[8] **Кудрявцева Р. А.** Периодизация литератур финно-угорских народов в контексте сравнительной филологии // Litera. 2017. № 1. С. 149 — 155. URL: https://nbpublish.com/library_read_article.php?id=22415 (дата обращения: 11.11.2021).
[9] См.: **Каторова А. М.** Мордовская литература как явление культуры в современном финно-угорском мире // Центр и периферия. 2020. № 3. С. 42 — 47.
[10] **Глухова Н. Н., Глухов В. А.** Указ. соч.
[11] Пермские литературы…
[12] Большая литература малых народов…
[13] **Кубанцев Т. И.** Указ. соч. С. 233.
[14] См.: **Кудрявцева Р. А.** Указ. соч. С. 150.
[15] Там же.
[16] См.: **Каторова А. М.** Указ. соч.
[17] См.: Мокша. 2000. № 6. С. 38, 39, 41 ; Сятко. 2000. № 6. С. 55, 57.
[18] См.: Сятко. 2000. № 6. С. 31 — 62.
[19] См.: Сятко. 2001. № 9. С. 76 — 85.

Поступила 30.03.2022 г.

УДК 811.511.152

О. Г. Борисова, А. В. Чернов
O. G. Borisova, A. V. Chernov

ОБЗОР МОРДОВСКИХ ЯЗЫКОВЫХ КОРПУСОВ

OVERVIEW OF MORDOVIAN LANGUAGE CORPORA

Ключевые слова: мокшанский язык, эрзянский язык, цифровизация, языковой корпус.
В статье дается обзор лингвистических платформ, содержащих корпуса эрзянского и мокшанского языков, а также определяется социальная значимость цифровизации материалов на мордовских языках.

Key words: the Moksha language, the Erzya language, digitalization, language corpus.
The article provides an overview of linguistic platforms containing the corpus of the Erzya and Moksha languages, as well as the social significance of digitalization of materials in the Mordovian languages are determined.

Стремительное развитие в последние два десятилетия отечественной корпусной лингвистики ставит и перед так называемыми малыми языками новые задачи. Цифровой формат языков знаменует следующий этап в их развитии. Современные цифровые лингвистические технологии — это не только мощный инструмент исследования языков и широкие возможности форсирования многих практических и трудоемких задач, например, составления словарей, но и сохранение функциональности, престижа и, наконец, витальности языков, находящихся под угрозой исчезновения. Говоря о функциях и социальной важности цифровизации материалов языков этнических меньшинств, ученые, в частности, прогнозируют, что в Сети будут представлены и доступны только те языки, для которых разработаны адекватные языковые ресурсы, продукты и системы. В ином случае языки, для которых

лингвистические технологии не будут должным образом развиты, рискуют утратить статус средств коммуникации в Сети, что серьезно угрожает языковому и культурному многообразию[1]. Существуют и более критические прогнозы относительно цифровой жизнеспособности уральских языков[2].

В целях дальнейшего развития, а также популяризации мордовских (мокшанского и эрзянского) языковых корпусов требуется детальный обзор и анализ уже имеющихся. Языковой корпус — собрание текстов в цифровом формате на том или ином языке, предназначенное для его изучения. По эрзянскому и мокшанскому языкам в Сети существуют несколько крупных ресурсных платформ, содержащих разнотипные корпуса эрзянского и мокшанского языков. Среди них: Kielipankki (Языковой банк) с сервером Korp (Финляндия) и тесно связанная с ней платформа Giellatekno (GiellaLT) Арктического университета (Норвегия), языковой портал UTU-Digilang Туркуского университета (Финляндия), ресурсы «Корпуса уральских языков Поволжья» исследователя Т. А. Архангельского и «ЛингвоДок» РАН.

Одна из лингвистических платформ, широко представляющая корпуса эрзянского и мокшанского языков, — современный ресурс Kielipankki=Language Bank of Finland (Языковой банк), поддерживаемый консорциумом FIN-CLAR (включает в себя CSC — ИТ-компанию, Институт языков Финляндии и группу финских университетов). Языковой банк предлагает многочисленные фонды лингвистических материалов в цифровом формате, например, по эрзянскому языку материал находится на 16 ресурсах, по мокшанскому — на 13[3]. Пользователями платформы являются исследователи, преподаватели и студенты, интересующиеся материалами, содержащими текст, речь или видео на естественном языке, а также инструментами для их обработки. В основном с фондами Языкового банка работают лингвисты, но сервисы также подходят для исследований и в других гуманитарных областях. Ресурсы Языкового банка имеют 3 режима доступа: доступен для он-лайн поиска, по личному запросу и для зарегистрированных пользователей, что не всегда удобно для исследователей не из европейских университетов. Основной сервер, где располагаются цифровые материалы, — Korp, имеющий шведскую[4] (здесь он разработан), норвежскую[5] и финскую[6] версии.

Материал по мордовским языкам размещен на финском сервере Korp.csc.fi в подразделе Muut kielet (Другие языки)[7], а также на норвежском[8]. В открытом доступе находятся 4 подкорпуса, или коллекции: ERME, Fenno-Ugrica, SUS-kenttätyö и Wanca 2016. По статистическим данным (охвату предложений и слов) на сервере Korp.csc.fi коллекции по эрзянскому и мокшанскому языкам занимают вторую и третью позиции (после марийского) среди 22 уральских языков и содержат: эрзянская — 201 104 предложения и 2 042 388 слов, мокшанская — соответственно 142 248 и 1 673 480[9]. Рассмотрим их подробнее.

