ИНСТИТУТ ЯЗЫКОЗНАНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

НАУЧНЫЙ ЦЕНТР ПО СОХРАНЕНИЮ, ВОЗРОЖДЕНИЮ И ДОКУМЕНТАЦИИ ЯЗЫКОВ РОССИИ

Итоговые документы, подготовленные в ходе научно-исследовательской работы «Разработка теоретических принципов и практических аспектов сохранения и возрождения языков России»

6. Описание методик создания ориентированных на языковые сообщества онлайн-инструментов поддержки языков

Е. М. Будянская
А. П. Евстигнеева

2022

СОДЕРЖАНИЕ

Введение 4

1. Цели, задачи и актуальность исследования 4

2. Обзор проблематики 6

3. Существующие лингвистические онлайн-инструменты 9

3.1. Онлайн-инструменты, ориентированные на документацию 9

3.1.1. Архивы 10

3.1.2. Работа с метаданными 13

3.1.3. Программное обеспечение для лингвистического анализа 14

3.1.4. Платформы и приложения для собственно документации 16

3.1.5. Корпусные платформы 18

3.1.6. Коллекции документационных материалов для широкой аудитории 19

3.1.7. Использование архивных материалов 19

3.2. Онлайн-инструменты для овладения языком 20

3.2.1. Работа с текстами 21

3.2.2. Словари 21

3.2.3. Языковые курсы 22

3.2.4. Инструменты для учителей 23

3.2.5. Примеры платформ с обучающими материалами 23

3.3. Онлайн-инструменты для создания контента на языке 24

3.3.1. Инфраструктура 24

3.3.2. Контент и общение на языке в сети Интернет 25

3.3.3. Пример платформы с коллекцией материалов на языке 26

4. Методики создания ориентированных на языковые сообщества онлайн-инструментов поддержки языков 26

4.1. Лингвистические корпуса текстов 27

4.2. Приложение Language Forge 28

4.3. Платформа ЛингвоДок 30

4.4. Платформа Living Dictionaries 32

4.5. Приложение LIG-Aikuma 34

4.6. Платформа Ревита 35

4.7. Приложения Динары Стёпиной 37

4.8. Перевод мультфильмов «Гора самоцветов» и «Лео и Тиг» 39

4.9. Субтитры в YouTube 40

4.10. Википедия на языке 42

5. Пилотная онлайн-платформа по ревитализации и документации 43

5.1. Структура онлайн-платформы 44

5.2. Раздел онлайн-платформы «Базы данных» 46

6. Перспективы дальнейшей работы 49

7. Отражение результатов работы в проекте Федеральной программы по сохранению и развитию языков России 50

Заключение 51

Список литературы 52

ПРИЛОЖЕНИЕ: Ориентированные на языковые сообщества онлайн-инструменты поддержки языков 58

Введение

Настоящий документ подготовлен группой исследователей, работающих в Научном центре по сохранению, возрождению и документации языков России Института языкознания РАН. Перед сотрудниками была поставлена задача изучить существующие лингвистические онлайн-инструменты с точки зрения их возможного использования для поддержки, ревитализации и документации языков, описать базовые методики создания ориентированных на языковые сообщества онлайн-инструментов. Кроме того, была запланирована разработка интерфейса пилотной онлайн-платформы по документации и ревитализации, содержащей в том числе базы данных по состоянию языков России, по практикам ревитализации языков, а также по задокументированности языков России, подготовленных Научным центром в 2021 году. Данная задача выполнялась в тесной взаимосвязи с подготовкой проекта Федеральной программы по сохранению и развитию языков России, разработанной осенью 2022 года в Институте языкознания РАН.

В ходе работы был проведен подробный обзор актуальных и поддерживаемых лингвистических онлайн-инструментов, так или иначе связанных с документацией и поддержкой языков. Исследование включало уточнение различных параметров таких инструментов: конкретная направленность, основной функционал и ключевые возможности, для каких языков реализовано, на какие языки переведен интерфейс, разработчики и возможность связаться с ними, необходимые ресурсы, чтобы воспользоваться инструментом, в том числе финансирование, а также наличие обучающих материалов в открытом доступе.

В результате исследования было проанализировано более 90 онлайн-инструментов и составлена база данных с описанием данных инструментов по выработанным параметрам.

В итоговом документе, к которому прилагается аналитическая таблица (Приложение), рассматривается общая проблематика онлайн-технологий, применяемых в документации и ревитализации языков (Раздел 2), обсуждаются параметры, релевантные для описания онлайн-инструментов, представляется классификация инструментов по сферам применения и общий обзор инструментов (Раздел 3), а также методология создания таких инструментов с приведением конкретных примеров (Раздел 4). Далее приводится описание создаваемой пилотной онлайн-платформы (Раздел 5) и обозначаются перспективы развития темы и планы на будущее (Раздел 6).

Кроме того, исследовательская группа подготовила перечень мероприятий для поддержки языков в области онлайн-технологий, которые нашли свое отражение в проекте Федеральной программы по сохранению и развитию языков России (Раздел 7).

1. Цели, задачи и актуальность исследования

Можно выделить четыре ключевые цели для данного направления исследований Научного центра и соответствующие им задачи.

Первая цель заключалась в том, чтобы изучить существующие лингвистические онлайн-инструменты с точки зрения их возможного использования для поддержки, ревитализации и документации языков. Для этого необходимо было выполнить следующие задачи: собрать коллекцию релевантных инструментов; выделить параметры описания инструментов; описать собранные в коллекцию инструменты по выработанным параметрам; классифицировать инструменты по сферам использования, а затем структурировать полученную информацию в формате базы данных.

Второй целью было описать базовые методики создания ориентированных на языковые сообщества онлайн-инструментов. Для этого необходимо было выполнить следующие задачи: связаться с разработчиками, обсудить принципы устройства и работы таких инструментов, а также уточнить необходимые для их создания навыки, оборудование,

языковые материалы и финансирование; отразить полученные сведения в виде справочных материалов.

Третья цель носила более практический характер и заключалась в разработке интерфейса пилотной онлайн-платформы по документации и ревитализации, содержащей в том числе базы данных по состоянию языков России, по лучшим российским и зарубежным практикам ревитализации языков, а также по задокументированности языков России, подготовленных Научным центром в 2021 году. Для реализации данной цели необходимо было выполнить следующие задачи: разработать структуру пилотной онлайн-платформы; продумать интуитивно понятный интерфейс; наполнить онлайн-платформу информацией; разработать структуру и способ онлайн-представления всех трех баз данных; продумать механизм пополнения информацией баз данных посетителями сайта; спроектировать структуру пополнения базы данных по состоянию языков России, позволяющую отслеживать изменения в языковой ситуации; разработать механизм взаимосвязей между базами данных; вместе с программистом отслеживать процесс реализации и при необходимости вносить корректировки.

Кроме того, с опорой на результаты исследования данного направления, необходимо было подготовить список мероприятий, касающихся использования онлайн-технологий в сфере поддержки языков России, для включения в проект Федеральной программы по сохранению и развитию языков России, разработанный осенью 2022 года в Институте языкознания РАН.

Каждая из этих целей обладает значимой актуальностью в современном контексте документации, сохранения и развития языков России. Развивающиеся быстрыми темпами онлайн-технологии становятся все более доступными для широкого круга людей, что делает их эффективным инструментом поддержки и развития языков (см. [Харитонов и др. 2021: 18–20; Корнев 2021: 38–41; Винклер, Павлова 2021: 44–48]). Это связано не только с постоянно растущим функционалом онлайн-платформ и мобильных приложений, но и с распространением смартфонов и расширением зоны покрытия сети Интернет.

Кроме того, онлайн-инструментов в последнее время появляется довольно много, при этом ориентироваться в них становится проблематично (см., например, Рисунок 1).

С этой точки зрения задача создания структурированной базы данных по таким инструментам является своевременной и актуальной. Немаловажным является удобство такой базы, ее интерфейса, и ее доступность для языковых активистов и представителей сообщества. Кроме того, получили русскоязычное описание некоторые англоязычные инструменты, что может оказаться удобным для первого знакомства.

Среди представителей языковых сообществ в России растет интерес к созданию онлайн-инструментов для своих языков, появляется все больше подобных инициатив. С этой точки зрения актуально создание справочной системы, которая сделала бы процесс выбора и разработки таких инструментов проще и доступнее для широкого круга людей. Онлайн-платформа с такой базой данных могла бы стать местом встречи специалистов из разных областей науки, языковых активистов, педагогов, представителей языковых сообществ для обмена опытом и, возможно, создания команды для разработки того или иного инструмента для поддержки языков России.

Отдельно стоит отметить, что описание онлайн-инструментов по овладению языками или для создания контента на языках актуально и для документационной лингвистики, поскольку такие инструменты создают новые возможности для использования языковых материалов, записанных в ходе документационных проектов. В современной документации языков развивается тенденция максимально широкого использования языковых материалов, в том числе с учетом потребностей языковых сообществ, о чем см., например, [Сюрюн и др. 2022].

