

КОРПУСА И КОРПУСНЫЕ ИССЛЕДОВАНИЯ ЯЗЫКОВ РОССИЙСКОЙ ФЕДЕРАЦИИ

© 2024 г. Т.И. Давидюк^{a,b,*}, А.А. Кибрик^{a,b,**}, Д.Д. Мордашова^{a,***}

^aИнститут языкознания РАН, Москва, Россия

^bМосковский государственный университет имени М.В. Ломоносова, Москва, Россия

*E-mail: davidyuk@iling-ran.ru

**E-mail: aakibrik@iling-ran.ru

***E-mail: d.mordashova@iling-ran.ru

Поступила в редакцию 09.07.2024 г.

После доработки 09.07.2024 г.

Принята к публикации 23.07.2024 г.

Статья посвящена описанию корпусных ресурсов по языкам России и их использованию в лингвистических исследованиях. Языковое разнообразие РФ весьма значительно — в настоящее время насчитывается 155 языков России. Многие из них находятся под угрозой исчезновения, что делает особенно актуальной задачу создания корпусов — инструмента, позволяющего сохранять языки и языковые данные. В рамках данной работы был проведён опрос среди сотрудников Института языкознания РАН и других лингвистов, собраны сведения о 73 корпусных ресурсах, представляющих различные языки России и их диалекты. Выборка охватывает как крупные языки, так и языки с небольшим числом носителей, включая бесписьменные языки.

В статье подробно рассматриваются параметры, по которым могут различаться корпуса, приводятся примеры исследований, проведённых на основе материалов корпусов. Обсуждаются организационные аспекты создания и поддержания корпусных ресурсов, которые не только играют важную роль в сохранении языкового разнообразия России, но и представляют собой ценный инструмент для решения различных исследовательских задач, а также для создания других языковых ресурсов. Статья основана на научном сообщении, с которым А.А. Кибрик выступил на заседании Президиума РАН 9 апреля 2024 г.

Ключевые слова: лингвистические корпуса, языковые ресурсы, языки России, сохранение языков.

DOI: 10.31857/S0869587324090039, **EDN:** FCFZQL

Известно, что языковое разнообразие России велико. Однако определить точное число языков страны — непростая задача. Во-первых, часто бывает трудно установить, является ли определённый идиом самостоятельным языком или диалектом

какого-то языка. Во-вторых, необходимо разработать критерии для включения того или иного языка в список языков России, поскольку многие языки, распространённые на территории нашей страны, используются также и в других государствах.



ДАВИДЮК Татьяна Игоревна — младший научный сотрудник ИЯз РАН, программист НИВЦ МГУ имени М.В. Ломоносова, аспирант МГУ имени М.В. Ломоносова. КИБРИК Андрей Александрович — доктор филологических наук, директор ИЯз РАН, заведующий отделом типологии и ареальной лингвистики ИЯз РАН, профессор МГУ имени М.В. Ломоносова. МОРДАШОВА Дарья Дмитриевна — младший научный сотрудник ИЯз РАН.

В-третьих, иногда в случае исчезающих языков сохраняется неопределённость относительно того, исчез какой-либо язык полностью или же знания о нём ещё сохраняются в сообществе его носителей (на этом основано противопоставление живых и так называемых “заснувших” языков).

В исследовании [1], проведённом сотрудниками Института языкознания РАН, указанные вопросы находят конкретные решения. Чтобы понять, является ли некоторый идиом языком или диалектом, было предложено использовать три критерия: традиционное описание идиома как языка или диалекта; степень взаимопонимания между носителями и процент сходства в базовой лексике; наличие различных письменных традиций и различия в языковом и/или этническом самосознании носителей. Чтобы определить, можно ли считать конкретный язык языком России или нет, рекомендовано учитывать компактность проживания его носителей (в том числе в недавнем прошлом) и количество носителей, живущих в России. На основании названных критериев, а также данных, полученных от лингвистического сообщества, был составлен список из 155 языков России.

Российская Федерация предоставляет большой простор для изучения языкового разнообразия. Согласно современным лингвистическим представлениям, одним из направлений исследовательской работы по изучению языков является создание *корпусов*. Корпус — это обширная коллекция собранных, систематизированных и аннотированных письменных и/или устных текстов на некотором языке, позволяющая анализировать данные о лексических и грамматических свойствах языка, о его употреблении, а также о разнообразных проявлениях внутриязыковой вариативности. Корпусная лингвистика — важный и бурно развивающийся

раздел современного языкознания, перед которым ставится широкий спектр задач [2].

Особую значимость корпусные исследования приобретают в условиях угрозы исчезновения языков. Для более чем 43% языков России характерна прерванная межпоколенческая передача; такие языки могут быть утрачены после ухода последнего поколения их носителей [1]. Именно поэтому создание корпусов становится важным шагом в сохранении языков, предотвращении их полного исчезновения, а это важная социальная задача [3, 4]. Имеется успешный опыт создания специализированного корпуса для языка лабрадорский инутитут [5]. Этот корпус стал важным инструментом для обучения языку. Подобные инициативы показывают значимость корпусных ресурсов не только для документации языков, но и для их развития и возрождения. Более того, в случае полного исчезновения языка корпусные ресурсы предоставляют основу для восстановления и возрождения того или иного языка благодаря записанным текстам и аудиоматериалам.

