2015. № 2 ВОПРОСЫ ЯЗЫКОЗНАНИЯ С./Pp. 117—135
Voprosy Jazykoznanija

РЕЧЕВЫЕ ТЕХНОЛОГИИ ДЛЯ МАЛОРЕСУРСНЫХ ЯЗЫКОВ МИРА*

© 2015 г. **Алексей Анатольевич Карпов**[а,б,@], **Василиса Олеговна Верходанова**[а]

[а] Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Россия; [б] Университет ИТМО, Санкт-Петербург, 197101, Россия; [@] karpov@iias.spb.su

В последнее десятилетие активно развивающаяся область компьютерной обработки речи для малоресурсных и миноритарных языков испытывает значительный подъем. В статье представлен аналитический обзор существующих проблем, подходов и решений в области распознавания речи для многочисленных разговорных языков с недостаточными речевыми и текстовыми данными, в том числе языков Российской Федерации. Дается определение и характеристика малоресурсных языков, описываются трудности, связанные с их автоматической обработкой, также представлены ведущиеся в этой области исследования и проекты, направленные на изучение и сохранение малоресурсных языков мира.

Ключевые слова: малоресурсные языки, речевые технологии, распознавание речи, модели языка

SPEECH TECHNOLOGIES FOR UNDER-RESOURCED LANGUAGES OF THE WORLD

Alexey A. Karpov[a,b,@], **Vasilisa O. Verkhodanova**[a]

[a] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, 199178, Russia; [b] ITMO University, St. Petersburg, 197101, Russia; [@] karpov@iias.spb.su

Over the past decade, computer speech processing of under-resourced and minority languages has experienced a significant progress. In this paper, we present an analytical review of existing problems and approaches in the field of speech recognition for many spoken languages lacking speech and text resources, including languages of the Russian Federation. The definition and characteristics of under-resourced languages and challenges connected with their automatic processing are presented, as well as projects and investigations dealing with analysis and preservation of under-resourced languages of the world are described.

Keywords: under-resourced languages, speech technology, speech recognition, language models

Введение

В настоящее время персональные компьютеры и смартфоны очень активно используются для текстового и речевого общения между людьми. Разговорный язык является основным средством межчеловеческой коммуникации, а языковое разнообразие мира — это основа богатого культурного наследия человечества. К языковому разнообразию можно

* Исследование проводится при частичной финансовой поддержке фонда РФФИ (проект № 15-07-04415-а), Совета по грантам Президента РФ (проект № МД-3035.2015.8) и государственной поддержке ведущих университетов РФ (субсидия 074-U01).

относиться так же, как и к биологическому разнообразию [Crystal 2000]. Уже сейчас для ряда наиболее используемых языков доступны различные компьютерные средства обработки текста, электронные словари, машинные переводчики, системы синтеза и распознавания речи. Однако в мире насчитывается более 7 000 живых разговорных языков, и только для небольшого числа из них существуют необходимые информационные ресурсы и программное обеспечение для реализации естественно-языковых и речевых технологий. Современные информационные технологии в основном связаны с теми естественными языками, для которых доступны необходимые языковые и речевые электронные ресурсы, или же с языками, которые стали по какой-либо экономической или политической причине представлять интерес для мирового сообщества. Большая же часть языков развивающихся стран и малочисленных народов на сегодняшний день изучена недостаточно. Один из способов исправления данной ситуации заключается в создании лингвистических и речевых ресурсов, технологий и приложений для работы с такими языками. Таким образом, есть веские основания для разработки речевых технологий (систем автоматического распознавания речи, синтеза речи по тексту, машинного перевода речи) практически для всех языков мира.

В данной статье представлен аналитический обзор проблем, методов и систем автоматического распознавания речи (САРР) для малоресурсных языков (МРЯ, under-resourced languages), который демонстрирует возросший в последнее время интерес к этой области. Несмотря на то что задача автоматического распознавания речи достаточно специфична, ряд затрагиваемых в этой статье аспектов актуален и для других задач прикладной лингвистики и информационных технологий. В целом данный обзор является обобщением и развитием исследований и статей, опубликованных в 2014 г. в специальном выпуске международного научного журнала «Speech Communication» (http://www.sciencedirect.com/science/journal/01676393/56/supp/C), который был посвящен компьютерной обработке МРЯ и приглашенными редакторами которого выступали Л. Безасие (Франция), Э. Барнард (ЮАР), А. Карпов (Россия, СПИИРАН) и Т. Шульц (Германия, президент ассоциации ISCA) [Besacier et al. 2014]. Также в данном обзоре анализируются доклады, представленные на международных семинарах по речевым технологиям для малоресурсных языков (SLTU), последний из которых впервые проходил в России (Санкт-Петербург) в мае 2014 г. (http://www.mica.edu.vn/sltu2014).

1. Общая характеристика языков мира

1.1. Разнообразие разговорных языков

Актуальную оценку количества живых языков мира можно найти на языковедческом интернет-ресурсе Ethnologue (http://www.ethnologue.com), где приводится следующее определение живого языка: «Язык, на котором говорит хотя бы один человек и для которого этот язык является родным». Таким образом, мертвые языки и неродные языки не учитываются при подсчете. На основе этого определения Ethnologue по состоянию на 2014 г. выделяет более 7 100 известных науке живых языков, на которых разговаривает около 6,3 млрд. человек. Причем этот список включает в себя около тысячи языков, которые классифицируются как почти вымершие (находящиеся под угрозой исчезновения), т. е. только несколько пожилых носителей живы. Также известно, что на 96 % из числа всех известных языков разговаривает лишь 4 % человечества, а более половины жителей нашей планеты говорит на одном из пяти крупнейших мировых языков [Плунгян 2010]. Нужно еще отметить, что каталог Ethnologue включает в себя как вербальные, так и визуально-кинетические разговорные языки. Последние называют жестовыми, и они используются для повседневной коммуникации глухими и слабослышащими людьми и объединяют в себе жесты, мимику и артикуляцию губ [Карпов 2011]. В данном обзоре рассматриваются только вербальные разговорные языки, которые имеют звуковую форму.

Отдельно ведется также подсчет языков, имеющих письменную форму. Так, Фонд вымирающих языков FEL (http://www.ogmios.org/home.htm) указывает цифру примерно в 2 тыс. письменных языков по количеству опубликованных библий (полностью или частично), однако эта оценка включает и уже мертвые языки. Другой интернет-ресурс Omniglot — энциклопедия систем письма и языков (http://www.omniglot.com) — перечисляет около тысячи письменных языков и приводит описание более 180 разных систем письма.

В то время как общий учет языков мира является непростой научной задачей, количество достаточно хорошо исследованных языков с необходимыми языковыми и речевыми корпусами легко перечислить, назвав число языков, которые учитываются в современных информационных технологиях по обработке естественного языка и речи, таких как переводчик Google Translate (80 языков в 2014 г.), словарь Wiktionary и энциклопедия Wikipedia (более сотни языков), интернет-поисковик Google Search (более сотни языков), голосовой поисковик Google Voice Search (35 языков и их региолектов), голосовой помощник Siri для iPhone от Apple/Nuance (9 языков в 2014 г., причем, русский язык в них не входит).

Проблема сохранения и обработки МРЯ является насущной также и для России. Как известно, в республиках Российской Федерации используется свыше 150 различных языков [Potapova 2011], многие из которых являются государственными и официальными языками России. При этом каждая республика в дополнение к государственному русскому языку имеет возможность определять на своей территории другие государственные языки (наибольшее количество языков со статусом государственного зарегистрировано в Дагестане — 14). Помимо русского языка, который обязано знать все население страны, наиболее распространенными по числу носителей являются татарский (свыше 5 млн носителей), чеченский, башкирский и чувашский языки (не менее 1 млн носителей). Все данные языки по международной классификации считаются МРЯ; причем русский язык и его диалекты за рубежом тоже иногда относят к таковым [Lamel et al. 2012] (современный обзор САРР для русского языка представлен в работах [Кипяткова, Карпов 2010; Кипяткова и др. 2013; Vazhenina et al. 2012]). При этом лингвистические и фонетические корпусные исследования и сбор речевых баз данных и словарей проводятся в России для многих из данных языков (например, языков Кавказа [Potapova 2011]), но систематические работы по созданию компьютерных речевых технологий были начаты, пожалуй, только для татарского [Хусаинов 2014].

1.2. Процесс вымирания естественных языков

В современном мире с его растущей глобализацией языки исчезают с высокой скоростью. В начале этого века было спрогнозировано [Crystal 2000], что через столетие половина из ныне исчезающих языков будут мертвыми. Можно сказать, что в среднем каждые две недели один язык вымирает. Как показывает эволюция, даже если на языке говорит 100 тыс. человек, он не защищен от вымирания [Crystal 2000], так как выживание определенного языка зависит от оказываемого давления на язык и его носителей. Подобное давление может возникать из-за природных катастроф (так, сильные землетрясения в Папуа — Новой Гвинее убили несколько живых языков), геноцида народов (около 90 % американских аборигенов погибли в период 200-летнего покорения Америки Европой) или просто от тотального доминирования одного языка над другим [Besacier et al. 2014]. Последнее может вылиться также в культурную ассимиляцию (из-за социальных, политических и экономических преимуществ от использования доминантного языка), которая обычно приводит к потере подавляемого языка в течение жизни следующих поколений (например, второго поколения иммигрантов).

