, что недочеты в обучении будут решены оптимизационными алгоритмами. Процесс обучения ASR представляется более простой задачей, если речь идет об использовании материала разнообразных и широко распространенных вариантов. Так, например, для разработки ASR, способной анализировать несколько популярных разновидностей русского языка, можно объединить формы языка, на которых говорят, например, в республиках Северного Кавказа, на Урале, в странах Центральной Азии.

Классифицировать варианты русского языка по признаку его положения в социуме разных регионов достаточно сложно по причине того, что даже в пределах одной страны условия функционирования языка могут сильно отличаться. Так, условно можно выделить четыре разновидности русского языка: русский язык как первый язык носителя в русскоязычной среде (к примеру, большая часть Центральной России); русский как первый язык в иноязычном окружении (агломерации республик России и стран бывшего Советского Союза: Киев, Нур-Султан, Ташкент, Таллин, Махачкала, Владикавказ и др.); русский как неродной язык (постсоветские республики и некоторые регионы России); «островные» варианты русского языка [7. С. 5—7].

Подобная классификация полезна, потому что она показывает современный статус языка. Для прикладных задач, однако, порой такой критерий использовать неудобно. В частности, если рассматривать приложения ASR как массовый и доступный продукт, то это разделение окажется излишним — например, среди пользователей из Таджикистана будут как те, для которых русский является первым, так и те, для кого русский — неродной; а в крупных агломерациях в целевую аудиторию продукта будет входить большое число мигрантов. При создании системы ASR принципиально важно лишь иметь данные о языковой ситуации в определенном регионе, роли русского языка в жизни общества и его взаимодействие с местными идиомами.

Особенности вариантов русского языка и их обработка моделями ASR

Распознавание речи производится через разделение длительных звуковых сигналов на гораздо более короткие, что формирует некоторые малые звуковые на-

боры, в которых с помощью механизмов акустического моделирования воспроизводится определенная последовательность векторов признаков. Задачей языкового моделирования, в свою очередь, является вычисление вероятности конкретных последовательностей элементов — в зависимости от языка это могут буквы, слоги, слова и фразы.

Фонетика вариантов и диалектов языка взаимодействует по большей части с акустическим моделированием; другие особенности разновидностей русского языка, представленные вне его фонетической системы, оказывают влияние на иные алгоритмы приложения ASR.

Например, для украинского варианта русского языка характерны такие черты, как специфическая интонация, высокая частотность ненормативных ударений, нестандартная реализация некоторых фонем. Данные феномены возникают в результате русско-украинской интерференции, а их присутствие в речи носителя данной формы языка отражается на работе алгоритмов акустических моделей.

В Центральной Азии интерференция местных идиомов относительно ограничена, так как русский язык в этом регионе, в отличие от Украины, окружен неродственными языками; можно говорить о том, что формы русского языка в регионе богаты регионализмами и заимствованными из местных языков словами, нехарактерными для других разновидностей русского. При этом такая лексика используется не только полилингвами, но и монолингвами, говорящими исключительно на русском.

Важно уточнить, что данные характеристики вариантов языка относятся к таким свойствам идиомов, реализация которых настолько значительна и заметна, что они влияют на то, как речь обрабатывается современной вычислительной электроникой. С уверенностью можно сказать, что специфика языка жителей Центральной Азии, Украины и других регионов регистрируется сервисами ASR, а их речь, отличающаяся от стандартного русского языка практически во всех аспектах, по-особому интерпретируется алгоритмами NLP. При детальном анализе различных нюансов процесса распознавания речи носителей форм русского языка обнаруживается, что главными причинами некорректной работы ASR в подобных условиях являются такие факторы, как недостаточная оптимизация и ориентированность данных для обучения на приближенную к стандарту разновидность языка.

Способы реализации ASR для распознавания вариантов русского языка

На сегодняшний день в индустрии существует достаточное количество инструментов и методов реализации ASR-приложений, способных распознавать речь носителей нескольких вариантов одного языка. Построение такой системы возможно различными способами.

Одними из наиболее оптимальных и удобных для использования программ ASR являются такие приложения, которые ограничиваются обслуживанием единственного языкового пакета в системе. Такой продукт легче интегрируется в иное программное обеспечение и при этом дает гарантии того, что ASR всегда исполь-

зуют правильный набор алгоритмов для обработки конкретного варианта языка, исключая случаи, когда определенный диалект некорректно распознается моделями, предназначенными для другого идиома.

Сегодня подобные ASR-продукты реализуются с помощью *sequence-to-sequence* моделей, для обучения которых требуются такие переменные, входные данные, включающие в себя параметры всех идиомов, на распознавание которых запрограммирована система. Иной способ создания такого рода приложения предполагает использование различных моделей для нескольких вариантов языка — подходящие для конкретного случая использования программы настройки определяются с помощью специального механизма, который способен узнавать каждый конкретный идиом и подбирать для него требуемые модели [8].