Обзор корпусных ресурсов. В связи с отсутствием единой базы данных, содержащей информацию о корпусных ресурсах языков России, был проведён опрос сотрудников Института языкознания РАН и других специалистов в этой области. Опрашиваемым было предложено предоставить информацию об известных им корпусах, включая следующие сведения: название языка, для которого составлен корпус; адрес ресурса в Интернете; объём корпуса; его особенности и характеристики; примеры исследований, проведённых на основе материалов корпуса. Список корпусов, полученный в результате опроса, оказался достаточно репрезентативным. Получены сведения о 73 ресурсах, включающих различные типы корпусов. Учитывая значительное языковое разнообразие России, а также децентрализованный

Таблица 1. Крупные корпуса по языкам России

Язык	Адрес корпуса в Интернете	Количество словоупотреблений, млн
Татарский	https://www.corpus.tatar/	больше 620
Коми-зырянский	https://komicorpora.ru/	82.8
Лугово-восточный марийский	https://gtweb.uit.no/u_korp/?mode=mhr	57.4
Удмуртский	https://udmcorpus.udman.ru/	45.0
Чувашский	https://ru.corpus.chv.su/	16.1
Эрзянский	https://gtweb.uit.no/u_korp/?mode=myv#!lang=en	14.0
Мокшанский	https://gtweb.uit.no/u_korp/?mode=mdf#!lang=en	12.8
Адыгейский	http://adyghe.web-corpora.net/	10.7
Удмуртский	http://udmurt.web-corpora.net/udmurt_corpus/search	9.6
Калмыцкий	http://kalmcorpora.ru/	8.7
Даргинский (литературный)	http://lingconlab.ru/standard_dargwa/	6.4
Бурятский	http://buriat.web-corpora.net/	2.8

Таблица 2. Примеры небольших по объёму корпусов

Язык	Адрес корпуса в Интернете	Количество словоупотреблений, тыс.
Камасинский	https://inel.corpora.uni-hamburg.de/KamasCorpus/search	63.8
Горномарийский	https://hillmari-exp.tilda.ws/corpus	63.5
Хакасский	http://lingconlab.ru/spoken_khakas/	57.6
Северноселькупский	https://corpora.iling-ran.ru/selkup	52.0
Эвенкийский	https://inel.corpora.uni-hamburg.de/EvenkiCorpus/search	47.8
Кетский	https://corpora.iling-ran.ru/ketsk/	15.9
Телеутский	https://corpora.iea.ras.ru/corpora/search.php	9.0
Ботлихский	http://lingconlab.ru/botlikh/	1.0

характер работы лингвистов, не остаётся сомнений, что это неполный список. Его расширение — задача дальнейших исследований.

В таблице 1 представлены наиболее крупные корпуса по языкам России (за исключением русского языка) с информацией о количестве словоупотреблений. В него входят государственные языки нескольких республик. Такие значительные объёмы корпусов достигаются во многом благодаря наличию большого количества письменных текстов на языках и автоматического морфологического анализатора для аннотирования корпуса.

Имеются и более скромные по объёмам корпуса (табл. 2), которые нередко представляют собой коллекции текстов, собранных в рамках полевых исследований. В отличие от крупных языков, многие из которых обладают развитой письменной традицией, языки с небольшим числом носителей часто являются бесписьменными или младописьменными (или имеют прерванную письменную традицию). Поэтому корпуса по таким языкам состоят преимущественно из образцов устных жанров — рассказов, песен, легенд и сказок. В связи с отсутствием или ограниченными возможностями автоматической разметки аннотация текстов в таких корпусах часто проводится вручную. Это снижает возможный объём корпуса, поскольку ручная разметка требует значительных временных и человеческих затрат. Тем не менее, несмотря на технические ограничения, небольшие корпуса играют важную роль в сохранении и изучении уникальных языковых и культурных характеристик отдельных сообществ.

Существует особый тип корпусов, ориентированных на диалектные варианты языков. В диалектных корпусах собраны тексты на вариантах языка, которые могут значительно отличаться по фонетике, грамматике и лексике от литературного стандарта. Примерами таких корпусов могут служить корпуса по диалектам даргинского языка¹, цналь-

скому говору лезгинского языка², татышлинскому говору удмуртского языка³ или малокарачкинскому говору чувашского языка⁴. Такие корпуса дают возможность не только исследовать характеристики конкретного диалекта и диалектное разнообразие языка в целом, но и зафиксировать отличительные особенности региональных этнических групп.

Следует отметить ресурс ЛингвоДок⁵, обладающий развитой корпусной инфраструктурой. На этой платформе собраны корпуса (в том числе с привлечением диалектного материала) практически всех уральских, а также некоторых тюркских языков. Эти корпуса содержат как современные звуковые записи, так и письменные тексты, включая тексты первых книг, написанных в XIX и начале XX вв. На платформе представлены корпуса литературных языков, глоссированные со снятой омонимией⁶ для большинства уральских языков, а также для некоторых тюркских языков (башкирского и якутского). Создание глоссированных корпусов со снятой омонимией стало возможным благодаря функционалу платформы, который позволяет работать одновременно нескольким пользователям, наличию парсеров для целого ряда языков, а также онлайн-интерфейсу снятия омонимии. На основе ЛингвоДока создана обучающая платформа⁷, которая работает на базе корпусов для девяти языков Российской Федерации. Необходимо подчеркнуть, что все данные, выложенные на ЛингвоДоке, находятся в открытом доступе. Наличие такого крупного ресурса существенно ускоряет анализ и обработку лингвистических данных.