Сегодняшнее языковое разнообразие мира довольно зыбко, поскольку многим языкам угрожает вымирание в силу многих причин [Мурадова 2008]. Возникает вопрос: как можно замедлить вымирание языков, и каковы связанные с этим затраты? Во-первых, язык можно

сохранить, только если само общество этого хочет и окружающая культурная и социальная среда уважает это стремление. Это может выражаться, например, в финансовой поддержке образовательных курсов, учебников, пособий и учителей. А также в том, что лингвисты выезжают в полевые экспедиции, собирают, публикуют и делают общедоступной информацию о языке: его грамматику, словарь, аудиовизуальные записи его носителей и т. д. Требуемые расходы, связанные с сохранением языка, зависят также от определенных условий, таких, например, как наличие у языка письменности. Так, в [Crystal 2000] расходы на сохранение языков оценены приблизительно в 80—100 тыс. долларов на один язык в год. Если же рассматривать все вымирающие ныне языки, то цифра необходимых расходов может вырасти до миллиарда долларов.

В мире для привлечения общественного внимания и финансов для решения проблемы вымирания языков были запущены масштабные проекты ЮНЕСКО (Atlas of the World's Languages in Danger и Red Book of Endangered Languages), а также созданы такие организации, как, например, Фонд исчезающих языков (Foundation of Endangered Languages, http://www.ogmios.org). Советом Европы была подготовлена и действует в ЕС «Европейская хартия региональных языков или языков меньшинств», к которой присоединилась и Россия (http://www.terralegis.org/terra/act/e474.html).

Некоторые научные организации и ассоциации также весьма активны в исследовании МРЯ и вымирающих языков. Так, например, международная ассоциация Sorosoro (http://www.sorosoro.org) занимается организацией программ «оживления» и документирования исчезающих языков. Крупный французский проект PI-languages (poorly computerized language) посвящен разработке инструментов распознавания речи для МРЯ, в особенности языков юго-восточной Азии (http://pi.imag.fr/xwiki/bin/view/Main/). Другая международная ассоциация African Language Technology (AfLaT, http://aflat.org) занимается исследованиями МРЯ восточной Африки. В рамках международной ассоциации по речевой коммуникации ISCA создана специальная группа по речевым технологиям для миноритарных языков SALTMIL (Speech and Language Technologies for Minority Languages, http://ixa2.si.ehu.es/saltmil). Проводится международная конференция по миноритарным языкам (http://icml14.uni-graz.at).

В России также иногда проходят научные мероприятия, посвященные изучению языков этнических меньшинств. Например, в 2011 г. в Йошкар-Оле проходила международная конференция «Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы» (http://www.marlamuter.ru/downloads/rezolution.pdf). Выходят также научные издания по результатам изучения миноритарных языков России и стран бывшего СССР [Казакевич, Кибрик 2005; Потапова 2009; Челышева 2009].

Однако важно подчеркнуть, что языки национальных меньшинств и малочисленных этносов (minority languages) не то же самое, что малоресурсные языки (under-resourced languages), которые могут быть национальными и официальными языками стран, причем иногда с очень большим количеством носителей. С другой стороны, некоторые языки национальных меньшинств можно рассматривать как языки с достаточными ресурсами (например, каталанский язык доступен на Google Search и Google Translate). МРЯ не обязательно являются вымирающими, хотя обратное утверждение чаще всего верно.

1.3. Речевые технологии для оживления языков

С помощью современных речевых и языковых технологий (speech technologies или spoken language technologies) можно задокументировать разговорные языки и таким образом прекратить или замедлить их вымирание. Языковое разнообразие — это основа нашего богатого культурного наследства. Если в мире пропадает язык, то воспоминания, опыт и культура уходят вместе с ним. Существование соответствующей компьютерной технологии может пробудить интерес к языку у его носителей и у людей, изучающих его. Кроме того, в свете необходимости оживления языков развитие речевых технологий для их транскрибирования

является важным шагом к их сохранению (в основном это касается разговорных бесписьменных языков), так как помогает облегчить доступ к аудиоданным на этих языках.

Другой веской причиной, почему естественно-языковые и речевые технологии должны быть доступны для всех языков, является то, что политическое и социальное влияние языка в глобальном плане может быть изменчиво. Различие языков — это также одно из препятствий, затрудняющих взаимодействие между людьми из разных стран. Во время вооруженных конфликтов или природных катастроф может потребоваться вербальное взаимодействие с носителями МРЯ. Например, разрушительное землетрясение на Гаити в 2010 г. подчеркнуло необходимость технологий для работы с креольским языком [Besacier et al. 2014]. Часто местное население, которому в подобной ситуации необходимо общаться с иностранными спасателями или врачами, говорит только на родном языке, незнакомом для приезжих. Для таких случаев переводчиков зачастую не хватает или они не могут вовремя добраться в нужное место. В подобных ситуациях такие информационные технологии, как машинный перевод, синтез и распознавание речи были бы очень полезными. Возможно, современные технологии еще далеко не совершенны, но при отсутствии альтернативы и в экстренной ситуации даже неидеальная система была бы крайне важна.

Кроме того, некоторые МРЯ и вымирающие языки при должной поддержке доступных информационных технологий могут расцвести и в дальнейшем стать сильными в социальной, политической и экономической сфере. В качестве примера можно отметить языки быстро развивающихся стран (в том числе стран БРИКС): малайский, вьетнамский, бенгальский, урду, суахили и т. д., некоторые из которых уже находятся в двадцатке наиболее общеупотребительных языков мира.

2. Анализ особенностей малоресурсных языков

2.1. Определение малоресурсных языков

Термин «малоресурсные языки» (under-resourced languages или low-resourced languages) первоначально был предложен нидерландским ученым С. Краувером [Krauwer 2003]. Это понятие относится к естественным языкам с некоторыми (или всеми) из следующих свойств: недостаток своей системы письменности или устойчивой орфографии; нехватка квалифицированных лингвистов и переводчиков для данного языка; ограниченное распространение в сети Интернет; нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфографических и фонетических транскрипций речи, словарей произношения и т. д. Как уже говорилось, МРЯ — это не всегда миноритарный язык (minority language), на котором говорит меньшая часть населения на определенной территории. Некоторые МРЯ в действительности являются государственными или официальными языками стран, и на них говорит большая часть населения (например, бенгальский, на котором говорит около 250 млн человек, тамильский, урду, малайский, казахский, белорусский и т. д.). С другой стороны, некоторые малые языки можно рассматривать как языки с достаточными ресурсами (well-resourced languages). Скажем, каталанский, баскский или галисийский языки сейчас доступны в интернет-сервисах Google Voice Search и Translate, а также в Wikipedia и Wiktionary, поэтому они уже не могут считаться МРЯ.

В течение последних лет все больше внимания научного сообщества уделяется созданию и адаптации текстовых и речевых ресурсов и моделей для автоматической обработки МРЯ. Этой теме посвящен, например, международный семинар по речевым технологиям для малоресурсных языков (International Workshop on Spoken Language Technologies for Under-resourced Languages — SLTU; http://www.mica.edu.vn/sltu), который проходил в 2008 г. в Ханое (Вьетнам), затем в 2010 г. в Пенанге (Малайзия), в 2012 г. в Кейптауне (ЮАР) и наконец в мае 2014 г. в Санкт-Петербурге [Карпов 2015].

2.2. Оценка статуса языка

Для того чтобы объективно оценить текущий статус доступности ресурсов языка, по инициативе ассоциаций ELSNET и ELRA был введен критерий BLARK (Basic LAnguage Resource Kit, http://www.blark.org) [Krauwer 2003]. В рамках предложенной концепции был определен минимальный набор языковых ресурсов, которые нужно сделать доступными для максимально большого числа языков. Критерий BLARK позволяет учитывать экспертные оценки наличия и доступности электронных ресурсов и сервисов для конкретного языка с вычислением средней оценки. В [Berment 2004] приводится пример оценки по данному показателю кхмерского языка Камбоджи (6,2 из 20 баллов), а также вьетнамского языка (10 из 20). В данной работе МРЯ определен как язык, для которого значение этой оценки ниже 10 из 20 возможных баллов.

Недавно европейский проект METANET опубликовал ряд аналитических документов «Языки в Европейском информационном обществе» (http://www.meta-net.eu/whitepapers/overview), которые описывают состояние каждого европейского языка с учетом наличия для него естественно-языковых технологий (Human Language Technology) и объясняют актуальные риски и возможности развития для языков Европы. Анализ показывает, что целый ряд государственных языков ЕС до сих пор является МРЯ: латышский, литовский, исландский, мальтийский, ирландский и т. д. (http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison).