Расширенный корпус эрзянского и мокшанского языков (ERME)=Ersän ja mokšan laajennettu korpus=Erzya and Moksha Extended Corpora, Korp. версия (erme-s-korp)[10] — многоязычный (мокшанский, эрзянский, финский и английский) текстовый корпус, содержащий преимущественно материал из эрзянской и мокшанской литературы, а также из СМИ. Временная соотнесенность текстов корпуса — XIX — XX вв. Основным используемым форматом является XML. Как указывается в аннотации, цель проекта состоит в том, чтобы создать корпуса с детализацией на уровне слов. На уровне предложения используется контекстный перевод (англий-

ский или финский), на уровне слова — морфологическая разметка, соответствующая каждому контексту. Морфологический анализ проводится на основе HFST (Helsinki Finite-State Technology — библиотеки и набора утилит для обработки естественного языка с конечными автоматами). Грамматический анализ и маркировка подобны методам, которые были разработаны в Giellatekno — Центре лингвистической технологии для саамских языков при университете Тромсё (Арктический университет, Норвегия). Данные методики применяются при документации и других уральских языков. Объем обработанного материала в корпусе составляет 155 100 предложений (эрзянских — 81 824, мокшанских — 73 276) и более 1,5 млн слов (эрзянских — 781 958, мокшанских — 797 852). Норвежская версия Korp (Giellatekno/Divvun) ERME[11] содержит тексты из эрзянской Википедии (46 041 предложение и 498 157 токенов), мокшанской (соответственно 8 420 и 93 654), из эрзянской художественной литературы (57 139 и 468 689) и эрзянских (1 311 742 и 13 083 739) и мокшанских (12 699 и 1 274 181) новостных статей. В дальнейшем количество обрабатываемого материала будет увеличено. Авторы корпуса: Д. Рютер и О. Ерина. ERME доступен для он-лайн поиска.

Следующий ресурс, расположенный на сервере Korp.csc.fi и содержащий мордовские языковые корпуса, — Fenno-Ugrica — цифровая коллекция финно-угорских публикаций Национальной библиотеки Финляндии[12]. В нее входят материалы 10 языков, в том числе монографии на эрзянском, мокшанском, ингерманландском, вепсском, марийском (горно-марийский и луговой), а также газеты с 1920-х по 1939 г. на мордовских, марийском и других языках. Всего в коллекции более 120 монографий и около 20 тыс. страниц периодических изданий. Многоязычная цифровая коллекция Fenno-Ugrica содержит 1 022 767 предложений и 8 654 289 слов, из них эрзянских — соответственно 88 617 и 899 106, мокшанских — 47 131 и 617 934. В подготовке материалов Fenno-Ugrica участвовала Российская национальная библиотека (г. Санкт-Петербург), где публикации были переведены в электронный формат в рамках проекта оцифровки материалов родственных языков, который являлся частью Языковой программы финского Фонда Kone. В ходе оцифровки Национальная библиотека Финляндии разработала редактор на основе OCR с открытым исходным кодом, который позволяет редактировать машинно-кодированный текст для лингвистических исследований. Ресурс доступен для он-лайн поиска.

Корпус полевых исследований Финно-угорского общества=SUS-kenttätyö=SUS-fieldwork[13] — многоязычный ресурс (пока 3 корпуса: эрзянский, мокшанский и коми-зырянский), основанный на опубликованных Финно-угорским обществом текстах, в основном из научной серии «Труды Финно-угорского общества (Suomalais-Ugrilaisen Seuran Toimituksia=SUST)», предназначен для размещения транскрибированных текстов, содержащих один или несколько параллельных переводов. Переписанные тексты имеют нормализованные варианты поиска. В метаданных корпуса можно найти информацию о публикации. Ядром мордовской части корпуса является труд Mordwinische Volksdichtung («Мордовская народная поэзия») Хейкки Паасонена в 8 томах. Корпус будет доступен для он-лайн поиска.

С 2013 по 2019 г. в университете г. Хельсинки на факультете цифровых гуманитарных наук под руководством К. Линдена велась работа над исследовательским проектом «Финно-угорские языки и Интернет»[14], целью которого являлось создание системы с автоматическим определителем языков для поиска в Сети текстов, напи-