² http://lingconlab.ru/tsnal_lezgi/

³ https://udmurt.web-corpora.net/tatyshly_corpus/search

⁴ <http://lingconlab.ru/chuvash/>

⁵ <https://lingvodoc.ispras.ru/>

⁶ В обычных текстах встречаются случаи омонимии словоформ (например, *три* — повелительное наклонение ед.ч. от глагола *тереть*, а также числительное 3), и снятие такой неоднозначности является специальной технической задачей. В корпусах со снятой омонимией каждой неоднозначной словоформе соответствует один разбор, соответствующий контексту.

⁷ <https://edu.ispras.ru/>

¹ <https://www.dargwa.ru/corpora>

Языки России представлены также на таком крупном ресурсе, как Национальный корпус русского языка⁸, где, помимо прочего, имеются параллельные корпуса, например бурятско-русский параллельный корпус, башкирско-русский параллельный корпус и хакасско-русский параллельный корпус [6].

Существуют также корпуса различных вариантов русского языка, распространённых в многоязычных регионах, например корпус дагестанского русского⁹. Наряду с корпусами региональных разновидностей русского языка разрабатываются и устные корпуса стандартного русского. Так, в состав Национального корпуса русского языка входят устный и мультимедийный подкорпуса. Корпус русской устной речи с детальной дискурсивной разметкой был создан в рамках проекта «Рассказы о сновидениях и другие корпуса звучащей речи»¹⁰. Достаточно подробной дискурсивной разметкой обладает и ресурс «Русские рассказы и разговоры о грушах»¹¹.

Нужно сказать и о русском жестовом языке (РЖЯ), используемом глухими и слабослышащими людьми для общения при помощи визуального канала. Как и другие жестовые языки, РЖЯ является полноценным языком, обладающим своей лексикой и грамматикой, и служит основным средством коммуникации внутри сообществ глухих и слабослышащих людей. Сегодня мы располагаем корпусом русского жестового языка¹².

Типы корпусов. Корпусные ресурсы различаются по множеству параметров, которые будут проиллюстрированы на примере нашей выборки. Учёт этих различий позволяет исследователям выбирать подходящие корпусные ресурсы в зависимости от конкретных целей и задач исследования.

Важный параметр — доступ к корпусу, который может быть либо свободным, либо требующим предварительной регистрации. Примеры открытых корпусов вместе со ссылками на их адреса в Интернете приведены в таблицах 1 и 2. Система с разными уровнями доступа представлена на платформе Endangered Languages Archive (ELAR), где имеются в том числе материалы по некоторым языкам России (например, коллекция текстов на чирагском даргинском¹³), большинство из которых имеют помету U, означающую, что эти тексты доступны для скачивания только зарегистрированными пользователями (регистрация на сайте бесплатна).

Типы материалов, содержащихся в корпусах, разнообразны: существуют письменные корпуса, корпуса с устным компонентом и корпуса, состоящие только из устных материалов, представленных либо со звуком, либо только в виде транскриптов.

Так, корпуса, приведённые в таблице 1, преимущественно письменные. Аудиозаписи нередко сопровождают корпуса диалектной речи: например, корпус татышлинского говора удмуртского языка¹⁴ и корпус муиринского даргинского¹⁵ являются звучащими корпусами. В окне выдачи при поиске можно нажать на найденное предложение и прослушать его. Отдельные корпуса могут содержать также видеоматериалы, примером такого ресурса может служить коми медиатека¹⁶.

Корпуса также могут включать в себя разметку жанров. К примеру, лугово-восточно-марийский корпус, размещённый на платформе Корр, включает в себя тексты различных жанров, которые можно выбрать для исследования — нон-фикшн, тексты законов, научные тексты, тексты новостей и тексты Википедии. Иногда корпуса содержат тексты только какого-то определённого жанра. Это, в частности, Башкирский поэтический корпус¹⁷, текстовую коллекцию которого представляют стихотворные произведения башкирских поэтов XX и начала XXI в.

Как правило, корпуса включают морфологический анализ словоформ — указание на основную форму и пометы, которые говорят о том, какая именно форма представлена в предложении. На рисунке 1 показано одно из предложений устного корпуса диалектов хакасского языка¹⁸. Например, в словоформе *абам* выделяется суффикс *-(b)м-*, который является посессивным (притяжательным) показателем 1 лица единственного числа (помета 1ros. sg). Нередко корпуса предоставляют также перевод предложений на русский или английский язык, что облегчает их использование широким кругом лиц. Так, в устном корпусе хакасских текстов (рис. 1) имеется перевод на русский язык.

Интерфейс доступа к корпусу также может быть разным. Для размещения корпусов в Интернете существуют специальные платформы. Многие корпуса по языкам России размещены на платформе Tsakorpus, разработанной Т.А. Архангельским. Данная платформа предлагает пользователям понятный интерфейс и достаточно богатый функционал для работы с корпусом [7]. Некоторые корпуса по языкам России размещены на платформе Корр, в частности корпуса по лугово-восточному марийскому, эрзянскому и мокшанскому языкам.