2.3. Проблемы автоматической обработки малоресурсных языков

При создании системы автоматического распознавания речи (САРР) для некоторого МРЯ необходимо использовать методы и подходы, которые выходят за рамки простого перевода, переобучения и адаптации моделей от другого языка (например, английского). При автоматизированной обработке любого нового языка часто приходится сталкиваться с новыми трудностями, которые проистекают из специфических фонетических систем, проблем с сегментацией на слова, размытых грамматических структур фраз, наличия бесписьменных языков и т. д. Необходимы также и новые методики сбора данных, например, краудсорсинг (crowdsourcing) [Gelas et al. 2011]. Помимо этого нужны универсальные модели, в которых информация разделена между языками, например, многоязычные акустические или языковые модели [Schultz, Kirchhoff 2006; Le, Besacier 2009]. Часто социальные и культурные аспекты, связанные с рассматриваемым языком, являются источником дополнительных трудностей: языки с многочисленными диалектами / регионолектами в разных регионах, постоянное переключение носителей с одного языка / диалекта на другой (англ. «code-switching speech», например, смешанная (английский и африкаанс) речь в ЮАР), перемешивание языков / диалектов в речи говорящих (англ. «code-mixing speech», например, суржик на Украине), массовое присутствие неносителей языка (как в случае с суахили). Другой сложностью является необходимость свести вместе требования самих носителей и разработчиков автоматических систем: крайне сложно найти носителей языка с необходимыми техническими навыками для разработки САРР для их родного языка. Кроме того, МРЯ обычно недостаточно исследованы учеными и слабо описаны в лингвистической литературе.

Для того чтобы начать создавать речевые технологии для МРЯ, нужно сначала постараться заимствовать подходящие информационные ресурсы и знания из родственных языков (одних языковых семей и групп). Для этого необходимы междисциплинарные исследования, в том числе с привлечением лингвистов, диалектологов (чтобы найти коэффициенты близости между языками), фонетистов (чтобы соотнести фонетические алфавиты рассматриваемых МРЯ и языков с достаточными ресурсами), технических специалистов и т. д. Более того, для некоторых языков важно на ранней стадии работ проверить адекватность использования базовых подходов и методов распознавания речи: например, является ли слово

оптимальной единицей для языкового моделирования либо лучше использовать морфемы, и нужно ли добавлять тональные характеристики речи при акустическом моделировании аудиосигналов? Вдобавок для некоторых малых или вымирающих языков разработчикам технологий зачастую необходимо работать в контакте с полевыми лингвистами, чтобы получить доступ к носителям языка и собрать необходимые обучающие и тестовые данные в соответствии с принятыми техническими требованиями и правилами.

2.4. Речевые и текстовые информационные ресурсы

Процесс создания речевых технологий (в частности, САРР) требует наличия речевых корпусов с аудиоданными и транскрипциями речи многих дикторов-носителей языка, словарей произношений, охватывающих всю лексику как минимум обучающего речевого корпуса, а также больших объемов текстовых данных для создания статистических языковых моделей. Несмотря на то, что за последние годы возросло количество языков, для которых речевые корпуса и языковые ресурсы были систематически собраны и стали доступны, этих языков до сих пор не более сотни. При этом во многих географических регионах при сборе речевых данных полевые исследователи сталкиваются с различными политическими и культурными сложностями, вследствие чего стоимость лицензирования и продажи баз данных для ряда языков может быть непомерно высокой.

Из практики речевых исследований известно, что обучающие речевые корпуса должны предоставлять аудиоданные для разных языков с сопоставимым качеством записи (тип микрофона, уровень шума, частота дискретизации сигнала и т. д.), стилем речи (чтение текста, разговор или спонтанная речь), форматом транскрипции и словарями. Такие речевые корпуса подходят для обучения многоязычных акустических моделей и для быстрой адаптации технологий под новые языки и прикладные области.

Для работы над базами данных был создан международный консорциум лингвистических данных LDC (https://www.ldc.upenn.edu), запустивший проект по созданию текстовых и речевых корпусов для многих языков, в рамках которого открыт коммерческий доступ к обучающим корпусам для распознавания и синтеза речи, машинного перевода и т. д. Европейская ассоциация по языковым ресурсам ELRA (http://www.elra.info) также имеет в своем каталоге базы данных и информационные ресурсы для десятков языков мира.

Одним из хороших примеров является и корпус речи и текста GlobalPhone [Schultz et al. 2013]. Эта база данных содержит аудиозаписи и транскрипции речи для разработки и тестирования речевых технологий для 20 распространенных языков и МРЯ, включая русский, арабский, болгарский, мандаринский китайский, шанхайский китайский, хорватский, чешский, французский, немецкий, хауса, японский, корейский, польский, португальский (бразильский вариант), испанский (латиноамериканский вариант), шведский, тамильский, тайский, турецкий и вьетнамский. Корпус содержит свыше 400 часов записи речи в исполнении более 2 000 носителей языков. В эту базу данных входят также словари произношений и языковые модели. GlobalPhone создана единообразно для всех включенных в нее языков в плане объема текста и речи для каждого языка (100 дикторов на язык), качества записей (одинаковые микрофоны, низкие шумы), методики сбора (окружающая обстановка, стиль речи, условия записи), а также фонетических транскрипций речи и обозначений фонем.

Кроме того, в последнее время в мире для систем автоматического распознавания речи для МРЯ значительную популярность приобрели также многоязычные речевые базы данных, созданные в рамках проекта Babel американского агентства IARPA (http://www.iarpa.gov/index.php/research-programs/babel). Речевой корпус Babel включает транскрибированные записи речи на нескольких десятках МРЯ, его используют в своих исследованиях научные коллективы и коммерческие организации из ряда зарубежных стран [Gales et al. 2014; Hartmann et al. 2014]. За последние годы также были собраны и стали доступны для исследований корпуса для 11 официальных южноафриканских языков, включая базу данных

телефонной речи AST [Roux et al. 2000], многоязычный корпус Lwazi [Barnard et al. 2009] и большой корпус речи NCHLT [de Vries et al. 2014].

Среди других коммерческих организаций, занимающихся сбором речевых и текстовых корпусов, стоит упомянуть компанию AppenButlerHill, которая имеет около 80 языков в своем каталоге (http://catalog.appenbutlerhill.com/), китайскую компанию SpeechOcean, предоставляющую речевые данные для САРР на 30 языках (http://www.speechocean.com/en-Product-Catalogue/). Тем не менее, во многих регионах при сборе баз данных составители сталкиваются с политическими и культурными трудностями, вследствие чего стоимость баз данных для некоторых языков может быть очень высокой (особенно для коммерческих задач). А для некоторых языков нельзя даже говорить о составлении произносительных словарей и больших корпусов.

В России и странах бывшего СССР проблема сбора баз данных для МРЯ стоит также остро из-за значительного языкового разнообразия. Собрано уже достаточно много больших текстовых корпусов для русского языка (например, Национальный корпус русского языка, http://www.ruscorpora.ru; Корпус русского литературного языка, http://www.narusco.ru), а также русскоязычные речевые базы данных (представлены, в частности, в работах [Ронжин и др. 2006; Кибрик, Подлесская 2009; Асиновский и др. 2010; Arlazarov et al. 2004]), чего, однако, нельзя сказать про остальные языки Российской Федерации. Среди известных электронных языковых ресурсов для официальных МРЯ можно отметить, например, лингвистические корпуса, созданные при поддержке программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (http://web-corpora.net). Так, примерами таких корпусов являются корпус литературного бурятского языка объемом более 2,2 млн словоупотреблений с начальной морфологической разметкой входящих в него слов на основе словоизменительных характеристик; корпус калмыцкого языка объемом около 800 тыс. словоупотреблений с морфологической разметкой; корпус осетинского литературного языка объемом 10,5 млн словоупотреблений с автоматической разметкой на русском и английском языках; корпус литературного лезгинского языка объемом около 4,5 млн. словоупотреблений с морфологической разметкой; национальный корпус татарского языка «Туган тел», объем которого составляет более 26 млн слов с автоматической морфологической разметкой. Также известен текстовый «Вепсский корпус» (http://vepsian.krc.karelia.ru), содержащий диалектные бытовые, этнографические и фольклорные тексты, тексты на младописьменном вепсском языке. Известен также российский интернет-проект по сравнительно-историческому языкознанию «Вавилонская башня» (http://starling.rinet.ru) [Старостин 2007], предоставляющий электронные лексико-грамматические базы данных по русским народным говорам. Кроме того, стоит упомянуть о создании «Письменного корпуса татарского языка» (http://corpus.tatfolk.ru) и о проекте «Онлайн-словари финно-угорских народов» (http://dict.marlamuter.ru), включающем словари для марийского и удмуртского языков. Существуют также публикации о разработке компьютерного фонда звучащей речи «Языки народов России» [Асиновский и др. 2007].

3. Автоматическое распознавание речи для малоресурсных языков

Любая современная САРР содержит в своем составе три основные модели: акустическая (акустико-фонетичекая) модель, лексическая модель (словарь произношений) и языковая модель (либо грамматика) [Besacier et al. 2014]. Общая архитектура современной САРР, включающая в себя три данные модели, представлена на рисунке 1. Такая САРР функционирует в двух основных режимах: обучение (сбор обучающих баз данных, создание и обучение вероятностных моделей) и распознавание (декодирование речевого сигнала и отбор наилучших гипотез распознавания). Далее в этом разделе будут описаны основные проблемы и возможные решения при создании всех этих моделей для САРР с учетом специфики МРЯ.