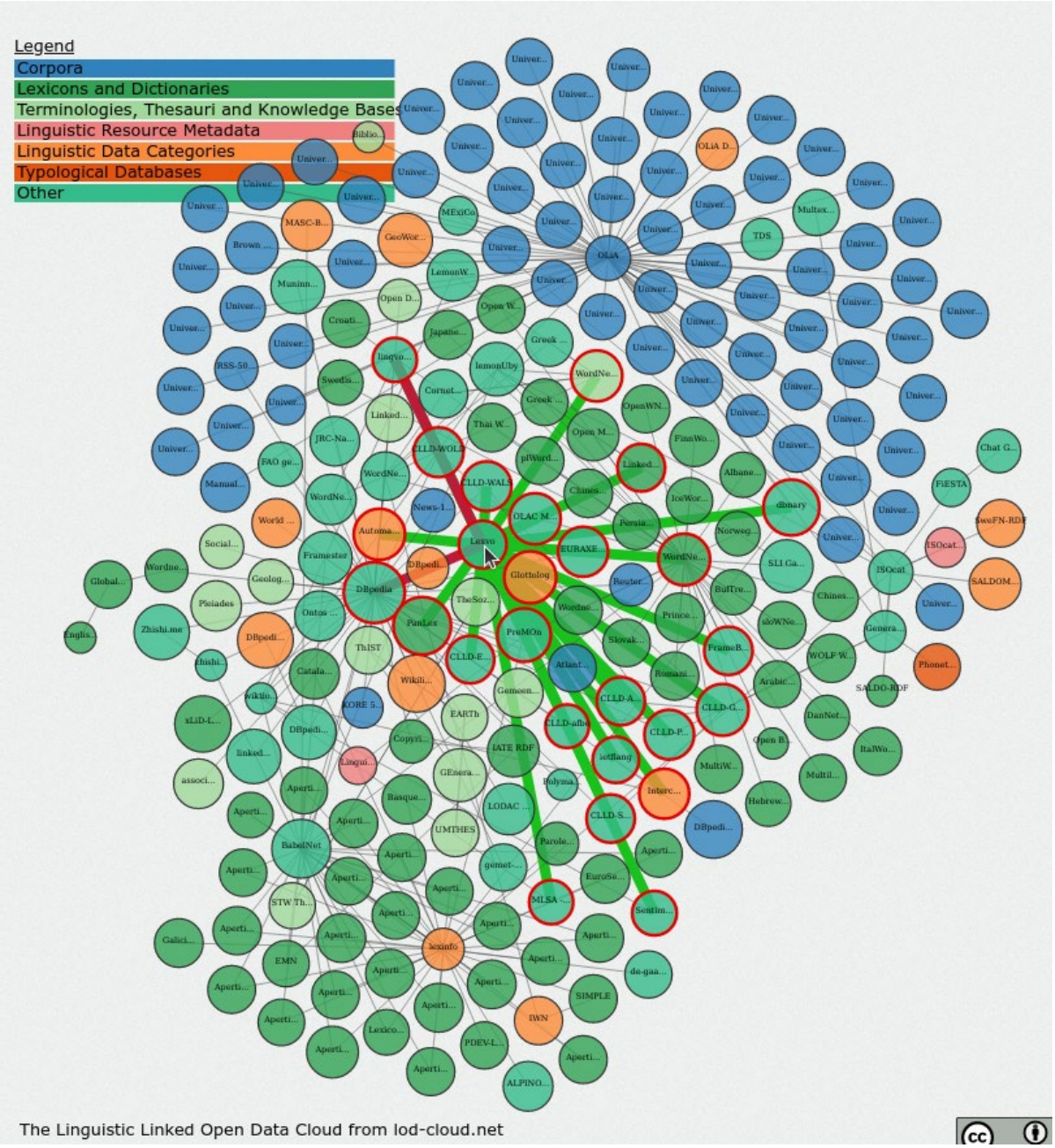


Рисунок 1. Облако связей лингвистических открытых данных, выполненное на сайте «Linguistic Linked Open Data Cloud», доступном по ссылке: https://linguistic-lod.org/

2. Обзор проблематики

Цифровизация в разных сферах науки и жизни в последние десятилетия стала неотъемлемой частью развития этих областей. В области лингвистики и сохранения языков цифровизация расширяет старые и открывает новые возможности как для лингвистов, так и для носителей языков. С одной стороны, оцифрованные материалы становятся более доступными, с другой стороны, хранение в электронном виде уменьшает риски потери информации. Кроме того, развитие и распространение цифровых технологий делает их приоритетным средством получения информации и инструментом работы с ней. Соответственно, эффективность цифровых инструментов в сфере сохранения и развития языков становится выше, чем эффективность аналоговых средств (см., например, [Galla 2018]). Так, все больше людей переходит на использование электронных словарей, а смартфон оказывается чаще под рукой, чем бумажный словарь.

О положительном отношении среди представителей сообществ к использованию цифровых ресурсов для овладения своим языком говорится и в [Li et al. 2021]. Представители языковых сообществ часто озвучивают потребность в интерактивных обучающих онлайн-инструментах, которые позволяли бы изучать свой язык с таким же техническим удобством, как и крупные иностранные языки. Например, проект «FirstVoices» (с англ. 'Главные голоса') уже около 20 лет ведет работу по технической поддержке инициативных групп, работающих над созданием онлайн-инструментов для своих языков [Chase, Borland 2022].

Наличие контента на языке в сети Интернет повышает видимость языка, его престиж и создает языковую среду. У представителей языкового сообщества появляется возможность смотреть мультфильмы, новости или любимых блогеров на своем языке. Причем это может быть как оригинальный контент на языке, так и переводной (автоматические или ручные субтитры, а в перспективе и автоматическая озвучка). В случае, когда представители языкового сообщества проживают далеко друг от друга, языковая среда, создаваемая онлайн, может объединить людей, дать им возможность общаться и взаимодействовать на своем языке и, таким образом, сыграть ключевую роль в сохранении языка. Вдобавок ко всему вышеперечисленному в статье [Huilcán Herrera 2022], например, раскрывается представление о том, как использование технологий может не просто быть использованием технологического устройства, а поддерживать сообщество в обновлении или восстановлении их национальной идентичности.

В рамках нашего исследования к онлайн-инструментам мы относим платформы и приложения, доступ к которым осуществляется с подключением к сети Интернет. Таким образом, мы различаем инструменты, функционирующие через браузер (платформы), и инструменты, требующие установки на электронные устройства (приложения).

В фокусе нашего внимания были онлайн-инструменты, которые направлены на поддержку языка и ориентированы на языковое сообщество. Такие инструменты должны быть доступными, простыми в использовании и, желательно, кросс-платформенными. О необходимости доступности и распространенности информации о подобных инструментах много говорится как среди разработчиков и лингвистов, так и среди представителей сообществ, в которых появляются инициативы по созданию различных инструментов. Так, в докладе «Digital preservation for the masses: Creating and publishing workflows and guidelines for institutions and smaller organisations interested in digital preservation, data management, and archiving» (с англ. 'Цифровое хранение для широкой аудитории: создание и публикация пошаговых инструкций и руководств для учреждений и небольших организаций, заинтересованных в цифровом хранении, управлении данными и архивировании'), представленного на конференции «Документация и архивация. Перспективы развития» (англ. «Language Documentation and Archiving. Where do we go from here?») (Берлин, октябрь 2022 года) [Miller 2022] рассказывается об инициативе лингвистов создать на платформе GitHub коллекцию инструкций по использованию архива PARADISEC[1] (об этом архиве см. подробнее в Разделе 3.1.1).

Важным фактором при отборе инструментов было и то, чтобы в качестве языковых данных для них могли бы быть использованы материалы, полученные в ходе документационных исследований. В рамках подхода к документации языков с акцентом на ревитализации одним из ключевых моментов является обеспечение видимости и доступности для представителей языкового сообщества записанных материалов, но на данный момент многие коллекции языковых материалов оказываются по разным причинам недоступными не только носителям языка, но и широкому кругу исследователей. К этой проблеме привлекают внимание авторы многих докладов, представленных на конференции «Документация и архивация. Перспективы развития», например, «How can endangered language collections be made more discoverable?» (с англ. 'Как сделать так, чтобы

[1] Архив, организованный консорциумом австралийских исследователей: https://catalog.paradisec.org.au/.

коллекции материалов по языкам, находящимся под угрозой исчезновения, было легче найти?') [Babinski et al. 2022], «Archiving the Ende Language Corpus» (с англ. 'Архивирование корпуса языка энде') [Garnica, Lindsey 2022]. В то же время, во многих документационных проектах исследователи начинают использовать онлайн-инструменты для работы с носителями языка. Собранные таким образом данные остаются доступными онлайн для участников исследования. Так, в статье [Bettinson, Bird 2021] описывается опыт не только работы с таким подходом, но и разработки специфически настроенных на конкретное сообщество мобильных приложений. В другой работе [Everson et al. 2019] говорится о создании такой платформы, которая объединит возможности записи языковых данных в аудиоформате и возможности хранения этих данных, снабженных метаданными.

Как показано выше, многие группы лингвистов и программистов разрабатывают различные лингвистические онлайн-инструменты. Для того, чтобы было удобно выбрать подходящий под поставленные задачи инструмент, необходима некоторая их систематизация. Подобные проекты по структурированию информации в области лингвистических технологий существуют (см., например, [Антопольский 2021]), но направлены они на другую выборку инструментов и описывают их по другим параметрам, нежели в нашем исследовании.