Несмотря на очевидные достоинства, подобные системы обладают очень запутанной структурой и представляются сложными в разработке — они очень требовательны к ресурсам и нуждаются в эффективной оптимизационной составляющей [9].

Более простой и популярный подход к созданию ASR-продукта, способного работать с несколькими вариантами одного языка, — применение индивидуальных языковых и акустических моделей для каждого из идиомов с помощью так называемых пакетов. Такие приложения предлагают пользователям самим выбирать наиболее подходящие для них механизмы распознавания. Так, для распознавания речи на английском языке доступны пакеты для диалектов Великобритании, США, Индии, ЮАР и др.; арабский язык в ASR обычно делится на язык Персидского залива, магрибский, египетский, левантийский варианты. Разделение мотивировано как тем, насколько формы языка различаются между собой, так и тем, насколько востребованным представляется продукт на конкретном рынке.

Подобные приложения, однако, гораздо менее удобны для пользователей. К тому же с многопакетными системами связана актуальная для определенных рынков проблема, когда ASR поставляются с предустановленными параметрами, которые не соответствуют языковой ситуации в регионе дистрибуции программы, что не только является причиной некорректной работы системы с конкретными формами языка, но и оказывает лишнюю нагрузку на облачные серверы. Кроме того, многопакетные модели хуже подстраиваются под малораспространенные, локальные формы, речь носителей которых не включается в обучающий набор данных [10].

Сам процесс разработки подобного рода структур представляется довольно гибким — образование системы подразумевает индивидуальную обработку моделей, что позволяет проще и точнее настроить выбранные алгоритмы на распознавание конкретных языковых форм. Данная особенность многопакетных ASR может оказываться полезной для таких вариантов, которые являются разнородными сами по себе, например, для русского языка Центральной Азии.

Важно понимать, что выбор стратегии построения ASR для распознавания нескольких вариантов языка должен отталкиваться от положения языка и его форм, диалектов, языковой ситуации в регионе. Обучение ASR распознаванию

нескольких форм языка — процесс, требующий большого количества ресурсов, а значит, создание подобных программ должно быть экономически оправданно. Разработка сложных алгоритмов приложения не интересна специалистам ни в академической сфере, ни в сфере самой индустрии ASR, если конечный продукт не востребован среди пользователей.

Эксперимент

Из-за технических и ресурсных ограничений опыт проводился с использованием уже готовых приложений с открытым кодом; система испытывалась на относительно небольшом объеме данных, чтобы создание выборки и ее машинная обработка не занимали много времени, но ее характеристики удовлетворяли бы требованиям эксперимента: в качестве инструмента был выбран программный интерфейс SOVA ASR, а для самого эксперимента был собран репрезентативный набор данных, в который входили 62 аудиозаписи с примерами речи носителей среднеазиатского и украинского вариантов русского языка и соответствующие им текстовые транскрипции.

Стратегия эксперимента предусматривает использование встроженных в программу инструментов для модификации ASR-моделей с помощью датасета, в дальнейшем используемого для тестирования приложения. Тестирование же и анализ работы системы реализуются методом валидации на отложенных данных, оценивания по алгоритму WER, а также самостоятельной ручной обработки полученных в ходе распознавания данных.

На основе проведенного эксперимента мы пришли к выводу, что обучение одних и тех же моделей распознаванию разных форм языка не принесет удовлетворительных результатов, если в структуре системы отсутствуют развитые оптимизационные алгоритмы; разумнее использовать многопакетный подход или механизм определения диалекта; наличие нормализованных и стандартизованных данных в выборках для обучения и тестирования ASR является залогом корректного обучения моделей для распознавания речи.

Заключение

Анализ языковой ситуации в постсоветских государствах показывает, что русский язык еще является распространенным языком на территории республик бывшего Советского Союза. С лингвистической стороны можно говорить о том, что построение многовариантной ASR для данных регионов будет востребовано, однако при образовании такого продукта стоит учитывать и некоторые социально-экономические факторы. Свои достоинства и недостатки имеют все подходы; предпочтительными являются те, которые предполагают использование единого языкового пакета для всех разновидностей языка, но многопакетный подход представляется гораздо более простым в реализации. Дальнейший выбор путей формирования моделей системы, написания алгоритмов приложения должен определяться качеством и доступностью ресурсов для разработки.