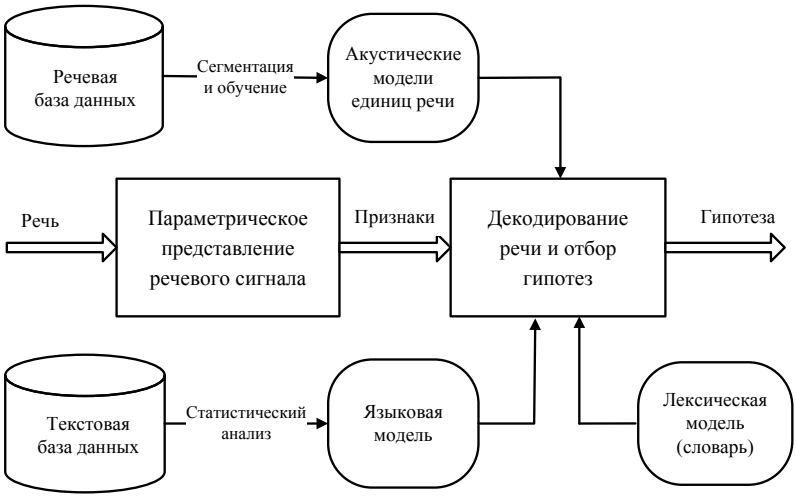

Рис. 1. Общая архитектура системы автоматического распознавания речи

Следует отметить, что в конце прошлого века на заре расцвета речевых технологий САРР, созданные первоначально для одного языка, были успешно адаптированы к ряду других языков. Среди таких разработок можно отметить САРР, созданные в IBM, BBN, Philips, CUED, MIT и LIMSI [Besacier et al. 2014]. Успешная адаптация САРР английской речи для немецкого, французского, чешского, японского языков показала адекватность методов совместного компьютерного моделирования разных языков и то, что речевые технологии для нескольких языков можно объединять. Тогда же исследователи начали систематически разрабатывать возможности создания языконезависимых акустических моделей для компьютерного анализа новых языков. Рассматривались различные аспекты этой задачи: влияние языковых семей и количества использованных языков на качество акустических моделей речи, количества обучающих данных на точность распознавания, возможность создания универсальных акустических моделей [Besacier et al. 2014].

3.1. Сбор и обработка обучающих данных

Как уже отмечалось выше, использование методов статистического анализа и вероятностных моделей в современных САРР требует значительного объема обучающих данных для создания акустических, лексических и языковых моделей. При этом для очень многих МРЯ вообще не существует корпусов, которые можно было бы использовать для разработки САРР. Поэтому сбор обучающих данных является первоочередной задачей при разработке САРР для нового языка. Подходы к сбору аудиоданных можно разделить на те, которые используют и адаптируют уже существующие речевые ресурсы, и те, в которых предполагается записывать речь дикторов. В первом случае в качестве отправной точки для создания аудиокорпуса используются доступные записи радио- и телепередач, выступления членов парламента или подобные речевые ресурсы. При этом основной трудностью является редактирование и транскрибирование записей, чтобы их можно было использовать для эффективного обучения вероятностных акустических моделей. Простое ручное транскрибирование речи осложняется нехваткой экспертов-фонетистов, владеющих МРЯ. В работах [Parent, Eskenazi 2010; Gelas et al. 2011] были достаточно успешно использованы методы краудсорсинга для транскрибирования аудиоданных с привлечением экспертов в удаленном режиме через Интернет. Другая сложность состоит в том, что часто существующие аудиоданные не отличаются необходимым для САРР разнообразием дикторов. Обычно корпус

для дикторонезависимой САРР должен содержать аудиозаписи как минимум 50 различных дикторов, но в теле- и радиопередачах или записях лекций чаще всего встречается до десяти говорящих, а то и меньше [Barnard et al. 2009].

Когда речевой корпус записывается по заранее подготовленным текстам, задача транскрибирования речи может быть существенно облегчена, поскольку есть возможность использовать подготовленные материалы для выравнивания речи и текстов. В этом случае начинают со сбора подходящего текстового корпуса, что возможно только для языков со стандартизированной письменностью. Из этого корпуса выбираются фразы, которые предоставляются для прочтения отобранным дикторам. Существуют и автоматические методы, которые показали свою эффективность для задачи проверки правильности произнесения слов дикторами [Davel et al. 2011]. При этом акустическая модель САРР обучается на начальных аудиоданных, считая, что все реплики записаны правильно, и эта система используется для итеративного обнаружения неправильно произнесенных дикторами фраз и увеличения вероятности распознавания обучающих фраз.

Для самого процесса сбора и записи аудиоданных сейчас часто используются телефонные сервисы, управляемые в режиме интерактивного меню. Инструкции с подготовленными репликами раздаются дикторам, которые звонят по бесплатному номеру телефона, где их просят воспроизвести фразы по порядку. Другой вариант сбора данных предполагает, что материал собирается во время записи разговоров или диалогов с инструктором (с использованием компьютера, диктофона или смартфона). Такой подход часто более выигрышен из-за того, что исследователь может давать указания информанту в ходе записи. Распространенность и доступность смартфонов подтолкнула ряд компаний к разработке специальных мобильных приложений, позволяющих информанту самому производить процесс записи и непосредственно контактировать с инструктором [Hughes et al. 2010; de Vries et al. 2014]. В этом случае полевой исследователь может использовать несколько смартфонов одновременно, что позволяет параллельно записывать речевые данные нескольких человек.

Однако нужно отметить, что в случае создания САРР для малоресурсных языков «с нуля» из-за нехватки ресурсов сначала обычно создаются относительно небольшие корпуса, где предпочтение отдается подготовленной читаемой речи, а не спонтанной, поскольку более четкое произношение слов в подобных корпусах важно в условиях нехватки ресурсов.

3.2. Акустическое моделирование речи

Современные САРР для языков с достаточными ресурсами, как правило, используют скрытые марковские модели (СММ) для моделирования фонем (фонов) языка. К МРЯ часто применяют этот же подход акустического моделирования на основе СММ. При этом непростой задачей является определение оптимального фонемного алфавита языка для акустико-фонетического моделирования. Даже когда набор фонем определен в МРЯ, он зачастую лишен эмпирического основания [Wissing, Barnard 2008]. Фонетический алфавит можно позаимствовать в родственных языках, но такой способ обязательно требует экспериментальной проверки. Существуют методы для объединения и сопоставления речевой и фонетической информации из родственных языков для создания акустических моделей [Le, Besacier 2009; van Heerden et al. 2010; Chan, Rosenfeld 2012].

На практике бывает весьма сложно получить транскрипции речи на МРЯ из-за того, что они еще недостаточно изучены. Для таких случаев разрабатываются контролируемые или полуавтоматические методы, использующие данные из родственных языков, например метод «Multilingual A-stabil» [Vu et al. 2010]. Подобные методы эффективны, когда доступна хоть какая-то информация о целевом языке (например, фонетический словарь или языковая модель), поскольку они сокращают затраты на создание САРР для нового языка.

В ряде работ для некоторых МРЯ применяются также СММ, описывающие акустические параметры не контекстно зависимых фонем (аллофонов), а больших единиц языка — слогов или дифтонгов [Gemmeke, van Hamme 2011; Tachbelie et al. 2014]. В этом случае из-за недостатка

обучающих данных можно исходить из того, что контекстные зависимости обычно менее значимы для слоговых моделей, чем для фонемных. В [Stuker et al. 2003; Siniscalchi et al. 2013] предложено моделировать любой разговорный язык универсальным набором основных единиц речи, который можно определить для всех языков. Набор моделей единиц речи для конкретного языка выбирается на основе характеристик фонем (таких как способ и место артикуляции) на основе классификации международного фонетического алфавита (МФА).

3.3. Лексическое моделирование

Одним из наиболее распространенных подходов к лексическому моделированию в САРР является использование словарей произношений (pronunciation vocabulary), где каждому слову ставится в соответствие его фонетическое (фонематическое) представление, т. е. то, как это слово произносится [Besacier et al. 2014]. Для преобразования «буква—фонема» существуют подходы, основанные на знаниях (правилах), либо подходы, основанные на данных. При этом далеко не для всех МРЯ существуют достаточно представительные словари произношений и не всегда возможно автоматически получить фонетическое представление из орфографических слов. В этом случае можно в словаре использовать графемное представление слов вместо фонетического [Le, Besacier 2009]. В случае акустико-графемного моделирования каждое слово в словаре произношений представляется в виде графем (букв), которые и являются базовыми единицами для акустического моделирования. Подобные словари дают достаточно хороший результат распознавания для тех языков, где графемы и фонемы в языке тесно связаны и имеют однозначное отображение.

Существуют также методы создания словарей произношений, использующие пары слово—транскрипция, найденные в Интернете [Ghoshal et al. 2009; Schlippe et al. 2014]. Например, открытый интернет-словарь Wiktionary содержит фонетическое представление слов в формате МФА. В работах [Schlippe et al. 2010; 2012] предложен автоматический метод извлечения фонетических транскрипций из словаря Wiktionary и словарных статей Wikipedia, который определяет, удаляет и заменяет противоречивые или некорректные пары слово-транскрипция, взятые в Интернете.

Кроме того, предложены также методы преобразования «буква—фонема», которые используют средства статистического машинного перевода [Laurent et al. 2009; Karanasou, Lamel 2010]. При этом графемы рассматриваются как «слова» во входном языке, а фонемы — как «слова» в выходном языке. Система машинного перевода обучается на основе исходного фонетического словаря для ограниченного количества слов, а после ее применяют для преобразования произвольного слова на этом языке в его фонетическую форму [Cucu et al. 2014].

Отдельную проблему представляет обработка языков, не имеющих письменности, для которых нет САРР и машинного перевода. Если язык бесписьменный, то работать можно только с аудиосигналами и фонетическими транскрипциями [Kempton, Moore 2014]. Эти транскрипции могут составляться экспертами-фонетистами или САРР для других родственных языков. В работах [Besacier et al. 2006; Stuker et al. 2009] предложены способы автоматического формирования лексических единиц (и их произношений) для неизвестного языка, что осуществляется с помощью автоматического объединения цепочек фонем для образования новых слов, например, на базе метода Model3P [Stahlberg et al. 2012].