Так, Навигатор информационных ресурсов по языкознанию (НИРЯЗ), разработанный в 2019–2020 годах группой сотрудников в ИНИОН РАН, содержит широкий перечень ресурсов по языкознанию, а также специфических лингвистических ресурсов, создаваемых учреждениями РАН и республиканских академий наук. При этом он включает сведения не только об электронных ресурсах (электронные библиотеки, базы данных, информационные системы, сайты, социальные сети и др.), но и о традиционных (библиотечных, архивных, музейных). Типы инструментов, на которые направлено наше исследование, частично пересекаются с тем, что в данном Навигаторе называется лингвистическими ресурсами. В НИРЯЗ предусмотрено несколько видов навигации и поиска ресурсов. Информационные ресурсы классифицируются в соответствии со следующими параметрами: владелец, тип, тематика ресурса и язык, которому посвящен ресурс. Поиск можно осуществлять по этим параметрам, а также по названию ресурса.

Каталог ресурсов для обработки естественного языка (NLPub) появился в 2012 году (автор проекта — Д. А. Усталов) и создается на принципах краудсорсинга. Каталог организован по вики-принципу и содержит сведения об инструментах, ресурсах, методах и алгоритмах, необходимых для успешного построения систем автоматической обработки языка. Эти инструменты на данном этапе ориентированы на крупные государственные языки, применить их к материалу миноритарных языков представляется проблематичным. Информационные ресурсы, представленные в каталоге наряду с инструментами, не относятся к предмету нашего исследования.

Портал знаний по компьютерной лингвистике разрабатывается с 2007 года сотрудниками Института систем информатики имени А. П. Ершова СО РАН совместно с учеными из других организаций СО РАН, Москвы и Казани. Данный портал, как и предыдущие, пересекается с нашей деятельностью лишь частично (например, в разделах «Интернет-ресурсы» и «Научные результаты и продукты») и в основном состоит из перечисления информационных ресурсов, а не онлайн-инструментов.

В статье [van Esch et al. 2019] подобный нашему подход реализуется в области разработок и программного обеспечения для лингвистического анализа языковых данных, записанных в ходе документационных проектов. В работе предлагаются такие пути расширения функционала лингвистических инструментов, которые учитывали бы задачи сохранения и развития языков. Например, форматы, в которых данные экспортируются после анализа в каких-либо программах, должны быть совместимы с другими инструментами, нацеленными на поддержку языков. Авторы считают, что подобный обзор имеющихся инструментов поможет обозначить лакуны и дальнейшие пути развития в этой

области, а также предоставит возможность коммуникации между разработчиками и полевым лингвистами.

Более близким к нашим задачам оказался ресурс CLARIN (англ. *Common Language Resources and Technology Infrastructure* ‘Языковые ресурсы общего назначения и технологическая инфраструктура’) — это исследовательская инфраструктура, созданная в 2012 году. Она обеспечивает для ученых из области социальных и гуманитарных наук доступ, с одной стороны, к цифровым языковым данным (в письменной, устной или мультимодальной форме), а с другой стороны, к инструментам по исследованию, использованию, аннотированию и анализу таких данных. Это становится возможным благодаря сетевому объединению центров: хранилищ языковых данных, сервисных центров и центров знаний с единым входом в систему для всех членов академического сообщества во всех странах-участницах. Инструменты и данные из разных центров совместимы, поэтому наборы данных можно комбинировать, а инструменты из разных источников можно объединять в цепочки для выполнения сложных операций для поддержки исследователей в их работе. В разделе «Данные» представлены различные хранилища, содержащие коллекции языковых материалов. В разделе «Инструменты» собраны ссылки на различные центры, содержащие коллекции языковых инструментов. В разделе «Семейства ресурсов» приводится удобный для пользователя обзор доступных языковых ресурсов, которые отсортированы по типам данных (в настоящее время доступны обзоры для 14 семейств корпусов, 6 семейств лексических ресурсов и 4 семейств инструментов). В разделе «Услуги» представлены специализированные инструменты, упрощающие поиск необходимой информации и материалов в рамках данного ресурса. Кроме того, CLARIN также ежегодно проводит конференции, семинары и другие профессиональные встречи. Стоит отметить, что некоторые онлайн-инструменты мы почерпнули именно из этого ресурса.

Главным отличием настоящего исследования от предыдущих работ является отчетливо прикладная задача в нашем случае. Мы создаем не только и не столько перечень онлайн-инструментов, а предоставляем информацию о том, как ими пользоваться, какие материалы для этого нужны. В итоге получается не список, а методология создания инструментов. Пользователи получают ответ на вопрос, что можно сделать с языковыми данными, которые у них имеются, а также какие данные нужны, чтобы реализовать какую-то конкретную цель.

3. Существующие лингвистические онлайн-инструменты

Существующие лингвистические онлайн инструменты мы разделили на три большие группы: онлайн-инструменты, ориентированные на процесс документации, онлайн-инструменты, помогающие в процессе овладения языком и в развитии языковых компетенций, и онлайн-инструменты, ориентированные на создание контента на языке и языковой среды. В каждой группе мы выделили по несколько подгрупп.

3.1. Онлайн-инструменты, ориентированные на документацию

Среди онлайн-инструментов, ориентированных на документацию, можно выделить несколько групп, различающихся функционально и структурно. Это онлайн-архивы с различными уровнями доступа и системой навигации, программное обеспечение для анализа языковых данных, платформы и приложения для собственно документации, корпусные платформы для развертывания текстовых корпусов, информационные ресурсы с коллекциями языковых материалов, с удобным и простым интерфейсом, а также некоторые более узкозадачные инструменты. Остановимся на каждой из этих групп отдельно.

3.1.1. Архивы

Специализированные онлайн-архивы вносят ключевой вклад в сохранение и распространение языковых данных. Архивы, с одной стороны, являются важными участниками рабочего процесса документации в качестве ресурса по систематизации и хранению собранных материалов [Nathan 2008, 2011]. С другой стороны, они становятся источниками языковых данных, которые могут быть использованы для различных целей, как научных, так и ревитализационных. Так, на основе данных другие специалисты могут проводить свои исследования. Вместе с тем, на основе документационных материалов можно создавать учебные материалы и приложения, контент на языке, а также различные цифровые инструменты. Соответственно, языковые архивы, как и другие научные цифровые хранилища, создаются с ориентацией на две основные аудитории. Это хранители языковых данных, которые могли бы занести их в архив, и различные потенциальные пользователи этих материалов: исследователи (лингвисты и другие), языковые сообщества, студенты, преподаватели, художники и т. д.

Многие архивы работают над тем, чтобы пользоваться ими становилось более удобно: развивают интуитивно понятный интерфейс, удобную систему навигации и систему метаданных, позволяющие легко находить необходимую информацию [Yi et al. 2022; Nathan 2010; Sullivant 2020]. Так, в исследовании [Burke et al. 2022] предоставляются эмпирические данные о потребностях пользователей языковых архивов и рассказывается об известных проблемах, с которыми сталкиваются менеджеры языковых архивов, люди, готовые внести материал в архив, и конечные пользователи. В последние года четко осознается потребность в том, чтобы к языковым архивам имели легкий доступ представители языковых сообществ [Panigrahi 2022]. Необходимо выработать такую систему, при которой академические архивы, обладающие возможностями долговременного хранения, становились бы более доступными и открытыми площадками. Тогда языковые активисты будут обеспечены надежным местом для хранения тех данных, которые они собирают, а представители языковых сообществ — примерами использования их языков. Например, архив «Коренные языки Америки» (англ. Native American Languages, NAL) [2] (Оклахомский университет, США) при разработке онлайн-платформы для своего цифрового архива сразу закладывает в необходимые качества простоту использования и доступность для представителей языковых сообществ. В докладе [Heaton, Hooshiar 2022], посвященном этим разработкам, рассказывается о серии воркшопов, проведенных в 2020 году, в которых, совместно с представителями коренного населения Оклахомы, т. е. будущими пользователями онлайн-платформы, обсуждались основные вопросы, связанные с ее функционированием. Среди них основными были вопросы навигации в системе, возможность и механизм пополнения, условия сотрудничества между архивом и теми, кто хочет внести в архив свои данные.

Далее остановимся подробнее на некоторых примерах онлайн-архивов.

Крупнейшим международным языковым архивом является The Language Archive (TLA)[3], который располагается в Институте психолингвистики им. Макса Планка в городе Неймеген (Нидерланды) и поддерживается Обществом Макса Планка, Берлин-Бранденбургской академией наук, Королевской академией наук и искусств Нидерландов и некоторыми внешними проектами. В нем содержатся текстовые, аудио- и видеоматериалы на различных языках мира, в том числе записи повседневной речи и некоторые экспериментальные данные. Команда TLA также разрабатывает лингвистические инструменты для работы с языковым материалом, его организацией и архивированием.

[2] Сайт архива: https://samnoblemuseum.ou.edu/collections-and-research/native-american-languages/.
[3] Сайт проекта: https://archive.mpi.nl/tla/.

Частью TLA является архив программы DOBES[4] (Dokumentation bedrohter Sprachen), инициированной фондом «Фольксваген» и направленной на документирование языков, находящихся под угрозой исчезновения. За 11 лет финансирования этой программы было запущено 68 проектов по документации. Отметим, что в архиве содержится материал 4 языков, распространенных в России. К аудио- и видеозаписям, хранящимся в архиве, обычно также прилагаются базовые сведения о языке и фотографии, связанные со сделанными записями. Кроме того, уделяется внимание документации не только языка, но и культурно-обрядовой деятельности носителей. Важным аспектом программы DOBES являются специально разработанные этические и правовые документы: «Кодекс поведения», регулирующий работу всех заинтересованных сторон; «Доступ к данным и принципы защиты», описывающий деятельность архиватора; соглашение между размещающим лицом и архиватором с условиями предоставления данных другим пользователям. Обычно доступ к коллекциям требует регистрации и заполнения формы запроса на использование.