Список литературы

1. *Бахтикиреева У.М.* Русский — полинациональный язык? // Вестник Российского университета дружбы народов. Серия: Лингвистика. 2014. №. 2. С. 16–30.
2. *Ерина Т.Н.* Теоретические основы изучения регионального варьирования русского языка // Русский язык в условиях би- и полилингвизма: сб. научных трудов / отв. ред. З.Н. Якушкина. Чебоксары: Чуваш. гос. пед. ун-т, 2019. С. 54–57.
3. *Ерина Т.Н., Фомин Э.В.* Говорят Чебоксары: к проблеме изучения чебоксарского диалекта русского языка // Научное наследие В.А. Богородицкого и современный вектор исследований Казанской лингвистической школы. 2018. Т. 1.
4. *Жеребило Т.В.* Функционирование регионального варианта русского языка в Чеченской Республике // Рефлексия. 2016. № 5. С. 3–102.
5. *Степанов Е.Н.* Национальные варианты русского языка или русские территориальные койне? // Мова. Одесский национальный университет им. И.И. Мечникова. 2011. № 16. С. 9–14.
6. *Арефьев А.Л.* Русский язык на рубеже XX–XXI веков. М.: Центр социального развития и маркетинга. 2012.
7. *Вахтин Н.Б., Мустайоки А., Протасова Е.* Русские языки // *Slavica Helsingiensia* 40. *Istrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian*. 2010. С. 5.
8. *Li B. et al.* Multi-dialect speech recognition with a single sequence-to-sequence model // 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018. С. 4749–4753.
9. *Elfeky M. et al.* Towards acoustic model unification across dialects // 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016. С. 624–628.
10. *Diakouloukas V. et al.* Development of dialect-specific speech recognizers using adaptation methods // 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1997. Т. 2. С. 1455–1458.

References

1. Bakhtikireeva, U.M. 2014. “Russian is a Multinational Language?”. *Vestnik RUDN. Russian Journal of Linguistics* 2: 16–30. Print. (In Russ.)
2. Erina, T.N. 2019. “Teoreticheskie osnovy izucheniya regional’nogo var’irovaniya russkogo yazyka”. In *Russkii yazyk v usloviyakh bi- i polilingvizma: sb. nauchnykh trudov*. Ed. by Z.N. Yakushkina. Cheboksary: Chuvash. gos. ped. un-t. Pp. 54–57. Print. (In Russ.)
3. Erina, T.N., Fomin Eh.V. 2018. “Govoryat Cheboksary: k probleme izucheniya cheboksarskogo regiolekta russkogo yazyka”. In *Nauchnoe nasledie V.A. Bogoroditskogo i sovremenniyi vektor issledovaniy Kazanskoi lingvisticheskoi shkoly*. Vol. 1. Pp. 81–84. Print. (In Russ.)
4. Zherebilo, T.V. 2016. “Funktsionirovanie regional’nogo varianta russkogo yazyka v Chechenskoj Respublike”. *Refleksiya* 5: 3–102. Print. (In Russ.)
5. Stepanov, E.N. 2011. “Natsional’nye varianty russkogo yazyka ili russkie territorial’nye koine?”. *Mova* 16: 9–14. Print. (In Russ.)
6. Aref’ev, A.L. 2012. *Russkii yazyk na rubezhe XX–XXI vekov*. Moscow: Tsentr Sotsial’nogo razvitiya i marketinga. Print. (In Russ.)
7. Vakhtin, N.B., A. Mustaioki, and Protasova, E. 2010. “Russkie yazyki”. *Slavica Helsingiensia* 40: 5. Print. (In Russ.)
8. Li, B. et al. 2018. “Multi-dialect speech recognition with a single sequence-to-sequence model”. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) Proceedings. Pp. 4749–4753.
9. Elfeky, M. et al. 2016. “Towards acoustic model unification across dialects”. In 2016 IEEE Spoken Language Technology Workshop (SLT) Proceedings. Pp. 624–628.
10. Diakouloukas, V. et al. 1997. “Development of dialect-specific speech recognizers using adaptation methods”. In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. — IEEE Proceedings. Vol 2. Pp. 1455–1458.

Сведения об авторах:

Валуйцева Ирина Ивановна — доктор филологических наук, профессор, заведующий кафедрой теоретической и прикладной лингвистики Московского государственного областного университета. E-mail: irinaiv-v@yandex.ru

Филатов Игорь Евгеньевич — бакалавр кафедры теоретической и прикладной лингвистики Московского государственного областного университета. E-mail: imphilya_com@yahoo.com

Bio Notes:

Irina I. Valuitseva is a Doctor of Philology, Professor, Chair of the Department of the Theoretical and Applied Linguistics at Moscow Region State University. E-mail: irinaiv-v@yandex.ru

Igor Ie. Filatov is a Bachelor of the Department of Theoretical and Applied Linguistics at Moscow Region State University. E-mail: imphilya_com@yahoo.com