санных на миноритарных языках уральской группы. Проект финансировался Фондом Kone и поддерживался Национальной библиотекой Финляндии. Одной из основных задач являлся сбор материалов на малых уральских языках из ресурсов Интернета. Исследовательская группа в соответствии с языковым справочником «Ethnologue: Languages of the World» определила область поиска — 38 языков. В ходе интенсивного сбора веб-страниц (сотни миллионов) и их загрузки в корпус языки прошли двукратную идентификацию, в том числе и через разработанный Т. Яухиайненом идентификатор языка, который способен различать около 400 языков и диалектов[15]. Для поиска материалов, написанных на уральских языках, исследователи использовали Heritrix — поисковый робот для веб-архивирования с открытым исходным кодом. Результатом проекта стала платформа, или портал, Wanca (от прауральского ‘корень’)[16]. Wanca — это, с одной стороны, ссылки на веб-страницы, написанные на разных уральских языках. На данный момент портал имеет 57 491 ссылку на 919 сайтов, содержащих тексты на 36 малых уральских языках. В эрзянской части даны ссылки на 26 сайтов[17], в мокшанской — на 20[18]. Содержание многих веб-сайтов сегодня очень динамично. Так, многие ссылки в Wanca ведут на новостные сайты, структура которых регулярно обновляется и меняется, а иногда по разным причинам они полностью исчезают из Интернета. Портал Wanca способен сохранить некоторые устаревшие страницы, однако отдельные ссылки в Wanca также устаревают[19]. Напомним, что портал функционирует в режиме бета-версии, поэтому некоторые ссылки еще требуют уточнения. Другая составляющая проекта — создание открытых корпусов предложений и слов из электронных текстов для 29 уральских языков. Как подчеркивают составители, предоставление исследователям новых корпусов позволит провести актуальные лингвистические исследования и будет способствовать их развитию. Корпуса и их качество можно автоматически совершенствовать с помощью существующих языковых инструментов. Списки слов, составленные из корпусов, также можно использовать для лексикографической работы[20]. Эрзянская коллекция содержит 28 986 предложений, 294 065 слов и 4 231 531 символ, мокшанская — соответственно 21 571, 214 052 и 3 101 193. Корпус Wanca 2016 также разместился на сервере Korp.csc.fi в открытом доступе[21].

В режиме закрытого доступа, требующего личной регистрации, в Языковом банке на сервере Korp.csc.fi находятся подкорпуса, которые ранее располагались на языковом сервере Хельсинкского университета Multilingual Resource Collection of the University of Helsinki Language Corpus Server (UHLCS)[22]. Коллекция UHLCS содержит машиночитаемые лингвистические данные и базовые службы, инструменты для их использования. Языковой сервер был основан в конце 1980-х гг., его ранний вариант состоял из финских, английских и шведских корпусов. В настоящее время он содержит компьютерные корпуса из более чем 50 языков, включая образцы как миноритарных языков, так и обширные корпуса, представляющие различные типы текстов. В 2000 г. корпуса уральских, тюркских, тунгусских, монгольских, чукотско-камчатских, иранских и северо-восточных кавказских языков были отредактированы для публичного использования при финансовой поддержке Института эволюционной антропологии Макса Планка в г. Лейпциге. Операционная система UNIX составляет основу UHLCS[23].

В корпусе уральских, тюркских, индоиранских, монгольских языков, а также языков Сибири и Кавказа=Uralilaisia, turkkilaisia, indo-iranilaisia ja mongolikieliä sekä

Siperian ja Kaukasian kieliä (UHLCS), Helsinki-Korp-versio=Uralic, Turkic, Indo-Iranian and Mongol languages; languages of Siberia and Caucasia (UHLCS), Helsinki Korp Version имеется текстовый и словарный материал по мордовским языкам[24].

Корпус эрзянской и мокшанской литературы и журналов и коми-зырянской литературы=Ersän ja mokšan kirjallisuutta ja julkaisuja ja komisyrjäänin kirjallisuutta (UHLCS)=Corpus of Erzya and Moksha Mordvin Literature and Journals and Komi Zyrian Literature (UHLCS) имеет несколько версий, в том числе Helsinki-Korp[25]. Мордовская часть корпуса составлена на основе оригинальных произведений писателей К. Г. Абрамова, А. В. Арапова, М. И. Брыжинского, И. А. Калинкина, П. А. Ключагина, А. Д. Куторкина, А. Мокшони, А. П. Тяпаева, А. М. Шаронова и др.

Кроме того, на языковом сервере UHLCS располагается Корпус эрзянских и мокшанских слов=Ersän ja mokšan sanaluettelokorpus (UHLCS), Helsinki-Korp-versio=Erzya and Moksha Mordvin Word List Corpus (UHLCS), Helsinki Korp Version, созданный на основе труда «Русско-мордовский словарь. Из истории отечественной лексикографии» А. П. Феоктистова (1971). Лексикографический материал представлен в основном эрзянскими единицами — около 23 500 и мокшанскими — 300. Напомним, это исторический материал, составленный в 1785 г. епископом Дамаскиным по указанию Екатерины II. Список слов для корпуса был подготовлен Д. Эстиллом[26].

Корпус Параллельные библейские стихи для уральских исследований= Raamatun jakeita uralilaisille kielille, rinnakkaiskorpus, Korp=Parallel Bible Verses for Uralic Studies, Korp содержит 27 библейских (исторических и современных, 1821 — 2019 гг.) текстов на эрзянском, мокшанском, вепсском, олонецко-карельском (ливви), двино-карельском (северокарельский), удмуртском, коми-пермяцком, коми-зырянском, мансийском, хантыйском, финском и русском языках[27]. В их числе переводы «Нового Завета» и «Евангелия от Марка» на эрзянском и мокшанском языках, «Детской Библии» на эрзянском и книги «Иисус друг детей» на мокшанском, «Евангелия от Луки и Деяния Апостолов» и «Евангелия от Матфея» на эрзянском языке[28]. Целью параллельных корпусов является дальнейшее изучение процесса перевода на уральских языках. Одновременно параллельный корпус дает возможность отслеживать изменения в лексических и синтаксических стратегиях, используемых в разных версиях библейских стихов на родном языке, сравнивать лексику и структуру между языками. Лемматизация и морфологический анализ предусмотрены для всех языков, кроме двино-карельского, хантыйского, вепсского и русского. Финские тексты проанализированы с помощью программы TNPP (Turku Neural Parser Pipeline), которая включает лемматизацию, морфологический анализ, а также синтаксическую аннотацию.