Исследования на основе корпусов. Создание корпуса — одна из возможных целей документации того или иного языка. Зачастую корпус выступает не в качестве цели исследования, а как его инструмент. Корпусные исследования играют ключевую роль в современной лингвистике, предоставляя учёным доступ к обширным коллекциям текстово-

⁸ <https://ruscorpora.ru/>

⁹ <https://parasolcorpus.org/dagrus/#/>

¹⁰ <http://spokencorpora.ru/>

¹¹ <https://multidiscourse.ru/main/>

¹² <http://rsl.nstu.ru/site/index>

¹³ <https://www.elararchive.org/dk0339>

¹⁴ http://udmurt.web-corpora.net/tatyshly_corpus/search

¹⁵ http://lingconlab.ru/muira_dargwa/search

¹⁶ <http://videocorpora.ru/ru>

¹⁷ <http://web-corpora.net/bashcorpus/search/>

¹⁸ http://lingconlab.ru/spoken_khakas/search

Случаи с животными		2011		
Меең	абам	аңнаңаң	кізі	полған
мин п	аба п	аңнирға v	кізі п	поларға v
меең	аба-(Ы)м	аңна-ЧАң	кізі	пол-ГАН
я.Gen	отец-1pos.sg	охотиться-Hab	человек	быть-Past
☒ gen	1pos.sg	hab	человек	past
я	отец	охотиться		быть
Мой отец охотником был .				
мой	отец	охотник	быть	

Рис. 1. Пример морфологического разбора

го материала на различных языках и диалектах, на их основе лингвисты изучают отдельные аспекты языка, а также его вариативность в различных социокультурных и контекстуальных условиях. Эти исследования, в свою очередь, могут иметь широкий спектр дальнейшего применения.

Корпуса — достаточно мощный исследовательский инструмент. Во-первых, корпус предоставляет объективные данные о реальном использовании языка, что позволяет избежать субъективных оценок и выводов. Во-вторых, корпуса могут содержать достаточно большие объёмы языкового материала, что обеспечивает репрезентативность и достоверность исследования. В частности, достаточно большие корпуса записанных текстов повышают эффективность и качество работы лингвистов, в чьи задачи входит описание некоторого языка на всех его уровнях. Более того, авторы некоторых методических работ в области документационной лингвистики рассматривают грамматические описания, составленные с опорой на цифровые корпуса, как предпочтительные по сравнению с описаниями, построенными на традиционных методах работы с материалом (например, на методах интроспекции¹⁹ и элицитации²⁰) [8, 9]. Таким образом, корпус является удобным полигоном для проведения исследований по различным компонентам языка. Ниже мы представим примеры исследований, выполненных на материале корпусов по языкам России.

Достаточно репрезентативный корпус позволяет исследовать прежде всего лексику языка. Путём анализа больших объёмов текстового материала, собранного в корпусе, лингвисты могут изучать частот-

ность употребления слов, определять семантические поля и связи между различными лексическими единицами, а также исследовать изменения в лексике языка со временем. В качестве примера использования корпуса для изучения лексики можно привести работу С.Е. Бачаевой о лексической сочетаемости имён прилагательных, обозначающих малый размер, выполненную на материале Национального корпуса калмыцкого языка [10]. Обнаружено, что три таких калмыцких прилагательных — *уутъхн*, *ахр(хн)* и *бичкн* — разделяют лишь одно семантическое поле (части тела), а в других семантических полях употребляется только одна или две из исследуемых лексем.

Корпус звучащей речи позволяет проводить различные фонетические исследования. Примером является работа О.В. Ханиной об энецких передних гласных [11]. На корпусном материале было показано, что в энецком языке происходит фонетический процесс сужения открытого переднего /e/ в /i/, поэтому наблюдается вариативность в реализации [e]/[i] в ряде слов; на основании полученных данных был выработан фонологический принцип записи энецких слов.

Многие грамматические исследования также опираются на корпусные материалы. Так, в работе Н.В. Сердобольской убедительно показано, что корпусные исследования отдельных синтаксических феноменов (речь идёт о дифференцированном маркировании прямого дополнения в бесермянском удмуртском языке) позволяют значительно дополнить сведения, которые можно получить при направленном опросе носителей языка [12]. В частности, по данным анкетирования удаётся установить зависимость наличия маркирования от референциальных свойств и одушевлённости прямого объекта. Обращение к корпусным данным, в свою очередь, позволяет уточнить, что немаркированные одушевлённые объекты возможны для отдельных лексических классов существительных, а маркированные неодушевлённые именные группы наиболее часто используются в сочетании с определёнными моди-

¹⁹ Интроспекция (самонаблюдение) в лингвистике — это метод исследования, при котором лингвист в ходе исследования обращается к собственному сознанию как к источнику фактов и аргументов.

²⁰ Элицитация — метод получения данных о языке, при котором исследователь обращается к информанту (носителю языка) с просьбой перевести выражение с языка-посредника или оценить сконструированное исследователем выражение.

фикаторами. Корпусные данные сочетаются с данными, полученными методом элицитации, в работе А.А. Русских и С.А. Оскольской об аддитивной частице в тюркских языках Поволжья [13]. В данном исследовании были описаны функции употребления аддитивной частицы и продемонстрированы некоторые различия в её использовании в разных языках. Сочетание корпусных данных с данными элицитации позволило достичь более полного и точного понимания функциональных и семантических особенностей аддитивной частицы. Ещё один пример грамматического исследования, выполненного на корпусном материале одного из языков России, — работа Д.С. Ганенкова об инфинитивных синтаксических актах в лезгинском языке: показано, что конструкции с синтаксическими актами при разных матричных предикатах имеют разную синтаксическую структуру [14].