3.4. Языковое моделирование

Модели языка и грамматики фраз позволяют оценить вероятность появления некоторой цепочки слов, что дает возможность отфильтровать маловероятные последовательности слов в гипотезах распознавания. Один из наиболее эффективных подходов к статистическому моделированию основан на использовании n-грамм (биграмм, триграмм и т. д.), с помощью которых вычисляется вероятность нахождения любой последовательности слов в тексте. Распределение вероятностей n-грамм в модели зависит от наличия обучающих

текстовых данных, поэтому для статистического моделирования языка необходимы большие объемы данных, чтобы обеспечить статистическую значимость.

При этом для ряда синтетических языков с богатой морфологией разумно разделение слов на сублексические единицы (морфемы / морфы или другие) и использование их в качестве элементов словаря и языковой модели [Ronzhin, Karpov 2007]. Такой подход позволяет сократить размер словаря распознавания и обеспечить больший охват лексики за счет уменьшения количества внесловарных слов (out-of-vocabulary words). Тем не менее, в случае применения данного метода возникают иные проблемы, связанные с высокой фонетической неоднозначностью на уровне частей слов, с преобразованием графем в фонемы со множественными транскрипциями, с необходимостью после распознавания составлять целые слова из распознанных частей (с учетом ошибок распознавания), а также использовать n-граммы более высокого порядка (от 5- до 10-грамм) для охвата грамматических зависимостей между соседними словами во фразах [Besacier et al. 2014]. Модели языка, основанные на морфемах, успешно применялись для ряда языков с богатой морфологией, в частности, для агглютинативных и флективных языков (славянских), в том числе русского [Whittaker, Woodland 2003; Ronzhin, Karpov 2007] и словенского [Rotovnik et al. 2007].

Разбиение словоформ на морфемы возможно выполнить с помощью двух разных подходов: методов, основанных на знаниях (с использованием грамматических правил), и методов, основанных на данных (статистическом анализе текстовых данных) [Karpov et al. 2011]. Преимуществом первых является то, что они позволяют получить истинное разделение словоформ на грамматические морфемы (приставки, корни, суффиксы, окончания и т. д.). Особенностью вторых является то, что они опираются только на статистический анализ текста и не используют дополнительных лингвистических знаний, что позволяет обрабатывать тексты любого письменного языка. Однако с помощью статистических методов, как правило, слова можно разделить только на псевдоморфемы, которые не всегда совпадают с грамматическими единицами. Для данной задачи широко используются программные средства статистического анализа текста, например программа Morfessor [Smit et al. 2014], которая изначально была разработана для финского языка (http://www.cis.hut.fi/projects/morpho/). При статистическом моделировании языка важной проблемой также является недостаток обучающих текстовых данных. Однако подход, основанный на использовании сублексических единиц языка, требует значительно меньше обучающих данных, чем целословные модели языка [Pellegrini, Lamel 2008].

Кроме того, некоторые МРЯ, например славянские языки (в том числе русский), характеризуются весьма свободным порядком слов в предложении по сравнению со многими языками со строгим грамматическим порядком слов (как английский или немецкий). Поэтому в первых для определения истинного порядка слов и структуры предложения зачастую необходима синтаксическая, семантическая и прагматическая информация. При этом стандартные статистические модели для таких языков будут недостаточно эффективны, потому как n-граммы высокого порядка (3-граммы и более) обладают высоким коэффициентом неопределенности (perplexity) и низким процентом совпадения для n-грамм (n-gram hit) при распознавании, поэтому для надежного обучения таких языковых моделей необходимы огромные корпуса текстовых данных (сотни миллионов словоупотреблений). Для улучшения статистических моделей разработаны методы, учитывающие синтаксическую информацию и дальнодействующие грамматические связи между словами в предложениях, например, структурные языковые модели [Chelba, Jelinek, 2000] и модифицированные n-граммные модели (в том числе синтактико-статистические) [Kuo et al. 2009; Rastrow et al. 2012; Kipyatkova et al. 2013; Karpov et al. 2014]. Кроме того, синтаксическая информация, полученная с помощью парсеров, может быть использована для оценки грамматических связей в гипотезах распознавания фраз (результат работы CAPP) [Huet et al. 2010], что может привести к более высокой точности распознавания речи за счет выбора более осмысленных гипотез. В последнее время для улучшения статистических языковых моделей и уменьшения их коэффициента неопределенности применяются также факторные (factored) модели

языка, которые помимо n-грамм слов дополнительно включают в себя и различную лингвистическую информацию (часть речи словоформы, ее лемма, основа, грамматические показатели и т. д.) [Bilmes, Kirchhoff 2003; Kipyatkova, Karpov 2014].

Отдельной проблемой является сбор обучающих текстовых данных рассматриваемого языка. Для решения этой проблемы часто используют текстовые ресурсы из Интернета (например, новостные ленты или Wikipedia) [Cai 2008], но возможно использовать и системы машинного перевода для преобразования корпуса текста одного языка в другой [Jensson et al. 2008; Cucu et al. 2012]. При этом отдельной задачей является также необходимость нормализации обучающих текстовых корпусов (числа, акронимы, аббревиатуры и т. д.), а также нахождение орфографических ошибок и опечаток в интернет-изданиях.

При создании языковых моделей для некоторых МРЯ приходится сталкиваться и с другими специфическими проблемами, например, румынский или турецкий языки активно используют диакритические знаки. Даже несмотря на то, что для человека текст без диакритики почти всегда понятен (при наличии контекста), машинное восстановление диакритических знаков является сложной задачей. При этом у ряда таких языков в текстах, которые можно собрать в Интернете, отсутствует диакритика. Но результат работы САРР, в которой не используются соответствующие диакритические знаки, может быть неоднозначным или неверным из-за различного произношения слов с диакритическими знаками и без них. Поэтому для таких языков необходимы методы автоматического восстановления диакритических знаков в гипотезах распознавания фраз [Cucu et al. 2014].

Наконец, еще одной проблемой при языковом моделировании является то, что в письменности ряда МРЯ (вьетнамского, тайского, кхмерского языков и т. д.) нет разделения текста на слова или же это разделение неоднозначно. Причем даже для тех языков, где слова разделяются специальным знаком (обычно пробелом), сегментация на слова — не всегда простая задача. При этом в САРР необходимо определение слов как элементов текста, поскольку на них основываются словарь и модель языка. Для тех языков, в которых нет очевидного разделения на слова, n-граммы слов обычно определяются по сегментированному автоматическими методами корпусу текстов [Besacier et al. 2014]. Однако из-за лексических неоднозначностей в языке и наличия внесловарных слов появляются ошибки при автоматической сегментации текста. Возможной альтернативой является вычисление вероятностей появления лексических цепочек по логографическим символам, таким как, например, кандзи в японском или ханча в корейском языке [Denoual, Lepage 2006].

4. Технологии и средства для распознавания речи МРЯ

В настоящее время существует коммерческий интерес к созданию и использованию речевых технологий для примерно 200 наиболее распространенных разговорных языков, но существуют иные важные причины для работы с остальными коммерчески малоинтересными МРЯ, например, предоставление доступа к информации на всех государственных и официальных языках, документирование речи для обогащения лингвистического знания о них, сохранения культурного наследия и т. д. Применение САРР актуально для компьютеризированного изучения вымирающих языков, а разработка программных средств для полевых лингвистов (инструментов автоматической аннотации, сегментации речи и т. д.) важна для документирования вымирающих языков и их сохранения.

Далее представлено несколько хороших примеров разработанных и внедренных САРР для различных МРЯ в разных странах. Известно, что в ЮАР почти в равной степени используются 11 официальных языков и в этой стране в последние годы стали активно развиваться технологические проекты, направленные на решение социальных проблем и сглаживание языковых барьеров [Barnard et al. 2010]. В Южной Африке активно прогрессирует и сфера разработки речевых технологий и ресурсов, охватывающая почти все языки страны. Заметным и коммерчески значимым результатом этой деятельности стало создание приложений для интернет-поиска на основе речевых запросов на зулусском, африкаанс и южно-африканском

английском. Были разработаны многоязычные САРР на основе программных средств и языковых ресурсов, предоставленных компанией Google [Ibid.]. И, хотя созданные системы дают несколько меньшую точность распознавания, чем современные САРР для английского языка, их качество все же позволяет использовать эти системы на практике.

В 2008 г. в Индии IBM India Research Laboratory был запущен масштабный проект Avaaj Otalo [Patel et al. 2009]. Индийским фермерам, которые зачастую не имеют достаточного образования даже чтобы уверенно читать, было предложено использовать голосовые сообщения для интерактивного доступа к сельскохозяйственной информации. Эта речевая технология в настоящее время доступна посредством мобильных телефонов, которые быстро распространились в индийских сельских общинах [Mohan et al. 2014]. Наиболее популярной возможностью проекта является интерактивная вопросно-ответная система по различным сельскохозяйственным и социальным темам, например, информация о рыночной стоимости продукции, прогноз погоды, информация о приезде врача, поиск наиболее дешевого оборудования или семян и т. д. Голосовой запрос абонента отправляется в САРР IBM Websphere Voice. Изначально данная САРР была создана для американского английского с большим словарем распознавания, но в ходе проекта была адаптирована к гуджарати, бенгальскому и ряду других языков Индии. В созданной системе вербальные команды на гуджарати представлены в системе фонем английского языка. При использовании такого подхода точность распознавания слов составляет до 94 % в условиях тихого помещения [Besacier et al. 2014].