ELAR (The Endangered Languages Archive)[5] представляет собой цифровой архив с материалами языков, чье исследование было поддержано грантами ELDP (Endangered Languages Documentation Programme)[6]. С июля 2021 года архив базируется в Берлин-Бранденбургской академии наук. В ELAR публикуются коллекции аудио- и видеозаписей, их транскрипции и переводы, а также словари и учебники по этим языкам. К данным предоставляется доступ разных уровней [Nathan 2014]. Кроме того, сотрудники архива занимаются оцифровкой печатных коллекций текстов, написанных на угрожаемых языках, таким образом делая их доступными для носителей этих языков, исследователей и общественности в целом. Российские лингвисты также участвуют в проектах ELDP, поэтому в архиве ELAR содержится достаточно большая коллекция аудио- и видеоматериалов по языкам России.

У Летнего института лингвистики (англ. SIL International)[7] также имеется собственный архив[8], в котором, помимо материалов, записанных работниками организации, содержатся опубликованные ими научные труды.

В Национальном центре научных исследований во Франции развивается проект Pangloss Collection[9] — открытый архив, содержащий записи на слабо задокументированных исчезающих языках. По состоянию на 2020 год в архиве выложено около 780 часов аудиозаписей (каждая из которых расшифрована, аннотирована и переведена на какой-либо крупный язык) на более чем 170 языках.

С 2009 года по 2013 год функционировал World Oral Literature Project, основанный Кембриджским университетом. Проект представлял собой инициативу по документированию и распространению устной литературы на различных языках мира. Кроме того, в цели проекта входили финансовая поддержка полевых исследователей и языковых активистов, занимающихся сбором устной литературы, а также обучение методикам полевых исследований и способам цифрового архивирования. Несмотря на то, что финансирование проекта было завершено, его архивы все еще доступны[10].

Существуют также и региональные архивы, которые, как правило, поддерживаются и развиваются местными университетами. Примерами региональных архивов являются Архив коренных языков Аляски (англ. Alaska Native Language Archive, ANLA)[11], основанный в 2009 году при Центре коренных языков Аляски (Университет Аляски

[4] Сайт архива: https://dobes.mpi.nl/.
[5] Сайт архива: https://www.elararchive.org.
[6] С данной программой можно ознакомиться по ссылке: https://www.eldp.net/.
[7] Сайт организации: https://www.sil.org/.
[8] Архив SIL International доступен по ссылке: https://www.sil.org/resources/language-culture-archives.
[9] Сайт проекта: https://pangloss.cnrs.fr/.
[10] Сайт проекта: http://www.oralliterature.org/.
[11] Архив доступен по ссылке: https://www.uaf.edu/anla/.

в Фэрбэнксе), и Архив по языкам Южной Америки на базе Техасского университета в Остине (англ. Archive of the Indigenous Languages of Latin America, AILLA)[12]. Еще одним примером является Калифорнийский языковой архив Калифорнийского университета в Беркли (англ. The California Language Archive, CLA)[13], представляющий собой физический и цифровой архив материалов, связанных с языками коренных народов Америки. Из региональных архивов также широко известен PARADISEC[14] — архив цифровых материалов преимущественно о языках и культурах Тихоокеанского региона, находящихся под угрозой исчезновения [Thieberger et al. 2015]. Проект объединяет архивы университетов Мельбурна, Сиднея и Австралийского государственного университета. PARADISEC использует международные стандарты для описания метаданных и является частью мирового сообщества языковых архивов. В настоящий момент в архиве представлен материал 1281 языка; объем аудиозаписей составляет 14000 часов, объем видеозаписей — 2000 часов. Примером регионального архива, ориентированного на представителей сообщества, также можно назвать «Цифровой архив этнических групп: содействие защите и сохранению языкового и культурного разнообразия в Таиланде» [Ungsitipoonporn et al. 2021]. Этот архив включает в себя коллекции материалов о 18 этнических группах Таиланда. В архиве содержатся данные по 385 проектам: видео-, аудиозаписи, тексты, изображения и файлы ELAN.

Существуют также ресурсы, которые объединяют информацию о разных архивах в одном месте.

Сеть цифровых архивов языков и музыки, находящихся под угрозой исчезновения, (англ. The Digital Endangered Languages and Musics Archives Network, DELAMAN)[15] была создана в 2003 году. Это международная сеть архивов данных о языковом и культурном разнообразии, в частности, о малочисленных языках и культурах, находящихся под угрозой. Раз в год проходит встреча членов этого объединения, членом может стать любой желающий. Одна из главных задач DELAMAN — быть информационным центром в области цифрового архивирования и лингвистических полевых исследований. На сайте есть страничка с ресурсами на эту тему. Организацией также был разработан минимальный чек-лист для надлежащего сохранения материалов языковой документации.

Открытое сообщество языковых архивов (англ. Open Language Archives Community, OLAC)[16] — это инициатива, основанная в 2000 году с целью создания унифицированного средства поиска информации по языкам мира в базах данных различных учреждений (в том числе архивах) и частных лиц. Информационный ресурс представляет собой виртуальную библиотеку, в которой перечислены все доступные материалы по языку (список имеющейся литературы, ссылки на аудио и видео). Информация о ресурсах хранится в формате XML. Данные депонированы на веб-сервере Консорциума лингвистических данных (англ. Linguistic Data Consortium) в Пенсильванском университете, США. В докладе [Paterson et al. 2022] было озвучено предложение добавить к данным о языках еще и информацию об обучающих материалах по этим языка и на них. Таким образом, в рамках данного сообщества также прослеживается тенденция взаимосвязи между лингвистической деятельностью, в частности, документацией, и ревитализацией языков.

Говоря о хранении данных и обеспечении доступа к ним, нужно упомянуть о некоторой альтернативе архивам, выросшим на базе университетов и музеев. Это универсальный репозиторий с открытым доступом Zenodo[17], разработанный в рамках европейской программы OpenAIRE, объединяющей открытые репозитории, и управляемый ЦЕРН (фр. Conseil Européen pour la Recherche Nucléaire, CERN). Данный репозиторий

[12] Архив по языкам Южной Америки на базе Техасского университета в Остине: https://ailla.utexas.org/.
[13] Архив доступен по ссылке: https://cla.berkeley.edu/.
[14] Архив доступен по ссылке: https://www.paradisec.org.au.
[15] Сайт проекта доступен по ссылке: https://www.delaman.org/.
[16] Ресурс доступен по ссылке: http://www.language-archives.org/.
[17] Репозиторий находится по ссылке: https://zenodo.org/.

позволяет исследователям депонировать исследовательские работы, наборы данных, исследовательское программное обеспечение, отчеты и любые другие цифровые документы, связанные с исследованиями. Для каждого представления создается постоянный идентификатор цифрового объекта (DOI), что позволяет легко цитировать хранимые элементы. Серверные мощности ЦЕРН позволяют хранить большое количество данных и предоставлять одновременный доступ к ним большому количеству пользователей. На Zenodo представлены данные по различным научным дисциплинам, в том числе и документационные материалы [Englisch 2022].

3.1.2. Работа с метаданными

Большую долю работы с языковыми данными, в том числе собранными в ходе документационных проектов, составляет организация материалов. Только структурированным архивом можно пользоваться. Структура архива строится на метаданных к материалам. Сейчас не существует единой системы метаданных, которой следовали бы все, кто занимается документацией. В архивах есть свои выработанные правила, но они отличаются друг от друга. Если же брать личные коллекции, то разнообразия тут еще больше. Опрос, проведенный австралийским исследовательским центром CoEDL (англ. *Centre of Excellence for the Dynamics of Language* 'Центр передового опыта в области динамики языка') в начале 2016 года, дал около 135 ответов и показал, что большинство ответивших исследователей не используют существующие инструменты для организации метаданных, а полагаются на электронные таблицы, текстовые документы или ручку и бумагу. Для унификации метаданных, облегчения и частичной автоматизации работы с ними разрабатываются различные приложения и инструменты. В том числе это очень актуально для удаленной документации, когда участники проекта находятся далеко друг от друга и, может быть, не общаются напрямую. Унификация метаданных необходима для того, чтобы создавать архивы с прозрачной навигацией и доступом к различным коллекциям языковых данных по всему миру.

Одним из таких инструментов является LaMeta [18], разработанный в результате реализации проекта, поддержанного программой ELDP и центром CoEDL. Он ориентирован на систематизацию коллекций, созданных в ходе документационных проектов. LaMeta позволяет не только структурировать метаданные, записанные в едином формате, но и связывать все файлы, относящиеся к одной документационной сессии, в том числе и сам файл с метаданными. Кроме того, есть возможность экспортировать проекты в различных форматах. Сейчас интерфейс LaMeta представлен только на английском языке, но пользователям открыта возможность поучаствовать в разработке и предложить свой перевод интерфейса. Это инструмент с открытым кодом, распространяется бесплатно. LaMeta не предусматривает коллективной работы над одним проектом. Другой же инструмент для работы с метаданными, разработанный в виде приложения и платформы, позволяет заполнять метаданные в режиме онлайн и работать над одним проектом нескольким участникам. Это KoBoToolbox [19], также распространяемый бесплатно. Описание того, как можно интегрировать этот инструмент в проект удаленной документации, можно почитать в статье [Griscom 2022]. Экспортированные из KoBoToolbox данные можно импортировать в LaMeta и таким образом создать файлы, необходимые для депонирования коллекции в некоторых архивах.