Корпуса играют важную роль в типологических исследованиях, так как не по всем необходимым для типологической выборки языкам могут быть доступны полные грамматические описания, и наличие корпусов по таким языкам может восполнить этот пробел. Кроме того, эффективным инструментом типологических исследований являются параллельные корпуса, которые позволяют выявлять грамматические средства, используемые в разных языках для выражения того или иного контекста (фрейма). Примером использования параллельных корпусов в типологическом исследовании является проект по созданию базы данных универсальных элементов грамматического набора на основе переводов Нового Завета (см. обзорную статью В.А. Плунгяна [15]) и 3 выпуск (том 19) журнала *Acta Linguistica Petropolitana* со статьями, выполненными в рамках этого проекта²¹. Параллельные корпуса (в том числе по некоторым языкам России) в рамках данного проекта позволили исследователям проанализировать и сравнить ряд грамматических структур в некоторых языках.

Различные направления изучения билингвизма (такие как переключение кодов между родным языком и русским, особенности контактно-обусловленных вариантов русского языка, используемых носителями миноритарных языков России, контактное влияние русского на другие языки России) могут быть успешно реализованы с помощью корпусов. Подобные исследования, в частности, проводит группа по изучению контактного взаимодействия русского языка с языками народов России в Институте русского языка им. В.В. Виноградова РАН²². Участниками исследовательской группы разработан корпус звучащей русской речи носителей автохтонных языков Севера, Сибири и Дальнего Востока, а также малый корпус устных текстов с единообразной разметкой переключения кодов в четырёх языках России: нанайском и уль-

чском (тунгусо-маньчжурские), горномарийском и мокшанском (уральские). На материале корпусов изучены частные аспекты грамматики: употребление союзов и рефлексивных местоимений в полипредикативных конструкциях, дифференцированное маркирование объекта, посессивные конструкции с препозицией генитива и др.

В настоящее время активно изучаются не только звуковые, но и жестовые языки, и русский жестовый язык (РЖЯ) не является в этом отношении исключением. В частности, в Новосибирском государственном техническом университете под руководством С.И. Бурковой разработан аннотированный онлайн-корпус²³, на базе которого проведены исследования структурных и функциональных особенностей редупликации [16, 17], средств выражения аспектуальности, модальности, коммуникативных значений, отношений обусловленности и др. в РЖЯ. Благодаря тому, что в корпус вошли записи текстов от носителей РЖЯ, проживающих в Новосибирске и Москве, корпус предоставляет возможность исследовать особенности нескольких региональных вариантов языка в рамках одного ресурса. Кроме того, многие тексты содержат так называемую калькирующую жестовую речь, то есть наборы жестов или их комбинаций, которые являются эквивалентами слов звукового русского языка и подчиняются его грамматическим нормам. Калькирующая речь рассматривается в РЖЯ как разновидность кодового переключения, что значительно расширяет представления о механизмах смешения кодов и билингвизма в целом, основанные прежде всего на звуковых языках.

Корпусные материалы могут быть использованы для создания других языковых ресурсов, в частности словарей. К примеру, в бесермянском словаре²⁴ к лексемам в качестве иллюстративных примеров автоматически добавляются предложения из корпуса. Развитые корпуса также позволяют создавать автоматические парсеры²⁵ для отдельных языков. Пример создания автоматического парсера для хакасского языка с использованием корпусной информации (в частности, о сочетаемости различных морфем) описан в работах А.В. Дыбо и коллег [18, 19]. Материалы звучащего корпуса могут стать основой для разработки программных средств синтеза речи, как это было сделано для татарского языка [20]. Данные, собранные для параллельного корпуса, позволяют создавать машинный переводчик. Например, включение лугово-восточного марийского и горномарийского языков в Яндекс.Переводчик стало возможным благодаря таким языковым материалам [21]. Перечисленные примеры свидетельствуют, что корпусные данные способ-

²³ <http://rsl.nstu.ru/>

²⁴ <http://beserman.ru/?page=dictionary>

²⁵ Парсер (от англ. *parser*) — синтаксический анализатор, программа или часть программы, выполняющей синтаксический анализ.

²¹ <https://alp.iling.spb.ru/issues/xix3.ru.html>

²² https://ruslang.ru/departments/contact_group

ствуют созданию разнообразных лингвистических инструментов и технологий, расширяя доступность и сферы использования языков.

На основе корпусных материалов можно также создавать контент для носителей языка и интересующихся им людей. К примеру, многие тексты, собранные исследовательницей Д. Форкер в рамках проекта по документации одного из даргинских диалектов (санжинского), были опубликованы в виде книги «Санжинские сказки и рассказы» [22]. Издание этой книги не только вносит вклад в сохранение культурных особенностей даргинского языкового сообщества, но и делает их доступными для более широкой аудитории.

Корпус можно использовать при обучении языку. Например, в работе В.З. Тулумбаева описан опыт использования корпуса башкирского языка, опубликованного на сайте Машинного фонда башкирского языка²⁶, в рамках студенческих занятий [23]. Кроме того, на базе платформы ЛингвоДок в данный момент разрабатывается образовательная платформа по языкам России. Материалы этой платформы основаны на корпусных и словарных данных, собранных и представленных на ЛингвоДоке. На этой платформе будут доступны обучающие упражнения на языках России, созданные на основе корпусов ЛингвоДока, а также словари.

Таким образом, корпусные ресурсы не только помогают в анализе языкового материала, но и вносят вклад в сохранение языкового разнообразия. Благодаря современным технологиям корпусные данные становятся доступными для широкого круга пользователей и могут применяться для создания других языковых ресурсов, таких как словари, парсеры, переводчики и др.