Стоит упомянуть также про программный инструментарий SPICE и систему быстрого адаптирования моделей языка RLAT (The Rapid Language Adaptation Toolkit) [Schultz et al. 2007], которые предназначены для быстрого прототипирования САРР. Средства RLAT (http://csl.anthropomatik.kit.edu/rlat.php) представляют собой свободно доступные интернет-сервисы для создания акустических и языковых моделей для любого языка, а также для сбора и обработки речевых и текстовых данных. RLAT позволяет проектировать базы данных для новых языков, давая возможность пользователям создавать транскрипции к речевым данным, непрерывно собирать и обрабатывать большие объемы текстовых данных из Интернета, эффективно выбирать подходящие фонетические наборы для новых языков, использовать собранные данные при обучении акустических и языковых моделей для САРР, создавать словари произношений, а также разрабатывать модели синтеза речи по тексту.

Речевые технологии помимо САРР составляют также системы синтеза речи (text-to-speech, TTS) и машинного перевода речи (speech-to-speech translation, S2S) [Nakamura 2014]. Например, система синтеза речи от Acapela Group позволяет синтезировать речь на 30 языках (включая такие языки, как русский, чешский и каталанский) с возможностью использовать более 100 различных голосов с учетом различной интонационной окраски (http://www.acapela-group.com). Существует ряд мобильных приложений для синтеза речи по тексту. Так, сервис Google Text-to-Speech на данный момент поддерживает 14 языков и их диалектов: английский (британский, американский, индийский), итальянский, испанский, корейский, немецкий, нидерландский, польский, португальский (включая бразильский вариант), русский, французский и японский (https://play.google.com/store/apps/details?id=com.google.android.tts&hl=ru). Компьютерные системы синтеза речи по тексту также уже разработаны для нескольких африканских языков [van Niekerk, Barnard 2014; Ekpenyong et al. 2014]. Для малоресурсных языков стран СНГ тоже разработаны системы синтеза речи — для белорусского [Гецевич, Лобанов 2010] и казахского языков (http://www.speechpro.ru/media/news/2014-09-25). Кроме того, в последнее время машинный перевод речи для малоресурсных языковых пар становится все более актуальным. Проблемой систем статистического машинного перевода также являются скудные информационные ресурсы (параллельные и сопоставимые корпуса) для многих языковых пар. В качестве хороших примеров средств машинной обработки можно привести автоматические системы для вьетнамско-французского перевода речи в речь [Do et al. 2010], амхарско-английского машинного перевода [Gebreegziabher, Besacier 2012] и несколько других систем [Nakamura 2014].

Заключение

Представленный аналитический обзор и многочисленные работы по данной тематике показывают, что компьютерная обработка речи для МРЯ представляет собой активно развивающуюся в последнее десятилетие область исследований, которая сейчас переживает существенный подъем. В данном обзоре были представлены общие проблемы, подходы и достижения в области автоматического распознавания речи для МРЯ, поскольку именно на этой задаче фокусируется значительная часть сегодняшних исследований. Нужно отметить, что многие освещенные в статье вопросы и подходы применимы ко всем речевым технологиям в целом, включая системы синтеза и машинного перевода речи. Значительные успехи в этой сфере продиктованы результатами технических и междисциплинарных исследований, но для решения многих представленных проблем потребуются еще значительные совместные усилия разработчиков, лингвистов, фонетистов, математиков и т. д. Дальнейший прогресс САРР для МРЯ будет в значительной степени основываться на сборе и использовании многоязычных речевых и текстовых ресурсов. Таким образом, есть уверенность, что существующий во всем мире интерес к речевым и естественно-языковым технологиям для МРЯ будет способствовать новым исследованиям и успехам по сохранению и поддержанию таких языков и их носителей.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Асиновский и др. 2007 — Асиновский А. С., Раднаева Л. Д., Шерстинова Т. Ю. Разработка фонда звучащей речи «Языки народов России» // Альманах – 2007. Языки народов России. СПб.: Факультет филологии и искусств СПбГУ, 2007. С. 4—11. http://www.peliken.iphil.ru/Almanakh_2007.pdf [Asinovskii A. S., Radnaeva L. D., Sherstinova T. Yu. Development of the vocal speech fund «Languages of the Nations of Russia». *Almanac – 2007. Languages of the Nations of Russia*. St. Petersburg: Faculty of Philology and Arts, St. Petersburg State Univ., 2007. Pp. 4—11. Available at: http://www.peliken.iphil.ru/Almanakh_2007.pdf]

Асиновский и др. 2010 — Асиновский А. С., Богданова Н. В., Степанова С. Б., Шерстинова Т. Ю., Маркасова Е. В., Супрунова А. В. Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог’2010» (Бекасово, 26—30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 41—47. [Asinovskii A. S., Bogdanova N. V., Stepanova S. B., Sherstinova T. Yu., Markasova E. V., Suprunova A. V. Sound corpus of Russian “One day of speech”: ways of updating and intellectual technologies. *Komp’yuternaya lingvistika i intellektual’nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog’2010»* (Bekasovo, May 26—30, 2010). No. 9 (16). Moscow: Russian State Univ. for the Humanities , 2010. Pp. 41—47.]

Гецевич, Лобанов 2010 — Гецевич Ю. С., Лобанов Б. М. Система синтеза белорусской речи по тексту // Речевые технологии. 2010. № 1. С. 92—101. [Getsevich Yu. S., Lobanov B. M. A system of Belorussian text-to-speech synthesis. *Rechevye tekhnologii*. 2010. No. 1. Pp. 92—101.]

Казакевич, Кибрик 2005 — Казакевич О. А., Кибрик А. Е. Малые языки на постсоветском пространстве. Малые языки и традиции: существование на грани. Вып. 1. Лингвистические проблемы сохранения и документации малых языков М.: Новое издательство, 2005. С.13—39. [Kazakevich O. A., Kibrik A. E. *Malye yazyki na postsovetskom prostranstve. Malye yazyki i traditsii: sushchestvovanie na grani. Vyp. 1. Lingvisticheskie problemy sokhraneniya i dokumentatsii malykh yazykov* [Small languages on the post-Soviet territory. Small languages and traditions: existence on the verge. No. 1. Linguistic problems of preservation documenting of small languages. Moscow: Novoe Izdatel’stvo, 2005. Pp.13—39.]

Карпов 2011 — Карпов А. А. Компьютерный анализ и синтез русского жестового языка // Вопросы языкознания. 2011. № 6. С. 41—53. [Karpov A. A. Computer analysis and synthesis of Russian sign language. *Voprosy jazykoznanija*. 2011. No. 6. Pp. 41—53.]

Карпов 2015 — Карпов А. А. 4-й Международный семинар по речевым технологиям для малоресурсных языков SLTU-2014 // Вопросы языкознания. 2015. № 2. С. 150—152. [Karpov A. A. 4th International workshop on spoken language technologies for under-resourced languages. *Voprosy jazykoznanija*. 2015. № 2. Pp. 150—152.]

Кибрик, Подлесская 2009 — Кибрик А. А., Подлесская В. И. (ред.). Рассказы о сновидениях: корпусное исследование устного русского дискурса. М.: Языки славянских культур, 2009. С. 288—308. [Kibrik A. A., Podlesskaya V. I. (eds). *Rasskazy o snovideniyakh: korpusnoe issledovanie ustnogo russkogo diskursa* [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Yazyki Slavyanskikh Kul'tur, 2009. Pp. 288—308].

Кипяткова, Карпов 2010 — Кипяткова И. С., Карпов А. А. Аналитический обзор систем распознавания русской речи с большим словарем // Труды СПИИРАН. 2010. Вып. 12. С. 7—20. http://proceedings.spiiras.nw.ru/ojs/index.php/sp/article/download/1474/1337 [Kipyatkova I. S., Karpov A. A. An analytical survey of large vocabulary Russian speech recognition systems. *SPIIRAS Proceedings*. 2010. No. 12. Pp. 7—20. http://proceedings.spiiras.nw.ru/ojs/index.php/sp/article/download/1474/1337]

Кипяткова и др. 2013 — Кипяткова И. С., Ронжин А. Л., Карпов А. А. Автоматическая обработка разговорной русской речи. СПб.: ГУАП, 2013. [Kipyatkova I. S., Ronzhin A. L., Karpov A. A. *Avtomaticheskaya obrabotka razgovornoi russkoi rechi* [Automated processing of the conversational Russian speech]. St. Petersburg: St. Petersburg State University of Airspace Instrumentation, 2013.]