Иногда оказывается, что даже сформированные в LaMeta метаданные оказываются непоследовательными. Поскольку массив метаданных в объемных коллекциях становится существенным, проверить их качество и правильность вручную становится трудоемкой и времязатратной задачей. Для решения этой проблемы Ирен Йи и Клэр Бауэрн был

[18] Подробная информация и обучающее видео о LaMeta доступны по ссылке: https://sites.google.com/site/metadatatooldiscussion/home?authuser=0.

[19] Подробнее об KoBoToolbox можно почитать на сайте: https://www.kobotoolbox.org/.

разработан, например, скрипт на языке R «CheckR» [20], позволяющий определять незаполненные поля; поля, заполненные некорректно или непоследовательно; неправильно привязанные файлы и даже базовые несоответствия в файлах разметки ELAN.

3.1.3. Программное обеспечение для лингвистического анализа

С развитием технологий в документационной лингвистике развивались и технические способы анализа языковых данных, кроме того, появились и вошли в стандарт документации и медиаформаты (аудио и видео), что потребовало новых решений. Если изначально разбор текста, например, велся с помощью какого-либо текстового редактора, то сейчас разработаны гораздо более удобные программы, ориентированные именно на лингвистические задачи. Важной чертой этих программ является совместимость форматов, в которых можно экспортировать проанализированные материалы.

Широко распространенным среди лингвистов инструментом является ELAN (EUDICO Linguistic Annotator)[21], разработанный командой TLA. Эта программа позволяет производить многослойное аннотирование аудио- и видеозаписей. Содержание аннотаций может быть разное, в соответствии с задачами проекта. Ключевой возможностью является синхронизация аннотаций с медиафайлом. ELAN предоставляет множество дополнительных функций, делающих разметку более удобной. Кроме того, есть функция поиска по нескольким текстам какого-либо элемента аннотации. Данное программное обеспечение хорошо зарекомендовало себя не только в лингвистике, но и в медицине, психологии, психиатрии, изучении поведения и образовании.

SLEXIL (англ. *Software Linking ELAN XML to Illuminated Language* ‘Программное обеспечение, связывающее XML ELAN с освещаемым языком’) — это веб-приложение, позволяющее пользователям создавать анимированные HTML-файлы из синхронизированных со звуком аннотаций, сделанных в ELAN [Beck, Shannon 2020]. В отличие от более ранних проектов с аналогичными целями, SLEXIL — это веб-приложение без установки, разработанное строго на принципах, ориентированных на пользователя, от которого требуется минимальное количество технических знаний. Инструмент разработан для лингвистов, активистов языковых сообществ, учителей и других лиц, занимающихся вопросами устойчивого развития языков коренных народов. На выходе SLEXIL формирует интерактивную веб-страницу исходного многоуровневого XML-файла ELAN. Эта веб-страница представляет собой текст с аннотацией, которая была сделана в программе ELAN, с возможностью воспроизводить отдельные строки и с встроенным аудиоплеером для непрерывного воспроизведения записи. Во время непрерывного воспроизведения браузер выделит произносимую затранскрибированную строку и будет прокручивать экран, чтобы следить за воспроизведением по мере его продолжения. Далее эти страницы могут распространяться через Интернет для общего пользования, или по схеме защиты с паролем. Если требуется конфиденциальность и/или отсутствует подключение к Интернету, веб-страницу можно просмотреть как локальный файл в веб-браузере пользователя.

Еще одним широко используемым лингвистами программным обеспечением для анализа данных и составления словаря и корпуса является FieldWorks Language Explorer. Этот инструмент позволяет работать с языковыми данными в рамках документационного проекта: составлять словарь (с возможностью настраивать поля в словарной статье), проводить морфологический анализ текстов, составлять грамматический очерк. Возможен автоматический парсинг в соответствии с одобренными ранее решениями разбора. Кроме того, FLEx предлагает удобную систему для записи метаданных к материалам. Развивается и возможность совместной и удаленной работы нескольких участников проекта через функцию Send/Receive с регистрацией на сайте Languagedepot. Работать одновременно в онлайн-режиме над один проектом пока что

[20] Обучающее видео по применению этого скрипта доступно по ссылке: https://www.youtube.com/hashtag/lda2022.
[21] Описание данного программного обеспечения доступно на сайте TLA: https://archive.mpi.nl/tla/elan.

нельзя. Реализована возможность поиска по текстам по разным категориям данных. Таким образом, разобранные в данной программе тексты составляют корпус. Материалы, экспортированные из проекта, легко интегрировать в онлайн-корпус, развернутый на какой-либо платформе. Интерфейс выполнен на английском языке, но частично переведен и на некоторые другие. Пользователи могут участвовать в переводе интерфейса используя платформу Crowdin.

Еще одним инструментом для анализа данных и не только является программа SayMore, разработанная командой SIL International. Это приложение для сбора, организации, преобразования, архивирования и аннотирования данных. Оно также помогает отследить ваш прогресс и сделать транскрипцию аудио. Программное обеспечение SayMore можно назвать центральным инструментом для документационных проектов. Оно позволяет пользователю решать все основные задачи: аудиозапись, сбор метаданных, управление файлами, транскрибирование и перевод с синхронизацией по времени, а также упаковку для хранения в архиве. В то же время обеспечивается простой экспорт в приложения, которые выполняют более сложные задачи. Работа с программой требует минимального обучения для людей со средней компьютерной грамотностью.

FLEx и ELAN создают корпуса аннотированных текстов, но оставляют их в личных архивах исследователей. Инструмент LingView был разработан в ходе реализации проекта по документации языка кофан (англ. *A'ingae Language Documentation Project*, ALDP) как раз для того, чтобы результат анализа языковых данных в этих лингвистических программах можно было опубликовать в удобном виде в сети Интернет [Pride 2020]. Инструмент позволяет создавать веб-страницу, на которой представлен текст в аудио- или видеоформате вместе с разбором, выполненным в одной из программ. При этом разбор синхронизирован с медиафайлом в соответствии с разбивкой на аннотации или предложения.

В статье [van Esch et al. 2019] обсуждается, как исследователи в области машинного обучения могли бы помочь полевым лингвистам в сфере автоматической аннотации текстов. Это уменьшило бы нагрузку на последних, высвободив время для более глубокого лингвистического анализа и описательной работы. Авторы представляют примеры задач для компьютерных ученых; описание связанных с этим технологических проблем и предполагаемый уровень сложности; а также ссылки на соответствующую литературу. Кроме того, описывается ряд возможных последующих технических проектов, которые, по мнению авторов, было бы целесообразно и просто реализовать. Однако лингвисты, работающие в полевых условиях, уже могут использовать ряд существующих наборов инструментов, которые предлагают разные варианты автоматической транскрипции, чтобы ускорить процесс аннотирования. Такие наборы инструментов включают CoEDL Elpis [Foley et al. 2018], Persephone [Adams et al. 2018; Michaud et al. 2018] и SPPAS [Bigi 2015].

Для языков с небольшим объемом данных или для исследовательских ситуаций, которые запрещают использование облачных технологий, TAP (англ. *Transcription Acceleration Project* 'Проект ускорения транскрипции') разрабатывает Elpis — инструмент для получения транскрипции нерасшифрованного аудио с первого прохода. Результаты его работы лингвист может использовать в качестве чернового варианта и уточнять вручную. Elpis тренирует модель распознавания речи, используя файлы ELAN. Программа предлагает передовые технологии распознавания речи, доступные языковым специалистам и исследователям, не имеющим опыта работы в области речевой инженерии. Авторы задумывали Elpis как возможность для членов сообщества внести значительный вклад в расшифровку новых записей на их родном языке или вдохнуть жизнь в архивные материалы культурного наследия.

Persephone — инструмент автоматической фонематической транскрипции. Традиционные инструменты распознавания речи требуют большого словаря произнесений (описывающего, как произносятся слова) и большого количества обучающих данных, чтобы система могла научиться выводить орфографическую транскрипцию. Напротив,

Persephone предназначен для ситуаций, когда языковые данные, на которых можно производить обучение, ограничены, например, есть всего лишь час расшифрованной речи. Такие ограничения на данные распространены в документации миноритарных языков. К сожалению, данный ресурс не поддерживается сейчас регулярно.

Для создания фонетической транскрипции с использованием существующей многоязычной модели распознавания речи, можно использовать инструментарий аннотирования речи для малоресурсных языков (англ. *Speech Annotation Toolkits for Low Resource Languages*, SANTLR).

SPPAS предлагает кросс-платформенные, настраиваемые решения для автоматического аннотирования и анализа аудио- и видеоматериалов с открытым исходным кодом. SPPAS автоматически создает аннотации из записанного аудио или видеоматериала и его орфографической транскрипции; помогает с анализом аннотированных файлов (статистика, запросы, просмотр и редактирование файлов); производить конвертацию с широким спектром форматов: xra, TextGrid, eaf, trs и другие.