Организационные проблемы создания корпусов. Разработка языкового корпуса — важная, но весьма трудоёмкая задача, которая требует значительных временных и человеческих ресурсов. Необходимо взаимодействие лингвистов с программистами, отвечающими за техническую сторону создания корпуса. Сама по себе разработка подобного ресурса — довольно долгий многоступенчатый процесс: на первом этапе отбираются материалы для корпуса, которые затем размечаются лингвистом определённым образом (в зависимости от задач, для которых создаётся корпус). Затем размеченные материалы загружаются программистом на специализированную платформу (платформа может представлять собой уже готовую программу либо разрабатываться под нужды конкретного проекта совместно с его участниками-лингвистами).

Существенно различается процесс подготовки устных и письменных корпусов. Для записи устной речи необходимо соответствующее техническое обеспечение — аудио- и/или видеоаппаратура, а также

подходящие условия для записи (желательно минимизировать посторонние шумы, отвлекающие факторы). Записанная устная речь требует расшифровки с участием носителя языка и последующей вычитки лингвистом на предмет опечаток или ошибок в разборе. В настоящее время сформулированы весьма конкретные рекомендации по созданию дискурсивной транскрипции, включая разметку дискурсивной структуры, просодии, пауз, речевых сбоев и т.д. [24]. Хочется верить, что будущие устные корпуса будут следовать этим рекомендациям, что значительно повысит их валидность и достоверность.

В случае уже готовых письменных текстов первый этап работы требует несколько меньше времени, чем при подготовке корпуса устных текстов. Однако при отборе письменных текстов разработчик обычно также решает непростые задачи, связанные с полнотой и репрезентативностью созданного корпуса. Подразумевается, что текстовая выборка должна быть определённым образом сбалансирована с точки зрения различных металингвистических параметров (год создания текста, авторство, жанр и т.п.) [25].

На этапе добавления разметки возможно до некоторой степени облегчить работу лингвиста, поскольку для ряда языков существуют качественные морфологические парсеры (ср. пример хакасского выше), а также встроенные морфологические анализаторы в программах для работы с массивами текстов (таких как Toolbox или FieldWorks), для которых можно прописать правила разметки. Безусловно, качество автоматической разметки во многом зависит от характера связи между морфемами в конкретном языке: так, наличие нетривиальных чередований на стыках морфем может создавать препятствия для разметки с помощью морфологического анализатора. В любом случае, финальный шаг предполагает проверку размеченных данных вручную, исправление программных ошибок и снятие омонимии в неоднозначных случаях. Для малоописанных языков этот процесс занимает много времени, поскольку требует полного погружения лингвиста в грамматику и лексику языка, для которого создаётся корпус, и глубокого понимания системы правил, по которым строятся высказывания на этом языке.

Наконец, после загрузки текстов на электронный ресурс впоследствии необходимо регулярно поддерживать этот ресурс, своевременно устранять технические неполадки, отслеживать состояние сервера и пр. Решение этих задач, а также работа лингвиста-специалиста по языку и его консультантов (носителей языка, участвующих в записи, расшифровке и переводе текстов) требует финансирования. Сейчас финансирование науки ориентировано на рейтинговые публикации, однако публикационная активность на основе данных корпуса возможна лишь спустя несколько лет после его разработки. В течение собственно работы над корпусом полу-

²⁶ <http://mfbl2.ru/>

чить финансовую поддержку крайне затруднительно. Большим подспорьем в решении этой проблемы могли бы стать специализированные гранты, которые должны выдаваться с ориентиром на качество и объём корпусного ресурса и не зависеть (по крайней мере на этапе разработки) от количества опубликованных журнальных статей.

Подводя итоги, следует подчеркнуть, что корпусные ресурсы важны для изучения и сохранения языкового разнообразия России, то есть помогают решать важную социальную задачу. Кроме того, они являются ценным инструментом лингвистических исследований и создания других языковых ресурсов.

БЛАГОДАРНОСТИ

Авторы благодарят Т.А. Архангельского, А.А. Головину, И.А. Грунтова, А.В. Дыбо, Т.А. Майсака, В.С. Мальцеву, Ю.В. Норманскую, В.А. Плунгяна, А.В. Савельева, А.С. Сметину, Н.Р. Сумбатову, А.А. Сюрюн, В.С. Харитонову, Э.В. Хилханову за предоставленные сведения о различных корпусных ресурсах по языкам России.