Следующий подкорпус, в котором имеется эрзянский материал, но пока отсутствует мокшанский, — Количественные показатели и количественная оценка на финском языке и на языках Центрального Волго-Камского региона=Kvantifiointi suomessa ja keskisen Volgan ja Kaman alueella puhuttavissa kielissä (UHLCS), Helsinki-Korp-versio= Quantifiers and Quantification in Finnish and Languages Spoken in the Central Volga-Kama Region (UHLCS), Helsinki Korp Version содержит параллельные примеры из квантификаторов и количественного анализа на русском, финском, эрзянском, удмуртском и татарском языках, с их толкованием и переводом на английский язык. Ресурс содержит 8 000 слов[29].

На сервере Korp.csc.fi размещается подкорпус Разговорного эрзянского языка= Corpus of Colloquial Erzya [speech corpus] (длительностью 2 часа)[30], содержащий три

записанных интервью из с. Косогоры Большеберезниковского района и один образец из с. Мокшалей Чамзинского района. Интервью проводились в 1991 г. Материал корпуса собирали и аннотировали Р. Грюнталь, Д. Рютер, О. Ерина.

Лексические корпуса Эрзянско-финско-эрзянские словари Neahttadigisánit[31] и Мокшанско-финско-мокшанские словари Neahttadigisánit[32], имеющие свободный доступ, разработаны в Центре лингвистических технологий Giellatekno. Giellatekno — это одновременно и многоязычный ресурс, содержащий в том числе и двуязычные словари (более 15 языковых пар). Корпус предоставляет возможности перевода мордовских слов не только на финский, но и русский, английский, эстонский, немецкий и французский языки и в обратном направлении. Кроме перевода и морфологического анализа слова, здесь же даются его контекстные иллюстрации и переходы в текстовый корпус Korp на страницу простого поиска с конкордансом заданной лексемы и в этимологическую базу данных Álgu-tietokanta, которая составлена для саамских языков, но содержит, например, 443 мордовские этимологии. Мордовский подкорпус ресурса основан на материале многих словарей и художественной литературы XX и XXI вв. Однако первоначальной базой явились Ersäläis-suomalainen sanakirja=Эрзянь-финнэнь валкс Я. Ниеми, М. В. Мосина (1995) и Mokšalais-suomalainen sanakirja = Мокшень-финнонь валкс Е. Херралы, А. П. Феоктистова (1998).

В 2019 г. Институт эстонского языка выпустил первый эстонско-эрзянский словарь в электронном формате[33]. Для изучения фонетических особенностей эрзянского языка эстонскими специалистами подготовлен Эрзянский просодический корпус в аудиоформате, доступ к материалам ограничен[34].

В Туркуском университете работы по созданию финно-угорских языковых корпусов начались в конце 1970-х гг. Во второй половине 1990-х гг. активизировалась деятельность по составлению корпусов языков Волго-Камского ареала[35]. На сегодняшний день в университете разработан крупный языковой портал UTU-Digilang, содержащий 37 языковых корпусов[36], из которых 11 — финно-угорские[37]. Мордовских корпусов на портале 6. Доступ ко всем финно-угорским ресурсам требует регистрации.

Корпус Мормула : грамматически аннотированные мордовские тексты (эрзя, мокша)=Mormula : Grammatically annotated Mordvin texts (Erzya, Moksha)[38] был первым ресурсом по финно-угорским языкам, созданным еще в конце 1970-х гг. Содержит как литературные, так и диалектные тексты на эрзянском и мокшанском языках. Размер всего корпуса составляет 244 368 слов. Эрзянская и мокшанская части содержат соответственно 129 535 и 114 833 слова. Корпус имеет частеречную разметку и созданную вручную морфологическую аннотацию, куда входят 64 тега pos и 152 морфологических. Добавление синтаксических тегов в корпус находится в стадии разработки. Тексты снабжены переводом на немецкий или финский языки, расположены в 9 файлах (5 эрзянских и 4 мокшанских). В корпус вошли труды Х. Паасонена и П. Равилы, тексты из серии УПТМН, журнала «Сятко» и др. Мормула уже на протяжении 40 лет является ключевым научным ресурсом для изучения лексики, морфологии и синтаксиса мордовских языков зарубежными исследователями[39].

Диахронический корпус мордовского литературного языка=Diachronic Corpus of Literary Mordvin[40] состоит из газетных статей, охватывающих разные периоды развития мордовских литературных языков. Самые ранние тексты относятся к 1920 г., а новые — 2008 г. Статьи располагаются по периодам: 1920 — 1937 гг., 1938 — 1950 гг., 1960 — 1970 гг. и 2000-е гг., классифицируются по содержанию: политика и обще-

ство, экономика, культура и образование, художественная литература. В корпусе представлено 516 текстов, из них 281 — на эрзянском языке и 235 — на мокшанском. Общее количество слов составляет 336 000, 187 000 — на эрзянском и 149 000 — на мокшанском. По корпусу можно изучать изменения в мордовских литературных языках, влияние языковой политики на их развитие.