Мурадова 2008 — Мурадова А. Р. Как исчезают языки и как их возрождают. Языковое разнообразие в киберпространстве: российский и зарубежный опыт. М.: МЦБС. 2008. С. 70—75. http://www.philology.ru/linguistics1/muradova-08.htm [Muradova A. R. *Kak ischezayut yazyki i kak ikh vozrozhdayut. Yazykovoe raznoobrazie v kiberprostranstve: rossiiskii i zarubezhnyi opyt* [Disappearance and revival of languages. Linguistic diversity in the cyberspace: Russian and foreign practices]. Moscow: Interregional Centre for Library Cooperation, 2008. Pp. 70—75. Available at: http://www.philology.ru/linguistics1/muradova-08.htm]

Плунгян 2010 — Плунгян В. А. Почему языки такие разные. М.: АСТ-ПРЕСС КНИГА, 2010. [Plungian V. A. *Pochemu yazyki takie raznye* [Why languages are so different]. Moscow: AST-PRESS KNIGA, 2010.]

Потапова 2009 — Потапова Р. К. Основные тенденции развития многоязычной корпусной лингвистики // Речевые технологии. 2009. № 2. С. 92—114. [Potapova R. K. The main trends in the development of multilingual corpus-based linguistics. *Rechevye tekhnologii*. 2009. No 2. Pp. 92—114.]

Ронжин и др. 2006 — Ронжин А. Л., Карпов А. А., Лобанов Б. М., Цирульник Л. И., Йокиш О. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи // Информационно-управляющие системы. 2006. № 6 (25). С. 24—34. [Ronzhin A. L., Karpov A. A., Lobanov B. M., Tsirulnik L. I., Jokisch O. Phonetic-morphological mapping of speech corpora for recognition and synthesis of Russian speech. *Informatsionno-upravlyayushchie sistemy*. 2006. No. 6 (25). Pp. 24—34.]

Старостин 2007 — Старостин С. А. Сравнительное языкознание и этимологические базы данных // Старостин С. А. Труды по языкознанию. М.: Языки славянских культур, 2007. С. 770—778. [Starostin S. A. Comparative linguistics and etymological databases. Starostin S. A. *Trudy po yazykoznaniyu*. Moscow: Yazyki Slavyanskikh Kul'tur, 2007. Pp. 770—778.]

Хусаинов 2014 — Хусаинов А. Ф. Технология автоматизации создания и оценки качества программных средств анализа речи с учетом особенностей малоресурсных языков. Дис. … канд. техн. наук. Уфа, 2014. [Khusainov A. F. *Tekhnologiya avtomatizatsii sozdaniya i otsenki kachestva programmnykh sredstv analiza rechi s uchetom osobennostei maloresursnykh yazykov*. Kand. diss. [The technology of automation of speech analysis tools software development and quality control by reference to specific features of under-resourced languages. Cand. diss.]. Ufa, 2014.]

Челышева 2009 — Челышева И. И. (ред.). Миноритарные языки Евразии. Проблемы языковых контактов. М.: Тезаурус, 2009. [Chelysheva I. I. (ed.). *Minoritarnye yazyki Evrazii. Problemy yazykovykh kontaktov* [Minority languages of Eurasia. Problems of language contacts]. Moscow: Thesaurus, 2009.]

Arlazarov et al. 2004 — Arlazarov V., Bogdanov D., Krivnova O., Podrabinovich A. Creation of Russian speech databases: Design, processing, development tools. *Proc. International conference on speech and computer SPECOM-2004.* St. Petersburg, Russia, 2004. Pp. 650—656.

Barnard et al. 2009 — Barnard E., Davel M., van Heerden C. ASR corpus design for resource-scarce languages. *Proc. International conference Interspeech-2009*. Brighton, UK, 2009. Pp. 2847—2850.

Barnard et al. 2010 — Barnard E., Davel M., van Huyssteen G. Speech technology for information access: a South African case study. *Proc. AAAI spring symposium on artificial intelligence for development AI-D*. Palo Alto, USA, 2010. Pp. 8—13.

Berment 2004 — Berment V. Méthodes pour informatiser des langues et des groupes de langues «peu dotées». Thèse de doctorat, Université Joseph Fourier. Grenoble, 2004.

Besacier et al. 2006 — Besacier L., Zhou B., Gao Y. Towards speech translation of non written languages. *Proc. IEEE/ACL spoken language technology workshop SLT-2006*. Aruba, 2006. Pp. 222—225.

Besacier et al. 2014 — Besacier L., Barnard E., Karpov A., Schultz T. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*. 2014. Vol. 56. Pp. 85—100.

Bilmes, Kirchhoff 2003 — Bilmes J. A., Kirchhoff K. Factored language models and generalized parallel backoff. *Proc. Conference of the North American chapter of the association for computational linguistics on human language technology*. Stroudsburg, PA, USA. Vol. 2. 2003. Pp. 4—6.

Cai 2008 — Cai J. Transcribing Southern Min speech corpora with a web-based language learning system. *Proc. 1st International workshop on spoken language technologies for under-resourced languages SLTU-2008*. Hanoi, Vietnam, 2008. Pp. 659—664.

Chan, Rosenfeld 2012 — Chan H. Y., Rosenfeld R. Discriminative pronunciation learning for speech recognition for resource scarce languages. *Proc. 2nd ACM symposium on computing for development ACM DEV*. New York, USA, 2012. Article No. 12.

Chelba, Jelinek 2000 — Chelba C., Jelinek F. Structured language model. *Computer speech and language*. 2000. Vol. 14. No. 4. Pp. 283—332.

Crystal 2000 — Crystal D. *Language death*. Cambridge: Cambridge University Press, 2000.

Cucu et al. 2012 — Cucu H., Besacier L., Burileanu C., Buzo A. ASR domain adaptation methods for low-resourced languages: application to Romanian language. *Proc. European conference on signal processing EUSIPCO-2012*. Bucharest, Romania, 2012. Pp. 1648—1652.

Cucu et al. 2014 — Cucu H., Buzo A., Besacier L., Burileanu C. SMT-based ASR domain adaptation methods for under-resourced languages: application to Romanian. *Speech communication*. 2014. Vol. 56. Pp. 195—212.

Davel et al. 2011 — Davel M., van Heerden C., Kleynhans N., Barnard E. Efficient harvesting of Internet audio for resource-scarce ASR. *Proc. International conference Interspeech-2011*. Florence, Italy, 2011. Pp. 3153—3156.

de Vries et al. 2014 — de Vries N., Davel M., Badenhorst J., Basson W., de Wet F., Barnard E., de Waal A. A smartphone-based ASR data collection tool for under-resourced languages *Speech communication*. 2014. Vol. 56. Pp. 119—131.

Denoual, Lepage 2006 — Denoual E., Lepage Y. The character as an appropriate unit of processing for non-segmenting languages. *Proc. NLP annual meeting*. Tokyo, Japan, 2006. Pp. 731—734.

Do et al. 2010 — Do T., Besacier L., Castelli E. Unsupervised SMT for a low resourced language pair. *Proc. 2nd International workshop on spoken language technologies for under-resourced languages SLTU-2010*. Penang, Malaysia, 2010. Pp. 130—135.

Ekpenyong et al. 2014 — Ekpenyong M., Urua E.-A., Watts O., King S., Yamagishi J. Statistical parametric speech synthesis for Ibibio. *Speech communication*. 2014. Vol. 56. Pp. 243—251.

Gales et al. 2014 — Gales M., Knill K., Ragni A., Rath S. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. *Proc. 4th International workshop on spoken language technologies for under-resourced languages SLTU-2014*. St. Petersburg, Russia, 2014. Pp. 16—23.

Gebreegziabher, Besacier 2012 — Gebreegziabher M., Besacier L. Preliminary experiments on English-Amharic statistical machine translation. *Proc. 3rd International workshop on spoken language technologies for under-resourced languages SLTU-2012*. Cape Town, South Africa, 2012. Pp. 36—41.

Gelas et al. 2011 — Gelas H., Teferra Abate S., Besacier L., Pellegrino F. Quality assessment of crowdsourcing transcriptions for African languages. *Proc. International conference Interspeech-2011*. Florence, Italy, 2011. Pp. 3065—3068.

Gemmeke, van Hamme 2011 — Gemmeke J. F., van Hamme H. An hierarchical exemplar-based sparse model of speech with an application to ASR. *Proc. IEEE international workshop ASRU-2011*. Hawaii, USA, 2011. Pp. 101—106.

Ghoshal et al. 2009 — Ghoshal A., Jansche M., Khudanpur S., Riley M., Ulinski M. Web-derived pronunciations. *Proc. IEEE international conference on acoustics, speech, and signal processing ICASSP-2009*. Taipei, Taiwan, 2009. Pp. 4289—4292.

Hartmann et al. 2014 — Hartmann W., Lamel L., Gauvain J.-L. Cross-word sub-word units for low-resource keyword spotting. *Proc. 4th International workshop on spoken language technologies for under-resourced languages SLTU-2014*. St. Petersburg, Russia, 2014. Pp. 112—117.

Huet et al. 2010 — Huet S., Gravier G., Sebillot P. Morpho-syntactic postprocessing of N-best lists for improved French automatic speech recognition. *Computer speech and language*. 2010. Vol. 24. No. 4. Pp. 663—684.

Hughes et al. 2010 — Hughes T., Nakajima K., Ha L., Moreno P., LeBeau M. Building transcribed speech corpora quickly and cheaply for many languages. *Proc. International conference Interspeech-2010*. Makuhari, Japan, 2010. Pp. 1914—1917.

Jensson et al. 2008 — Jensson A., Iwano K., Furui S. Development of a speech recognition system for Icelandic using machine translated text. *Proc. 1st International workshop on spoken language technologies for under-resourced languages SLTU-2008*. Hanoi, Vietnam, 2008. Pp. 18—21.