Однако исследователи не останавливаются в своих разработках в области инструментов лингвистического анализа. В последнее время заметна направленность на все большую автоматизацию работы лингвиста. Так, около года назад начала свое функционирование рабочая группа об NLP (англ. *Natural language processing* ‘Обработка естественного языка’) в полевой лингвистике (контактным лицом рабочей группы является О. А. Сериков). В рамках нее исследователи стараются выработать методологию хранения и доступа к полевым данным, предложить средства снижения количества рутины в обработке материалов, проводят прикладные разработки и исследования. В планах исследовательской группы сформировать сообщество инженеров и полевых лингвистов, которые понимают ограничения и перспективы технологий, и разделяют стремление к доступности исследовательских данных. Технологической целью является создание удобной платформы, обеспечивающей единообразный доступ ко всем языковым данным и к инструментам их обработки с открытым кодом. На пути к этой цели исследователи уже добились определенных результатов. Так, они создали унифицированный поисковик по всем корпусам, созданным на платформе tsacorpus, и провели соревнование по распознаванию полевой речи, ai4talk.

3.1.4. Платформы и приложения для собственно документации

Активные разработки архивных онлайн-платформ и онлайн-инструментов для доступной документации вместе со стремлением к развитию новых взаимоотношений между лингвистами и представителями небольших языковых сообществ и к поддержке документационных проектов, инициированных изнутри сообщества, приводит к созданию инструментов, ориентированных как на сбор языковых данных, так и на хранение и обеспечение доступа к ним.

Примером такого инструмента можно назвать онлайн-платформу «Varieties in Contact» (VinKo)[22], разработанную в сотрудничестве университетами Вероны, Тренто и Больцано. Совместный проект направлен на популяризацию и изучение языков и диалектов, на которых говорят в регионах Венето, Трентино-Южный Тироль и Фриули-Венеция-Джулия, посредством анализа и распространения лингвистических данных [Cordin et al. 2018; Kruijt 2022]. Платформа VinKo представляет собой региональную краудсорсинговую платформу для сбора и распространения языковых данных ограниченного набора языков и диалектов. Желающие внести свой вклад в коллекцию языковых данных после регистрации могут через функционал платформы сделать аудиозапись, отвечая на вопросы лингвистического опросника. Полученные аудиоданные становятся доступны всем пользователям через интерактивную карту, составленную максимально интуитивно понятно. Таким образом примеры речи открыты всем интересующимся для просмотра. Кроме того, есть отдельный тип доступа «для исследователей», который позволяет работать

[22] Платформа доступна по адресу: www.vinko.it.

с архивом целиком [Rabanus et al. 2022]. VinKo стремится объединить цели научных исследований с интересами членов языковых сообществ и предоставить открытый и удобный доступ к собранным данным.

Подобным примером является и платформа «Живые словари» (англ. Living Dictionaries) — бесплатный ресурс, на котором собираются и представляются словари языков, находящихся под угрозой исчезновения. На этой платформе любой желающий может создать словарь, в котором словарные статьи могут включать транскрипцию слова, написание в разных орфографических системах, переводы на несколько языков, примеры употребления, указание на часть речи, семантическое поле, морфосинтаксический анализ и другие метаданные, а также аудиофайл и изображение. Собранный словарь можно сделать открытым для всех пользователей сети Интернет.

В похожем проекте, Tatoeba, пользователи онлайн-платформы могут записывать различные высказывания (в текстовом и аудиоформате), в том числе относящиеся к повседневному общению. Одновременно, можно снабжать переводами записи других пользователей. Краудсорсинговым приложением для сбора и распространения процедурных текстов в аудиоформате можно назвать The Zahwa Project, которое позволяет всем пользователям записывать и сразу публиковать различные рецепты и процесс приготовления блюд с использованием фотографий и аудиозаписи. С точки зрения документации, таким образом собирается корпус звучащей речи в жанре процедурных текстов. С точки зрения ревитализации, это приложение поддерживает межпоколенческую передачу языка, расширяет сферы функционирования языка, создает контент на языке, объединяет языковое сообщество. Пользователи могут просматривать рецепты других, а также переводить их на свои языки. Все материалы публикуются под Creative Commons licence, с чем пользователь должен согласиться до начала работы с приложением. Есть возможность скачивать рецепты в виде архива с файлами. Приложение выпущено в бета версии. На еще одной платформе такого типа, Lingua Libre, собирается коллекция аудиозаписей с произнесением отдельных слов или коротких фраз на различных языках. К материалам организован открытый доступ. Также после регистрации любой пользователь может добавить свой материал. Реализована функция записи аудиофайла прямо в интерфейсе платформы. Аудиозапись снабжается расшифровкой в графике языка, а также метаданными о языке и дикторе. Доступ к аудиозаписям осуществляется через механизм поиска по параметрам: язык, диктор, пол диктора, статус диктора на платформе. Кроме того, есть функция экспорта выдачи запроса в формате csv.

Еще несколько приложений разработано для облегчения процесса аудиозаписи языковых данных и проведения проектов с несколькими участниками, в том числе в дистанционном формате. Так, SIL International активно разрабатывает приложение для мобильных устройств Language Forge (с англ. букв. ‘языковая кузница’). Это онлайн-инструмент для создания и редактирования словарей, которые синхронизируются со словарями в проектах программного обеспечения FieldWorks Language Explorer (FLEx). Словарные статьи могут включать аудиозаписи как на языке-источнике, так и на языке перевода. Приложение предлагает широкий функционал для совместной работы над одним проектом. Отлажен простой механизм синхронизации версий, а также возможность видеть изменения, вносимые другими участниками, практически в реальном времени. Кроме того, доступна функция комментирования, которая позволяет пользователям оставлять заметки и предлагать изменения в отдельных полях словарной статьи, при этом менеджер словаря может просмотреть каждый комментарий и принять его или добавить в список задач на будущее (функция «TODO»). Существует три категории доступа к словарю для зарегистрированных пользователей: наблюдатель, комментатор или соавтор. Language Forge доступен как через браузер, так и в виде приложения для смартфона (работает в операционных системах Windows, Mac, Linux, iOS и Android). Создаваемый в приложении словарь доступен только для участников проекта, но если воспользоваться связкой с FLEx

и Webonary, то собранный таким образом словарь можно сделать доступным в виде приложения для мобильных устройств.

Еще одним инструментом, разработанным специально для облегчения сбора языковых данных и для доступа к этому процессу всех заинтересованных, является LIG-Aikuma (букв. с языка усаруфа (папуасские языки) ‘место встречи’). Это бесплатное приложение с открытым кодом для устройств на Android. Приложение является в первую очередь инструментом для проведения аудиозаписи звучащей речи. Все записанные файлы хранятся в памяти того устройства, на котором установлено приложение. В приложении реализованы следующие методы работы: запись (в том числе спонтанной речи), перезапись (при необходимости исправить исходный файл или заменить его файлом лучшего качества), перевод (внесение перевода уже записанного фрагмента), элицитация (с письменным стимулом, изображением или видео, которые подгружаются в приложение пользователем). В функционал приложения входит перенос файлов в облако, передача по Bluetooth и другие способы выведения данных с устройства, на котором производилась запись.

3.1.5. Корпусные платформы

Важную роль в документационном процессе играет создание корпуса проанализированных текстов. Такой корпус позволяет широко использовать записанные в каком-либо проекте данные и проводить корпусные исследования различных тематик. Для того, чтобы представлять такой корпус в сети Интернет и обеспечивать возможность онлайн работы с ним, создаются корпусные платформы. Изначально они создавались для академических нужд, но сейчас разработчики учитывают необходимость одновременного представления текстов как для членов языковых сообществ, так и для всех интересующихся. Кроме того, корпусные платформы доступны бесплатно и с открытым кодом, так что каждый может доработать функционал под свои задачи.

Одной из таких корпусных платформ является часто используемая Tsakorpus [23], разработанная Тимофеем Архангельским. В корпусе, развернутом на этой платформе, можно делать сложные запросы по разным параметрам разметки, в том числе с дистанцией между словами, можно задавать поиск в подкорпусе, можно синхронизировать выдачу с аудио- или видеоформатом, можно создавать интерфейс на разных языках. Кроме того, при доработке загруженные тексты могут быть доступны пользователю для просмотра целиком, в том числе видео с субтитрами. При этом страница с такой коллекцией будет создаваться автоматически без дополнительной верстки.

Еще одна бесплатная корпусная платформа с открытым кодом — NoSketchEngine. Это упрощенная версия коммерческой платформы SketchEngine. В этой версии представлены базовые функции, необходимые для проведения корпусных исследований. Доступны возможности создания конкорданса и параллельных конкордансов, списка слов, выделения терминологии и вывода статистики по корпусу.

Еще одним примером корпусных разработок является кросс-платформенная (Linux, Mac, Windows) система ANNIS, позволяющая осуществлять поиск по корпусу текстов и визуализировать результаты. Система разработана для многоуровневых лингвистических корпусов с различными типами аннотаций. Сложные лингвистические явления, такие как информационная структура, взаимодействуют на многих уровнях, а ANNIS позволяет одновременно аннотировать, запрашивать и визуализировать данные из таких различных областей, как синтаксис, семантика, морфология, просодия, референциальность, лексика и многое другое. Система поддерживает аннотацию аудио- и видеоматериалов. В систему внедрены механизмы автоматического разбора и парсинга. Таким образом, ANNIS предоставляет средства для визуализации и извлечения этих данных. Для ее работы пользователю необходимо иметь только веб-браузер.

[23] На странице по ссылке: https://tsakorpus.readthedocs.io/en/latest/examples.html, можно посмотреть, какие корпуса развернуты на этой платформе.