Корпус МокшЕр=MokshEr Corpus (Moksha, Erzya)[42] содержит собрание газетных («Эрзянь правда», «Эрзянь мастор», «Мокшень правда») и журнальных («Сятко», «Мокша») статей за 2002 — 2009 гг., а также несколько художественных произведений. Тексты даны без аннотаций. Эрзянская часть корпуса состоит из 2 991 текста, включающих около 2 785 000 токенов, мокшанская — 1 300 текстов и около 1 742 000 токенов.

Многоязычный корпус «Финляндия — прошлое и настоящее» (параллельные тексты)='Finland — Past and Present' Corpus (parallel texts)[43] составлен на основе материалов перевода научно-популярной книги «Финляндия вчера и сегодня» К. Хяккинен и С. Цеттерберга. Содержит параллельные тексты на 7 языках (финском, русском, эрзянском, мокшанском, луговом марийском, удмуртском и коми), около 24 000 слов и 1 800 выровненных предложений на одну языковую версию.

Многоязычный корпус параллельных текстов «Turku 'Pavlik Morozov' Corpus (parallel texts)»[43] построен на основе перевода романа В. Г. Губарева «Павлик Морозов» на следующие языки: эрзянский (версии 1953 и 2007 гг.), мокшанский, горно-марийский и луговой марийский, удмуртский, коми-пермяцкий, коми, венгерский (две версии), ханты, манси, финский, русский, чувашский и татарский. Размер ресурса — около 10 000 слов, 1 600 выровненных предложений на одну языковую версию.

Электронные списки слов: марийский, мордовский, удмуртский, коми, чувашский, татарский=Electronic Word Lists: Mari, Mordvin, Udmurt, Komi, Chuvash, Tatar[44], как отмечают составители, предназначены для изучения словообразования и структуры слов. Общее количество записей в шести списках составляет около 327 000 лексем: мордовский — 75 000, коми — 70 000, марийский — 54 000, удмуртский — 49 000, татарский — 46 000 и чувашский — 31 000 лексем. Каждое слово ввода снабжено метками, указывающими на язык, класс слов и источники словаря. Значения слов в списке не приводятся. Альтернативными языками пользовательского интерфейса программы являются английский, русский и финский. Существует две версии каждого списка слов: первая, построенная по алфавиту (начиная с первой буквы слова), вторая — с обратным алфавитным порядком (начиная с конца слова). Файлы, например, озаглавлены следующим образом: mordva_alph.txt, mordva_rev.txt. Для работы с такими списками разработана специально компьютерная программа SFOu WordListTool. Корпус общедоступен по ссылке: https://www.sgr.fi/fi/items/show/404.

Проект Universal Dependenci (UD) (Универсальных зависимостей) — это открытое научное сообщество, в рамках которого разрабатываются принципы универсального формата разметки для разноструктурных языков[45]. Как отмечают исследователи, он призван облегчить работу над мультиязычным машинным переводом и обеспечить лучшую совместимость банков синтаксических деревьев. Последнее, в свою очередь, позволит переносить инструменты обработки текста с одного корпуса на другой, не обращая внимания на конкретный язык[46]. В проекте участвуют более чем 100 языков, в их числе и мордовские. Эрзянская часть включает 1 704 предложения и 17 289 синтаксических слов, мокшанская — соответственно 342 и 3 181[47].

Следующая крупная языковая платформа, содержащая цифровой материал на мордовских языках, — Корпуса уральских языков Поволжья (или Корпуса уральских языков Волго-Камья)[48]. Ресурс разработан однотипно для 5 финно-угорских языков (удмуртский, коми-зырянский, луговой марийский, эрзянский и мокшанский) исследователем Т. А. Архангельским. Имеет морфологическую разметку и находится в открытом доступе. Корпуса используют техническую инфраструктуру Школы лингвистики ВШЭ. Материал по языкам собран, как и в коллекции Wanca, из Интернета в два корпуса (корпус соцсетей и основной) и отражает в основном современное состояние языков. Как подчеркивает автор, ресурс будет полезен при изучении лексики, грамматики, диалектологии и социолингвистических вопросов, а также «усилит интерес к этим языкам и приведет к росту количества посвященных им исследований»[49]. Подробную характеристику корпусов, в том числе эрзянского и мокшанского[50], методику их сбора и создания автор описал в своих публикациях[51]. Что касается мордовских языков, Т. А. Архангельский на основе данных разрабатываемого им проекта отмечает: «…положение эрзянского и мокшанского языков с точки зрения их представленности в цифровой сфере заметно хуже, чем у пермских языков и лугового марийского…»[52]. Этот факт также подтверждают статистические показатели платформы «Языки России»[53], данные проекта «Малые языки России»[54] по мордовским языкам пока отсутствуют в открытом доступе.