ЛИТЕРАТУРА

1. *Коряков Ю.Б., Давидюк Т.И., Харитонов В.С., Евстигнеева А.П., Сюрюн А.А.* Список языков России и статусы их витальности. Монография-препринт. М.: Институт языкознания РАН, 2023. [http://jazykirf.iling-ran.ru/\(2023\)_Spisok_jazykov_Rossii_Monograph.pdf](http://jazykirf.iling-ran.ru/(2023)_Spisok_jazykov_Rossii_Monograph.pdf) (дата обращения 25.05.2024).
2. *Koryakov Yu.B., Davidyuk T.I., Haritonov V.S., Evstigneeva A.P., Syuryun A.A.* A list of languages of Russia and their vitality statuses. Preprint. Moscow: Institute of Linguistics RAS, 2023. [http://jazykirf.iling-ran.ru/\(2023\)_Spisok_jazykov_Rossii_Monograph.pdf](http://jazykirf.iling-ran.ru/(2023)_Spisok_jazykov_Rossii_Monograph.pdf) (accessed 25.05.2024).
3. *The Routledge handbook of corpus linguistics* / Ed. by A. O'Keeffe, M.J. McCarthy. Abingdon, New York: Routledge, 2021.
4. *Kibrik A.A.* A program for the preservation and revitalization of the languages of Russia // *Russian Journal of Linguistics*. 2021, vol. 25, no. 2, pp. 507–527.
5. *Языковое многообразие России и возможности его сохранения* / Ред. Е.Ю. Груздева, А.А. Сюрюн. Препринт. М.: Институт языкознания РАН, 2023. https://iling-ran.ru/library/revitalization/gruzdeva_et_al_language_diversity_2023.pdf (дата обращения 25.05.2024).
6. *Linguistic diversity of Russia and opportunities for its preservation* / Ed. by E.Yu. Gruzdeva, A.A. Syuryun. Preprint. Moscow: Institute of Linguistics, Russian Academy of Sciences, 2023. https://iling-ran.ru/library/revitalization/gruzdeva_et_al_language_diversity_2023.pdf (accessed 25.05.2024).
7. *Gatbonton E., Pelczar I., Cook C., Venkatesh V., Nochasak C., Andersen H.* A pedagogical corpus to support a language teaching curriculum to revitalize an endangered language: the case of Labrador Inuttitut // *International Journal of Computer-Assisted Language Learning and Teaching*. 2015, no. 5(4), pp. 16–36.
8. *Сичинава Д.В.* Параллельные тексты в составе Национального корпуса русского языка: новые языки и новые задачи // *Труды Института русского языка им. В.В. Виноградова*. 2019. № 21. С. 41–60.
9. *Sichinava D.V.* On parallel texts within the Russian National Corpus: new languages and new challenges // *Proceedings of the V.V. Vinogradov Russian Language Institute*. 2019, no. 21, pp. 41–60.
10. *Архангельский Т.А.* Корпусная платформа Tsakorpus и языки России // *Электронная письменность народов Российской Федерации – 2021 и IWCLUL 2021. Материалы Международной научно-практической конференции, Сыктывкар, 23–24 сентября 2021 года. Сыктывкар: Коми республиканская академия государственной службы и управления, 2022. С. 23–24.*
11. *Arkhangelsky T.A.* The corpus platform Tsakorpus and the languages of Russia // *Electronic Writing Systems of the Peoples of the Russian Federation – 2021 and IWCLUL 2021. Proceedings of the International Scientific and Practical Conference, Syktyvkar, September 23–24, 2021. Syktyvkar: Komi Republic Academy of Public Administration and Management, 2022. P. 23–24.*
12. *Bright W.* Contextualizing a grammar // *Perspectives on grammar writing* / Ed. by Th. Payne, D. Weber. Amsterdam: John Benjamins, 2007. P. 11–17.
13. *Mosel U.* Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language // *The Art and Practice of Grammar Writing (LD&C Special Publication 8)* / Ed. by T. Nakayama, K. Rice. 2014. P. 135–157.
14. *Бачаева С.Е.* Лексическая сочетаемость имён прилагательных, обозначающих малый размер (на материале Национального корпуса калмыцкого языка) // *Известия ДГПУ*. 2016. Т. 10. № 4. С. 42–47.
15. *Bachayeva S.E.* Lexical collocations of adjectives denoting the small size (based on the materials of the National Corpus of the Kalmyk language) // *DSPU Journal*. 2016, vol. 10, no. 4, pp. 42–47.
16. *Ханина О.В.* Возможности цифровых технологий: описание аллофонов передних гласных, гортанного смычного и объектного согласования глагола в энецком языке // *Урало-алтайские исследования*. 2017. № 3(26). С. 186–207.
17. *Khanina O.V.* Advantages of digital technologies: a description of front vowels allophones, of a glottal stop, and of verbal object cross-reference in Enets // *Ural-Altaic Studies*. 2017, no. 3(26), pp. 186–207.
18. *Serdobolskaya N.* A corpus analysis of differential object marking in Beserman Udmurt // *Linguistica Uralica*. 2020, vol. 56, no. 4, pp. 275–308.