Karanasou, Lamel 2010 — Karanasou P., Lamel L. Comparing SMT methods for automatic generation of pronunciation variants. *Proc. International workshop IceTAL-2010*. Reykjavik, Iceland, 2010. P. 167.

Karpov et al. 2011 — Karpov A., Kipyatkova I., Ronzhin A. Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. *Proc. International conference Interspeech-2011*. Florence, Italy, 2011. Pp. 3161—3164.

Karpov et al. 2014 — Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A. Large vocabulary Russian speech recognition using syntactico-statistical language modeling. *Speech communication*. 2014. Vol. 56. Pp. 213—228.

Kempton, Moore 2014 — Kempton T., Moore R. K. Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech communication*. 2014. Vol. 56. Pp. 152—166.

Kipyatkova, Karpov 2014 — Kipyatkova I., Karpov A. Study of morphological factors of factored language models for Russian ASR. *Proc. 16th International conference on speech and computer SPECOM-2014*. Springer, Lecture Notes in Artificial Intelligence. Vol. 8773. 2014. Pp. 451—458.

Kipyatkova et al. 2013 — Kipyatkova I., Karpov A., Verkhodanova V., Zelezny M. Modeling of pronunciation, language and nonverbal units at conversational Russian speech recognition. *International journal of computer science and applications*. 2013. No. 10 (1). Pp. 11—30.

Krauwer 2003 — Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. *Proc. International workshop on speech and computer SPECOM-2003*. Moscow, Russia, 2003. Pp. 8—15.

Kuo et al. 2009 — Kuo H.-K. J., Mangu L., Emami A., Zitouni I., Lee Y.-S. Syntactic features for Arabic speech recognition. *Proc. International workshop ASRU-2009*. Merano, Italy, 2009. Pp. 327—332.

Lamel et al. 2012 — Lamel L., Courcinous S., Gauvain J.-L., Josse Y., Le V. B. Transcription of Russian conversational speech. *Proc. 3rd International workshop SLTU-2012*. Cape Town, South Africa, 2012. Pp. 156—161.

Laurent et al. 2009 — Laurent A., Deleglise P., Meignier S. Grapheme to phoneme conversion using an SMT system. *Proc. International conference Interspeech-2009*. Brighton, UK, 2009. Pp. 708—711.

Le, Besacier 2009 — Le V.-B., Besacier L. Automatic speech recognition for under-resourced languages: application to Vietnamese language. *IEEE transactions on audio, speech and language processing*. 2009. Vol. 17 (8). Pp. 1471—1482.

Mohan et al. 2014 — Mohan A., Rose R., Ghalehjegh S. H., Umesh S. Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. *Speech communication*. 2014. Vol. 56. Pp. 167—180.

Nakamura 2014 — Nakamura S. Towards real-time multilingual multimodal speech-to-speech translation. *Proc. 4th International workshop on spoken language technologies for under-resourced languages SLTU-2014*. St. Petersburg, Russia, 2014. Pp. 13—15.

Parent, Eskenazi 2010 — G. Parent, M. Eskenazi. Toward better crowdsourced transcription: transcription of a year of the Let's Go bus information system data. *Proc. IEEE workshop on spoken language technology SLT-2010*. Berkeley, USA, 2010. Pp. 312—317.

Patel et al. 2009 — Patel N., Chittamuru D., Jain A., Dave P., Parikh T. S. Avaaj Otalo: A field study of an interactive voice forum for small farmers in rural India. *Proc. ACM international conference CHI-2009*. Boston, USA, 2009. Pp. 733—742.

Pellegrini, Lamel 2008 — Pellegrini T., Lamel L. Are audio or textual training data more important for ASR in less-represented languages? *Proc. 1st International workshop on spoken language technologies for under-resourced languages SLTU-2008*. Hanoi, Vietnam, 2008. Pp. 2—6.

Rastrow et al. 2012 — Rastrow A., Dredze M., Khudanpur S. Fast syntactic analysis for statistical language modeling via substructure sharing and uptraining. *Proc. 50th Annual meeting of association for computational linguistics ACL-2012*. Jeju, Korea, 2012. Pp. 175—183.

Potapova 2011 — Potapova R. Multilingual spoken language databases in Russia. *Proc. International conference on speech and computer SPECOM-2011*. Kazan, Russia, 2011. Pp. 13—17.

Ronzhin, Karpov 2007 — Ronzhin A., Karpov A. Russian voice interface. *Pattern recognition and image analysis*. 2007. Vol. 17 (2). Pp. 321—336.

Rotovnik et al. 2007 — Rotovnik T., Maucec M. S., Kacix Z. Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech communication*. 2007. Vol. 49. No. 6. Pp. 437—452.

Roux et al. 2000 — Roux J. C., Botha E. C., du Preez J. A. Developing a multilingual telephone based information retrieval system in African languages. *Proc. 2nd International conference on language resources and evaluation LREC-2000*. Athens, Greece, 2000. Pp. 975—980.

Schlippe et al. 2010 — Schlippe T., Ochs S., Schultz T. Wiktionary as a source for automatic pronunciation extraction. *Proc. International conference Interspeech-2010*. Makuhari, Japan, 2010. Pp. 2290—2293.

Schlippe et al. 2012 — Schlippe T., Ochs S., Vu N. T., Schultz T. Automatic error recovery for pronunciation dictionaries. *Proc. International conference Interspeech-2012*. Portland, USA, 2012. Pp. 2298—2301.

Schlippe et al. 2014 — Schlippe T., Ochs S., Schultz T. Web-based tools and methods for rapid pronunciation dictionary creation. *Speech communication*. 2014. Vol. 56. Pp. 101—118.

Schultz, Kirchhoff 2006 — *Multilingual speech processing.* Schultz T., Kirchhoff K. (eds). Burlington, MA: Elsevier Academic Press, 2006.

Schultz et al. 2007 — Schultz T., Black A., Badaskar S., Hornyak M., Kominek J. SPICE: web-based tools for rapid language adaptation in speech processing systems. *Proc. International conference Interspeech-2007*. Antwerp, Belgium, 2007. Pp. 2125—2128.

Schultz et al. 2013 — Schultz T., Vu N. T., Schlippe T. GlobalPhone: a multilingual text & speech database in 20 languages. *Proc. International conference ICASSP-2013.* Vancouver, Canada, 2013. Pp. 8126—8130.

Siniscalchi et al. 2013 — Siniscalchi S. M., Reed J., Svendsen T., Lee C.-H. Universal attribute characterization of spoken languages for automatic spoken language recognition. *Computer speech and language*. 2013. Vol. 27. No. 1. Pp. 209—227.

Smit et al. 2014 — Smit P., Virpioja S., Grönroos S. A., Kurimo M. Morfessor 2.0. Toolkit for statistical morphological segmentation. *Proc. 14th Conference of the European chapter of the association for computational linguistics EACL-2014*. Gothenburg, Sweden, 2014. Pp. 21—24.

Stahlberg et al. 2012 — Stahlberg F., Schlippe T., Vogel S., Schultz T. Word segmentation through cross-lingual word-to-phoneme alignment. *Proc. IEEE workshop on spoken language technology SLT-2012*. Miami, USA, 2012. Pp. 85—90.

Stuker et al. 2003 — Stuker S., Schultz T., Metze F., Waibel A. Multilingual articulatory features. *Proc. IEEE international conference on acoustics, speech and signal processing ICASSP-2003*. Hong Kong, China, 2003. Vol. I. Pp. 144—147.

Stuker et al. 2009 — Stuker S., Besacier L., Waibel A. Human translations guided language discovery for ASR systems. *Proc. International conference Interspeech-2009*. Brighton, UK, 2009. Pp. 3023—3026.

Tachbelie et al. 2014 — Tachbelie M., Abate S. T., Besacier L. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language — Amharic. *Speech communication*. 2014. Vol. 56. Pp. 181—194.

van Heerden et al. 2010 — van Heerden C., Kleynhans N., Barnard E., Davel M. Pooling ASR data for closely related languages. *Proc. 2nd International workshop on spoken language technologies for under-resourced languages SLTU-2010*. Penang, Malaysia, 2010. Pp. 17—23.

van Niekerk, Barnard 2014 — van Niekerk D. R., Barnard E. Predicting utterance pitch targets in Yoruba for tone realisation in speech synthesis. *Speech communication*. 2014. Vol. 56. Pp. 229—242.

Vazhenina et al. 2012 — Vazhenina D., Kipyatkova I., Markov K., Karpov A. State-of-the-art speech recognition technologies for Russian language. *Proc. Joint international conference on human-centered computer environments HCCE-2012*. Aizu-Wakamatsu, Japan, 2012. Pp. 59—63.

Vu et al. 2010 — Vu N. T., Kraus F., Schultz T. Multilingual A-stabil: A new confidence score for multilingual unsupervised training. *Proc. IEEE workshop on spoken language technology SLT-2010*. Berkeley, USA. Pp. 183—188.

Whittaker, Woodland 2003 — Whittaker E. W. D., Woodland P. C. Language modelling for Russian and English using words and classes. *Computer speech and language*. 2003. Vol. 17. No. 1. Pp. 87—104.

Wissing, Barnard 2008 — Wissing D., Barnard E. Vowel variations in Southern Sotho: An acoustical investigation. *Southern African linguistics and applied language studies*. 2008. Vol. 26 (2). Pp. 255—265.

Статья поступила в редакцию 08.08.2014.