ЛингвоДок=LingvoDoc — крупнейшая, функционирующая с 2012 г. отечественная языковая платформа, призванная способствовать активной цифровизации материалов языков национальных меньшинств[55]. Проект «ЛингвоДок» реализуется в лаборатории «Лингвистические платформы» Института системного программирования им. В. П. Иванникова (ИСП РАН) в партнерстве с Институтом языкознания РАН. Лингвистическая платформа предназначена для составления, анализа и хранения в цифровом формате словарей и корпусов в основном уральских и алтайских языков, она также имеет возможности картографирования лингвистических характеристик. В настоящее время собрано более 1 000 словарей и 300 текстовых корпусов на диалектах уральских и алтайских языков России. По эрзянскому языку на платформе размещен 31 словарь разного объема и 6 конкордансов книг, изданных в дореволюционный период, по мокшанскому — 17 словарей и 2 конкорданса книг[56]. Кроме инструментов морфологического и синтаксического анализа, ЛингвоДок предлагает программы фонетического и фонематического анализа, поиска этимологий, анализа когнатов и др. Ресурс планирует централизованно аккумулировать большие базы данных по финно-угорским и алтайским языкам России и на их основе продолжать разрабатывать технологии искусственного интеллекта. В декабре 2021 г. ЛингвоДок сообщила о создании консорциума научно-образовательных организаций по изучению языков России[57]. В рамках его деятельности на платформе создаются глоссированные корпуса со снятой омонимией на башкирском, якутском, мордовских и удмуртском языках, которые в будущем также станут основой обучающих ресурсов.

Некоторый реестр корпусов финно-угорских языков дан на сайте Национального корпуса русского языка[58].

Таким образом, имеющиеся цифровые инструменты мордовских языков создавались в основном зарубежными институтами. К сожалению, доступ к ним иногда ограничен. Данная область требует большего внимания со стороны отечественных исследователей.

Библиографические ссылки

[1] Multilingual Resource Collection of the University of Helsinki Language Corpus Server (UHLCS). URL: http://www.ling.helsinki.fi/uhlcs/data/databank.html (дата обращения: 20.12.2021).

[2] См.: **Федина М. С.** О необходимости создания ресурсного центра по поддержке языков коренных народов России в цифровом пространстве // Лингвистический форум 2021 : Языковая политика и сохранение языков : тезисы докл. Междунар. конф. (Москва, 11 — 13 нояб. 2021 г.). М., 2021. С. 94 — 95 ; **Ее же.** Финно-угорские языки Российской Федерации в электронном информационном пространстве: опыт, проблемы, перспективы // Финно-угорский мир. 2016. Т. 3. № 28. С. 111 — 121 ; **Ácsjudit J., Pajkossy K., Kornai A.** Digital vitality of Uralic languages // Acta Linguistica. 2017. Vol. 64: Is. 3. URL: https://akjournals.com/view/journals/2062/64/3/article-p327.xml?body=pdf-23898 (дата обращения: 20.12.2021) .

[3] Kielipankki=Language Bank of Finland (Языковой банк). URL: https://www.kielipankki.fi/tuki/korp/ ; https://metashare.csc.fi/repository/search/?q=&selected_facets=languageNameFilter_exact%3AErzya (дата обращения: 20.12.2021).

[4] **Borin L., Forsberg M., Roxendal J.** Korp — the corpus infrastructure of Språkbanken // Proceedings of LREC. Istanbul. 2012. URL: https://spraakbanken.gu.se/en,= (дата обращения: 20.12.2021).

[5] URL: http://gtweb.uit.no/korp (дата обращения: 20.12.2021).

[6] URL: https://korp.csc.fi ; https://korp.csc.fi/korplab/#?cqp=%5B%5D&corpus= (дата обращения: 20.12.2021).

[7] URL: https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus= (дата обращения: 20.12.2021).

[8] URL: http://gtweb.uit.no/u_korp/?mode=myv#?lang=en (дата обращения: 20.12.2021).

[9] См.: **Рютер Д., Партанен Н.** Новые текстовые корпусы миноритарных языков на сервере Korp.csc.fi.=On New Text Corpora For Minority Languages On The Helsinki korp.csc.fi Server // Электронная письменность народов Российской Федерации : опыт, проблемы и перспективы : материалы II Междунар. науч. конф. (Уфа, 11 — 12 дек. 2019 г.). Уфа, 2019. С. 35.

[10] **Rueter J., Yerina O.** ERME—Erzya and Moksha Extended Corpora [text corpus]. Kielipankki. 2017. URL: http://urn.fi/urn:nbn:fi:lb-201407306 ; https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus=erme_myv ; https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus=erme_mdf (дата обращения: 20.12.2021).

[11] URL: http://gtweb.uit.no/u_korp/?mode=myv ; http://gtweb.uit.no/u_korp/?mode=mdf#?lang=en&stats_reduce=word&cqp=%5B%5D (дата обращения: 20.12.2021).

[12] Kansalliskirjasto. Fenno-Ugrica Kielipankin ladattava versio [tekstikorpus]. Kielipankki. 2013. URL: http://urn.fi/urn:nbn:fi:lb-2016092001 ; https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus=fennougrica_myv,fennougrica_kca,fennougrica_izh,fennougrica_mhr,fennougrica_mrj,fennougrica_mns,fennougrica_mdf,fennougrica_sel,fennougrica_yrk,fennougrica_vep. (дата обращения: 20.12.2021).

[13] The Finno-Ugrian Society Fieldwork Corpus [text corpus]. Kielipankki. URL: http://urn.fi/urn:nbn:fi:lb-2016092001 ; https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus=sust_myv,sust_kpv,sust_mdf (дата обращения: 20.12.2021).