Корпусная платформа Korp имеет открытый исходный код, и каждый может загрузить и настроить свой собственный корпус или внести свой вклад в разработку. Эта платформа позволяет осуществлять поиск одновременно по всем или выбранным корпусам, загруженным в систему.

3.1.6. Коллекции документационных материалов для широкой аудитории

Кроме доступных архивов и корпусов видится важным разрабатывать платформы, на которых пользователи смогут познакомиться с языком и культурой какого-либо сообщества в удобной форме. Материалами для таких сайтов могут стать документационные материалы в аудио- и видеоформате, снабженные расшифровкой и переводом. На данный момент ресурсов, где в такой форме были бы представлены именно собранные в ходе документации данные, мало. Это в том числе связано с проблемами с персональными данными и информированным согласием участников исследования на обработку и распространение своих данных, в том числе видеоизображения и другой личной информации. В то же время, кажется, что ценность и востребованность таких ресурсов велика, и это мотивирует научное сообщество решать возникающие этические и юридические вопросы.

Примером такого информационного ресурса является сайт «Малые языки Сибири: наше культурное наследие». Данный портал содержит мультимедиа материалы по языкам бассейна Среднего Енисея и Среднего и Верхнего Таза (селькупский, кетский, эвенкийский): социолингвистические данные, озвученные словари, аннотированные тексты, обучающие игры, документальные фильмы, видеозаписи и фотографии. Ресурс ориентирован на представителей языковых сообществ. Аннотированные тексты, записанные в ходе документационных проектов, представлены в видео- и аудиоформате и сопровождаются записью в графике языка, параллельным переводом на русский язык и поморфемной аннотацией, а также данными о рассказчике. В разделе со словарями можно осуществлять поиск слов по выбранному списку словарей.

Еще один интересный мультимедийный ресурс, предоставляющий доступ к коллекции материалов, — сайт «Chuklang». Этот ресурс ориентирован как на лингвистов, так и на представителей языкового сообщества и всех интересующихся языками Севера. На нем представлены результаты документационных проектов, проводившихся в селе Амгуэма Иультинского района Чукотского автономного округа. Помимо корпуса и грамматического очерка, доступны также тексты целиком в аудиоформате с поморфемной аннотацией. Аудиофайлы и файлы с аннотацией в формате pdf доступны для скачивания.

3.1.7. Использование архивных материалов

Немаловажной задачей как документации, так и лингвистики в целом является изучение письменных памятников языка ранних времен. Часто такие памятники хранятся в архивах, если и в оцифрованном виде, то в формате изображения. Таким образом, поиск нужных текстов напрямую по их содержанию невозможен. Чтобы это стало возможным, изображение с текстом (а часто это рукописный текст) необходимо превратить в файл в текстовом формате. Над автоматизацией этой задачи давно ведется работа, в том числе и со стороны разработчиков-лингвистов.

Примером информационного ресурса, который нацелен на обеспечение доступности архивных материалов для представителей сообщества и ученых, можно назвать сайт Ticha. На этом ресурсе размещены обширные архивные материалы на сапотекском языке (ticha происходит от сапотекского ‘слово, язык, текст’). Эти материалы содержат богатую лингвистическую, историческую и антропологическую информацию, но их чтение и интерпретация затруднены из-за специфики ранней сапотекской орфографии, отличающейся от современных лексики и грамматики. Ticha предоставляет пользователям доступ к данным и позволяет исследовать и анализировать их. Изображения рукописей сопровождаются транскрипцией, переводами на английский и современный испанский языки, лингвистическим анализом (включая поморфемную аннотацию) и комментариями.

Похожий информационный ресурс Digital Daisy Bates предоставляет доступ к архивам этнографа Дейзи Бейтс (1859–1951), работавшей в Австралии. В архивных материалах содержатся списки лексики с комментариями. Большая часть материалов рукописная. Команда создателей ресурса в сотрудничестве с Национальной библиотекой Австралии (NLA) оцифровала и перепечатала материалы и создала интуитивно понятный, удобный для исследования онлайн-архив.

Оба упомянутых ресурса обрабатывали исходные архивные материалы вручную. Но в проекте Nyingarn поставили задачу автоматизировать процесс распознавания и транскрибирования архивных материалов. Причем целью было создать такую открытую онлайн-платформу, на которой пользователи могли бы выполнять эти задачи для своих материалов, и таким образом пополнять архив материалов ранних лет, в том числе загружая все на платформу PARADISEC [Thieberger et al. 2022]. В широком смысле Nyingarn можно назвать рабочим пространством и хранилищем, где можно сохранять и распространять данные и вести совместные проекты. На платформу можно подгружать уже распознанные архивные материалы с оригинальными изображениями, а можно распознавать встроенными в платформу средствами. Данные снабжаются метаданными по одному из стандартов, чтобы легче встраиваться в поисковые системы. После проверки наличия всего необходимого, в том числе согласия представителей сообщества, данные помещаются в репозиторий платформы.

3.2. Онлайн-инструменты для овладения языком

Изучение языка с помощью компьютера (англ. *Computer-aided language learning*, CALL) начало зарождаться еще в середине прошлого века и с тех пор существенно продвинулось благодаря стремительному развитию информационных технологий. В настоящее время существуют сотни программ для изучения языка. Некоторые из них не просто используют компьютер как средство обучения, но опираются на современные достижения в области корпусной лингвистики и автоматической обработки языка. Такие системы могут быть названы интеллектуальными обучающими системами.

В этой области можно выделить сферу изучения языка с помощью мобильных устройств (англ. *Mobile-Assisted Language Learning*, MALL), опирающуюся на мобильные технологии. Основные характеристики мобильного обучения (m-learning) заключаются в том, что процесс обучения является персонализированным, спонтанным, неформальным и повсеместным. Хотя обучение с помощью мобильных телефонов может занять больше времени по сравнению с компьютерами, учащиеся чувствуют большую свободу в выборе времени и места. В отличие от традиционного обучения, в MALL учащимся не нужно присутствовать в классе или находиться за компьютером для участия в учебном процессе.

Приведем обзор некоторых исследований по данной проблематике. В работе [Miangah, Nezarat 2012] обсуждаются преимущества и недостатки, связанные с использованием мобильных технологий как для студентов, так и для профессионалов. Авторы рассматривают следующие области изучения языка с помощью мобильных устройств: словарный запас, аудирование, грамматику, фонетику и понимание прочитанного. В работе [Koole, Lewis 2020] поднимается глобальный вопрос о необходимости рассмотрения точки зрения коренных народов в технологическом образовании. В статье [Cassels, Farr 2019] авторы исследуют более узкую тему — использование мобильных приложений для изучения языков коренных народов. В работе приводятся обзор соответствующей литературы по теме и обзор 32 приложений, предназначенных для поддержки усилий по возрождению языков коренных народов в Канаде и на материковой части США. В работе [Begay 2013] исследуются новые разработки в области мобильных технологий и то, как они используются в сообществах коренных народов для изучения и преподавания языков. На основе анализа четырех мобильных приложений с помощью выработанных критериев оценки, онлайн-отзывов пользователей и консультаций с разработчиками автор определяет, какие основные темы важны при разработке эффективной и успешной модели

языкового приложения. В статье [Ovide, García-Peñalvo 2016] авторы описывают проект по изучению языка вичи (Аргентина), который включает в себя преподавание и изучение языка, совместную работу над языком вичи и создание мобильного приложения для изучения основ языка.

Далее приведены примеры онлайн-инструментов, ориентированных на представителей языковых сообществ и позволяющих разрабатывать образовательные ресурсы для различных языков.

3.2.1. Работа с текстами

Обучение языку часто включает в себя работу с текстами, которая может строиться на разных методиках, что позволяет практиковать определенные языковые компетенции. В этом разделе представлены онлайн-инструменты, реализующие некоторые из таких методик.

Ярким примером является обучающая платформа Revita, которая будет подробнее описана в Разделе 4.6. Во-первых, эта платформа позволяет работать с подгружаемыми пользователями текстами, а не с ограниченной библиотекой, собранной разработчиками. На основе текстов в режиме реального времени формируются разноформатные упражнения: заполнение пропусков в тексте (поставить слово в нужную форму, записать по произнесению, выбрать из вариантов, в форме кроссворда); работа с лексическими карточками (слово — перевод). Второй отличительной чертой платформы можно назвать персонализацию образовательного процесса, достигаемую средствами машинного обучения.

Другая платформа, Readlang, позволяет самостоятельно работать с текстом, определять и заучивать незнакомые слова, создавать словарик и флеш-карты для лексических тренировок. Пользователь имеет возможность подгружать свои тексты и работать с ними. Система реализована для многих языков с разными языками-посредниками. Требуется регистрация. Полный функционал доступен при оплате доступа. Необходимо установить приложение.

3.2.2. Словари

Неотъемлемым атрибутом обучения языку является, конечно же, словарь. Современные электронные словари далеко ушли по удобству и функционалу от печатных. А словари, доступные на мобильных устройствах, делают процесс обучения наиболее гибким. Кроме того, создание таких словарей сейчас доступно широкой аудитории, причем в основе могут лежать разные источники: как изданные словари, так и словари, собранные в ходе документационных проектов. Процесс оцифровки, распознавания и форматирования печатных словарей может быть полностью автоматизирован. Данные, полученные в результате обработки документационных материалов различными лингвистическими инструментами, также позволяют автоматически создавать мобильный словарь.