13. Русских А.А., Оскольская С.А. Аддитивная частица в тюркских языках Поволжья // *Oriental Studies*. 2021. Т. 14. № 6. С. 1324–1352.
Russkih A.A., Oskolskaya S.A. Additive particle in Turkic Languages of the Volga-Kama Sprachbund // *Oriental Studies*. 2021, vol. 14, no. 6, pp. 1324–1352.
14. Ганенков Д.С. Корпусное исследование инфинитивных конструкций в лезгинском языке // *Acta Linguistica Petropolitana. Труды Института лингвистических исследований*. 2016. Т. 12. Ч. 1. С. 310–322.
Ganenkov D.S. A corpus-based study of infinitive constructions in Lezgian // *Acta Linguistica Petropolitana. Transactions of the Institute for Linguistic Studies*. 2016, vol. 12, part 1, pp. 310–322.
15. Плунгян В.А. Параллельный корпус как грамматическая база данных и Новый Завет как параллельный корпус (предисловие) // *Acta Linguistica Petropolitana. Труды Института лингвистических исследований*. 2023. Т. 19. Ч. 3. С. 15–38.
Plungian V.A. The parallel corpus as a grammar database and the New Testament as a parallel corpus (Preface) // *Acta Linguistica Petropolitana. Transactions of the Institute for Linguistic Studies*. 2023, vol. 19, part 3, pp. 15–38.
16. Буркова С.И., Филимонова Е.В. Редупликация в русском жестовом языке // *Русский язык в научном освещении*. 2014. № 2(28). С. 202–258.
Burkova S.I., Filimonova E.V. Reduplication in Russian sign language // *Russian Language and Linguistic Theory*. 2014, no. 2(28), pp. 202–258.
17. Буркова С.И. Способы выражения именной множественности в русском жестовом языке // *Сибирский филологический журнал*. 2015. № 2. С. 174–184.
Burkova S.I. The ways of expressing nominal plurality in the Russian sign language // *Siberian Journal of Philology*. 2015, no. 2, pp. 174–184.
18. Дыбо А.В., Крылов Ф.С., Мальцева В.С., Шеймович А.В. Сегментные правила в автоматическом парсере Корпуса хакасского языка // *Урало-алтайские исследования*. 2019. № 1(32). С. 48–69.
Dybo A.V., Krylov Ph.S., Maltseva V.S., Sheimovich A.V. Segmental rules in the automatic parser for the Khakas corpus // *Ural-Altaic Studies*. 2019, no. 1(32), pp. 48–69.
19. Дыбо А.В., Мальцева В.С., Султрекова Э.В., Шеймович А.В., Крылов Ф.С. Структура хакасской словоформы и ограничения на сочетаемость аффиксов в автоматическом парсере хакасского языка // *Урало-алтайские исследования*. 2023. № 2(49). С. 42–75.
Dybo A.V., Maltseva V.S., Sultrekova E.V., Sheimovich A.V., Krylov Ph.S. The structure of the Khakas word form and restrictions on the compatibility of affixes in the automatic parser for the Khakas language // *Ural-Altaic Studies*. 2023, no. 2(49), pp. 42–75.
20. Хусаинов А.Ф., Сулейманов Д.Ш. Обзор созданных речевых корпусов и программных средств для синтеза татарской речи // *Речевые технологии*. 2020. № 1. С. 63–72.
Khusainov A.F., Suleymanov D.Sh. Overview of speech corpora and software for the Tatar speech synthesis // *Speech Technology*. 2020, no. 1, pp. 63–72.
21. Сабанцев Г.Л., Чемышев А.В. Яндекс.Переводчик и языки России // *Электронная письменность народов Российской Федерации - 2021 и IWCLUL 2021. Материалы Международной научно-практической конференции, Сыктывкар, 23–24 сентября 2021 года. Сыктывкар: Коми республиканская академия государственной службы и управления, 2022. С. 178–181.*
Sabantsev G.L., Chemyshev A.V. Yandex.Translate and the languages of Russia // *Electronic Writing Systems of the Peoples of the Russian Federation – 2021 and IWCLUL 2021. Proceedings of the International Scientific and Practical Conference, Syktvykar, September 23–24, 2021. Syktvykar: Komi Republic Academy of Public Administration and Management, 2022. P. 178–181.*
22. Форкер Д., Гаджимурадов Г.А. Санжинские сказки и рассказы. С приложением санжинско-русского и русско-санжинского словарей. Махачкала: Типография А4, 2017.
Forker D., Gadzhimuradov G.A. Sanzhi tales and stories. With Sanzhi-Russian and Russian-Sanzhi dictionaries. Makhachkala: A4 Printing House, 2017.
23. Тулумбаев В.З. Технологии корпусной лингвистики в обучении башкирскому языку // *Современные проблемы и перспективы развития естествознания. Материалы национальной научно-практической конференции. Уфа, 8–9 июня 2020 года. Уфа: Башкирский государственный педагогический университет им. М. Акмуллы, 2020. С. 309–312.*
Tulumbaev V.Z. Corpus linguistics technologies in teaching Bashkir // *Modern Problems and Prospects of Natural Sciences Development. Proceedings of a National Scientific and Practical Conference. Ufa, June 8–9, 2020. Ufa: Bashkir State Pedagogical University named after M. Akmulla, 2020. P. 309–312.*
24. Кибрик А.А., Майсак Т.А. Правила дискурсивной транскрипции для описательных и документационных исследований // *Рема*. 2021. № 2. С. 23–45.
Kibrik A.A., Maisak T.A. Discourse transcription rules for descriptive and documentary studies // *Rhema*. 2021, no. 2, pp. 23–45.
25. Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСС, 2001.
Baranov A.N. Introduction to applied linguistics. Moscow: Editorial URSS, 2001.

CORPORA AND CORPUS-BASED STUDIES OF THE LANGUAGES OF THE RUSSIAN FEDERATION

T.I. Davidyuk^{a,b,*}, A.A. Kibrik^{a,b,**}, D.D. Mordashova^{a,***}

^a*Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russia*

^b*Lomonosov Moscow State University, Moscow, Russia*

*E-mail: davidyuk@iling-ran.ru

**E-mail: aakibrik@iling-ran.ru

***E-mail: d.mordashova@iling-ran.ru

The article describes corpus resources for the languages of Russia and their use in linguistic research. The linguistic diversity of the country is quite substantial: currently 155 languages are identified as languages of Russia. Many of them are under threat of extinction, which makes the task of creating corpora particularly relevant as a tool for language preservation. In this study we conducted a survey among the staff of the Institute of Linguistics of the Russian Academy of Sciences and other colleagues, which helped us collect the data about 73 corpus resources representing various languages and dialects of Russia. The sample covers both major languages and languages with relatively few speakers, including unwritten languages. The article examines various parameters by which corpora may differ, and offers examples of research based on materials from the corpora. The final part of the article discusses the organizational aspects of creating and maintaining corpus resources. The results of the study suggest that corpus resources not only play an important role in preserving the linguistic diversity of Russia but also represent a valuable tool for various research tasks, as well as for creating other language resources.

Keywords: linguistic corpora, language resources, languages of Russia, language preservation.