[14] Finno-Ugric Languages and the Internet. URL: http://suki.ling.helsinki.fi/eng/project.html (дата обращения: 20.12.2021) ; **Jauhiainen H., Jauhiainen T., Lindén K.** Wanca in Korp : Text corpora for underresourced Uralic languages // Proceedings of the Research data and humanities (RDHUM) 2019 conference : data, methods and tools. P. 21 — 40.

[15] См.: **Jauhiainen T.** Tekstin kielen automaattinen tunnistaminen : Master's thesis. Helsinki, 2010 ; **Jauhiainen H., Jauhiainen T., Lindén K.** Wanca in Korp.

[16] Wanca. URL: http://suki.ling.helsinki.fi/wanca/ (дата обращения: 20.12.2021).

[17] URL: http://suki.ling.helsinki.fi/wanca/languages/22 (дата обращения: 20.12.2021).

[18] URL: http://suki.ling.helsinki.fi/wanca/languages/23 (дата обращения: 20.12.2021).

[19] См.: **Jauhiainen T., Jauhiainen H., Lindén K.** Suomalais-ugrilaiset kielet ja internet-projekti 2013 — 2019 // Multilingual Facilitation : This book has been authored for Jack Rueter in honor of his 60th birthday / M. Hämäläinen, N. Partanen, K. Alnajjar (eds.). University of Helsinki, 2021. P. 228 — 247. URL: https://doi.org/10.31885/9789515150257.21 (дата обращения: 20.12.2021).

[20] Ibid.

[21] Wanca 2016, Korp-versio [tekstikorpus]. Kielipankki. URL: http://urn.fi/urn:nbn:fi:lb-2019052401 ; https://korp.csc.fi/korplab/?mode=other_languages#?cqp=%5B%5D&corpus=wanca_2016_mdf_multili,wanca_2016_myv_multili (дата обращения: 20.12.2021).

[22] Multilingual Resource Collection of the University of Helsinki Language Corpus Server (UHLCS). URL: http://www.ling.helsinki.fi/uhlcs/ (дата обращения: 20.12.2021).

[23] См.: **Копотев М. В.** Корпусная лингвистика в Финляндии (обзор ресурсов) // Научно-техническая информация : Информационные процессы и системы. Сер. 2. 2003. № 6. С. 37 — 43 ; http://www.ling.helsinki.fi/uhlcs/data/helsinki-corpora-II.html ; http://metashare.ilsp.gr:8080/repository/browse/multilingual-resource-collection-of-the-university-of-helsinki-language-corpus-server/11cc1ddc33f211e2b67c005056be118e667f3ec574ed4aef98fde634b1e1af17/ (дата обращения: 20.12.2021).

[24] **Suihkonen P.** Uralic, Turkic, Indo-Iranian and Mongol languages; languages of Siberia and Caucasia (UHLCS), Helsinki Korp Version [text corpus]. Kielipankki. 2007. URL: http://urn.fi/urn:nbn:fi:lb-2017022808 (дата обращения: 20.12.2021).

[25] **Rueter J.** Ersän ja mokšan kirjallisuutta ja julkaisuja ja komisyrjäänin kirjallisuutta (UHLCS) [tekstikorpus]. Kielipankki. 2007. URL: http://urn.fi/urn:nbn:fi:lb-2014032612 ; **Его же.** Corpus of Erzya and Moksha Mordvin Literature and Journals and Komi Zyrian Literature (UHLCS), Helsinki Korp Version [text corpus]. Kielipankki. 2007. URL: http://urn.fi/urn:nbn:fi:lb-2017022816 (дата обращения: 20.12.2021).

[26] **Estill D.** Erzya and Moksha Mordvin Word List Corpus (UHLCS), Helsinki Korp Version [text corpus]. Kielipankki. 2007. URL: http://urn.fi/urn:nbn:fi:lb-2017022824 ; http://urn.fi/urn:nbn:fi:lb-2014032611 (дата обращения: 20.12.2021).

[27] **Rueter J., Axelson E.** Parallel Bible Verses for Uralic Studies, Korp [text corpus]. Kielipankki. 2020. URL: http://urn.fi/urn:nbn:fi:lb-2020021121 (дата обращения: 20.12.2021).

[28] URL: http://www.ling.helsinki.fi/uhlcs/readme-all/README-uralic-lgs.html# (дата обращения: 20.12.2021).

[29] **Suihkonen P.** Quantifiers and Quantification in Finnish and Languages Spoken in the Central Volga—Kama Region (UHLCS), Helsinki Korp Version [text corpus]. Kielipankki. 2016. URL: http://urn.fi/urn:nbn:fi:lb-2017030104 (дата обращения: 20.12.2021).

[30] Corpus of Colloquial Erzya [speech corpus]. Kielipankki. URL: http://urn.fi/urn:nbn:fi:lb-2014073034 (дата обращения: 20.12.2021).

[31] Эрзянско-финско-эрзянские словари Neahttadigisánit. URL: http://valks.oahpa.no/myv/fin/ (дата обращения: 20.12.2021).

[32] Мокшанско-финско-мокшанские словари Neahttadigisánit. URL: http://valks.oahpa.no/mdf/fin/ (дата обращения: 20.12.2021).

[33] URL: http://www.eki.ee/dict/ersa/index.cgi (дата обращения: 20.12.2021).

[34] URL: https://doi.org/10.15155/1-00-0000-0000-0000-0014AL (дата обращения: 20.12.2021).

[35] См.: **Moisio A., Luutonen J.** Turun yliopiston volgakielten korpukset // XXXIII Kielitieteen päivät, 123. Helsinki, 1996.

[36] UTU-Digilang : Kieliaineistoportaali. URL: https://digilang.utu.fi/ ; https://sites.utu.fi/utu-digilang/wp-content/uploads/sites/996/2021/11/UTU-Digilang-seka-Suomen-ja-sen-sukukielten-arkisto.pdf (дата обращения: 20.12.2021).

[37] URL: https://finno-ugric-corpora.utu.fi/cqpweb/ (дата обращения: 20.12.2021).

[38] URL: https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_mormula_erzya_v3&why=6 ; https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_mormula_moksha_v3&why=6 (дата обращения: 20.12.2021).

[39] См.: **Kurki T.** Digilang — Turun yliopiston digitaalisia kieliaineistoja kehittämässä / T. Kurki [et al] // Proceedings of the Research data and humanities (RDHUM) 2019 conference : data, methods and tools. P. 41 — 56.

[40] URL: https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_long_erzya_v2&why=6 ; https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_long_moksha_v2&why=6 (дата обращения: 20.12.2021).

[41] MokshEr Corpus (Moksha, Erzya). URL: https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_moksher_erzya_v3&why=6 ; https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_moksher_moksha_v3&why=6 (дата обращения: 20.12.2021).

[42] Finland — Past and Present' Corpus (parallel texts). URL: https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_p_sk_erzya_v1&why=6 ; https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_p_sk_moksha_v1&why=6 (дата обращения: 20.12.2021).

[43] Turku 'Pavlik Morozov' Corpus (parallel texts). URL: https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_p_pm_erzya_v1&why=6 ; https://finno-ugric-corpora.utu.fi/cqpweb/usr/index.php?ui=accessDenied&corpusDenied=t_p_pm_moksha_v1&why=6 (дата обращения: 20.12.2021).

[44] Electronic Word Lists: Komi, Chuvash and Tatar. With SFOu WordListTool 1.4. Ed. Jorma Luutonen et al. Lexica Societatis Fenno-Ugricae XXXI:2. 2016. URL: https://www.sgr.fi/fi/items/show/404 (дата обращения: 20.12.2021).

[45] Universal Dependenci (UD). URL: https://universaldependencies.org/introduction.html (дата обращения: 20.12.2021).

[46] **Ляшевская О. Н.** Корпуса Universal Dependencies, или зачем нам еще один формат синтаксической разметки? URL: https://ling.hse.ru/news/197443947.html (дата обращения: 20.12.2021).

[47] См.: **Рютер Д.** Корпус национальных мордовских языков: принципы разработки и перспективы функционирования // Финно-угорские народы в контексте формирования общероссийской гражданской идентичности и меняющейся окружающей среды : материалы Междунар. науч. конф., г. Саранск, 8 — 9 окт. 2020 г. Саранск, 2020. С. 118 — 127 ; URL: https://universaldependencies.org/treebanks/myv_jr/index.html ; https://universaldependencies.org/mdf/index.html ; https://universaldependencies.org/treebanks/mdf_jr/index.html (дата обращения: 20.12.2021).

[48] Корпуса уральских языков Поволжья. URL: volgakama.web-corpora/net (дата обращения: 20.12.2021).

[49] **Архангельский Т. А.** Интернет-корпуса финно-угорских языков // Ежегодник финно-угорских исследований. 2019. Т. 13, № 3. С. 534.

[50] URL: http://erzya.web-corpora.net/ ; http://moksha.web-corpora.net/ (дата обращения: 20.12.2021).

[51] **Arkhangelskiy T. A.** Corpora of social media in minority Uralic languages // Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, January 7 — January 8, 2019. Tartu, 2019. P. 125 — 140 ; **Его же.** Интернет-корпуса финно-угорских языков… С. 528 —536.

[52] **Архангельский Т. А.** Интернет-корпуса финно-угорских языков… С. 532.

[53] Языки России. URL: http://web-corpora.net/wsgi3/minorlangs/view/mdf (дата обращения: 20.12.2021).

[54] Малые языки России. URL: https://minlang.iling-ran.ru/langs ; https://minlang.iling-ran.ru/corpora (дата обращения: 20.12.2021).

[55] URL: http://lingvodoc.ispras.ru/ (дата обращения: 20.12.2021).

[56] URL: http://lingvodoc.ispras.ru/dashboard/dictionaries_all (дата обращения: 20.12.2021).

[57] URL: https://www.ispras.ru/groups/modis/laboratoriya-lingvisticheskie-platformy/ ; https://www.youtube.com/watch?v=s_Ntwgr1Zt0#t=1h19m00s&ab_channel=%D0%98%D0%A1%D0%9F%D0%A0%D0%90%D0%9D (дата обращения: 28.12.2021).

[58] Национальный корпус русского языка. URL: https://ruscorpora.ru/new/corpora-other.html (дата обращения: 20.12.2021).

Поступила 01.04.2022 г.