Так, программа Dictionary App Builder (DAB) позволяет создавать индивидуальные приложения словарей для смартфонов и планшетов на основе Android и IOS. Пользователь сам определяет, откуда взять файл со словником, название приложения, шрифты, описание программы, аудиодорожки, иллюстрации и иконки. DAB превратит все это в настроенное под пользователя приложение, которое затем можно будет установить на мобильное устройство или отправить другим по Bluetooth, с помощью карты памяти microSD, или загрузив приложение в интернет-магазины приложений.

Платформа Webonary дает возможность публиковать в нескольких форматах словарь из программы FLEx. Платформа автоматически преобразует загруженные словари (экспортированные из FLEx) в форматы pdf, Epub, apk. Есть возможность подгружать аудиозаписи и изображения. Необходимо заполнить форму-запрос на сотрудничество с платформой.

Про приложения Динары Стёпиной будет подробнее рассказано в Разделе 4.7. Она разработала и поддерживает приложения для словарей селькупского, ительменского,

талышского, ульчского и нанайского языков. Чаще всего приложение содержит не только словарь, но и разговорник. К разработчице можно обращаться, если есть желание сделать нечто подобное.

Одним из самых удобных способов создания словаря является бот в мессенджере Telegram. Такой словарь-бот может быть устроен просто, как чат-бот LuchadiBot Рустама Юсупова для русско-эвенкийского словаря: пишете ему слово на русском — получаете в ответ перевод на эвенкийский язык. А может быть устроен сложно, как чат-бот для англо-русского словаря StudyWords, и реализовывать следующие функции: интервальное повторение; изучение и понимание, в каких контекстах используется слово; конструирование предложений, а не просто заучивание карточек; выполнение упражнений с этими словами; периодические проверки, насколько усвоено слово (причем не просто кликая на карточку, а нужно будет написать перевод на русский и так же обратный перевод).

Онлайн-словари могут создаваться не только для звучащих языков, но и для жестовых. Например, словарь Dictio, созданный Люсией Влашковой, поддерживает как звучащие языки, так и жестовые. Доклад Люсии об этом на конференции по языковой документации и архивированию осенью 2022 года можно посмотреть по ссылке. Записи жестового языка доступны для поиска по нескольким параметрам: перевод на один из звучащих языков; фонологические параметры знаков (форма руки, место артикуляции, движение). Каждая запись устроена следующим образом: заголовок каждого знака содержит лицевое и боковое видео, его грамматические и региональные варианты, его социальное и региональное распространение, определение (в данном жестовом языке), примеры употребления (например, в коротких предложениях) и, наконец, переводы на несколько звучащих или жестовых языков. В настоящее время платформа поддерживает одиннадцать языков, большинство из которых жестовые, и содержит около 158000 записей.

3.2.3. Языковые курсы

Кроме отдельных компонентов процесса обучения языку, в приложениях могут быть реализованы комплексные языковые курсы для самостоятельного изучения языка. Сейчас такие курсы чаще всего построены на основе коммуникативного подхода и задействуют такие технологии, как автоматическое распознавание и порождение речи, машинное обучение. Далее приведем несколько наиболее популярных примеров таких приложений.

Бесплатные языковые курсы Duolingo ориентированы на учащихся начинающего уровня. Для разных языков обучения реализован разный набор языковых курсов. Курсы различных языков проработаны в разной степени. В максимальной версии содержатся такие типы упражнений: перевод слова (в обе стороны), перевод фразы, запись фразы на слух, перевод фразы на слух, произнесение слова или фразы; заполнение пропуска (в том числе на слух); работа с текстами. Задания разбиты по уровням владения, есть траектория процесса, задания более высоких уровней открываются после прохождения более простых. Есть возможность организации обучения в группе (есть функционал для педагога). Траектория курса персонализируется с использованием машинного обучения. Данные языковые курсы сфокусированы на разговорной практике и повседневном использовании языка. Кроме того, в миссию команды включены идеи поддержки и развития языков малочисленных сообществ. Команда Дуолинго открыта к сотрудничеству, если вы хотите создать курс по вашему языку!

Платформа Memrise обеспечивает изучение языка через видео на актуальные темы, подгруженные на платформу носителями. Есть вариант с языком-посредником русским. Нужна регистрация. Режимы обучения, доступные на сайте: учить новые слова; обычное повторение; быстрое повторение; трудные слова; навыки аудирования; учись с носителями языка (доступно для некоторых курсов). У приложения диапазон режимов шире. На платформе представлены и некоторые языки России: чеченский, чувашский, татарский, даргинский, якутский, лакский и другие.

Команда проекта 7000 Languages в сотрудничестве с небольшими языковыми сообществами создает бесплатные онлайн-курсы по изучению языков. Проект нацелен на поддержку языкового активизма и сохранение языкового многообразия мира. Для создания курсов используется программное обеспечение партнера «Transparent Language». Курсы доступны бесплатно после регистрации. Можно обратиться к команде проекта с инициативой создания обучающего курса по своему языку. Лекцию Стефани Витковски, исполнительного директора проекта, о некоторых бесплатных возможностях создания языковых ресурсов и о том, как создаются курсы изучения языков в 7000 Languages, можно посмотреть по ссылке.

3.2.4. Инструменты для учителей

В этом разделе представлены некоторые инструменты для преподавания языков, которыми могут воспользоваться учителя. Эти инструменты не специфичны для миноритарных языков, они могут использоваться в преподавании любого языка.

Платформа TeacherMade предоставляет возможность создания интерактивных упражнений на основе загруженных материалов. Есть автоматическая проверка и оценивание.

Платформа Wizer.me позволяет создавать упражнения с передвижением элементов, соединением текста и разделением его на фрагменты и прочее. Создавать разнообразные задания на разные темы становится просто. Подойдет для онлайн или образования через компьютеры.

Платформа Storybird помогает научиться создавать тексты разных жанров, в том числе эссе, рецензии и другие. Рассчитана на широкую аудиторию. Есть специальный функционал для педагогов, практикующих с учениками навыки письма. Платформа реализована также в формате мобильного приложения.

Открытая платформа LearningApps предназначена для создания разных заданий. Доступная и понятная. Можно адаптировать шаблоны под любые задачи. Можно создавать свои упражнения-игры по предоставленным шаблонам (21 шаблон). Например, «Найди пару», «Кроссворд», «Выбор ответа», «Хронологическая линейка», «Ввод текста», «Заполнить пропуски» и другие. Создавать задания можно бесплатно и без регистрации. Можно выкладывать свои задания в открытый доступ.

3.2.5. Примеры платформ с обучающими материалами

В заключительном разделе из сферы образования мы рассмотрим конкретные примеры платформ, на которых представлены уже готовые обучающие материалы и инструменты для их создания.

Платформа FirstVoices представляет собой коллекцию мультимедийных инструментов для обучения языку, направленную на языки коренных народов. Для каждого языка есть раздел по изучению (в том числе, с аудио), раздел с играми, фотогалерея и детский портал. Материалы по разным языкам доступны и в виде приложений (для Android и iOS). Есть приложение с раскладками для языков Канады, Австралии, Новой Зеландии и немного США. Проект сотрудничает с 70 сообществами.

Проект Mobile Language Team посвящен обучению миноритарным языкам Австралии. На платформе есть обучающие уроки, карточки со словами, списки слов, обучающие видео для восьми языков. Помимо этого, представлена коллекция видео и книг.

Проект Tusaalanga посвящен изучению диалектов эскимосского языка инуктитут. Уроки построены на основе озвученных диалогов на разные темы с грамматикой и озвученным словарем (два произнесения, мужским и женским голосом). Можно работать с грамматикой или словарем отдельно. Есть озвученные диалоги на бытовые темы. В уроках лексика и грамматика подобраны под диалог.

На сайте Learn Manx собраны различные ресурсы по сохранению и развитию мэнского языка. В одном из разделов собрана информация об образовательных организациях, где проводится обучение мэнскому языку (в том числе дошкольного уровня). Несколько

разделов посвящены онлайн-ресурсам по мэнскому языку: онлайн-словари, в том числе озвученные; статьи по грамматике; оффлайн и онлайн курсы мэнского языка разных уровней; приложения и платформы для самостоятельного изучения; ресурсы, на которых можно послушать мэнский язык (или посмотреть видео на мэнском). Есть раздел, посвященный сообществу, организациям и общественным инициативам по поддержке мэнского языка.

Уже несколько лет братья Тоби Адамс и Дарел Догерти занимаются документацией языка своих предков каллилли, в том числе с участием лингвиста Клэр Бауэрн и студентов Йельского университета. Они создали словарь в виде приложения для смартфона (инструкция по его использованию есть в этом видео), ведут блог и подкаст о том, как изучать этот язык. Все разработанные образовательные материалы собраны на сайте проекта Kullilli Ngulkana в разделе «Ресурсы».

Еще одним примером является веб-сайт Blackfoot language resources, посвященный языку блэкфут. Он построен на основе цифрового словаря, а также содержит грамматические зарисовки, библиотеку других ресурсов, связанных с языком, и архив рассказов. В работе [Genee, Junker 2018] подробно описывается этот проект со всеми этапами его реализации.

Что касается российской почвы, то интересным примером служит страница проекта «Компьютерные и лингвистические ресурсы для поддержки шугнанского языка», который реализуется студентами Школы лингвистики НИУ ВШЭ и Института классического Востока и античности. Проект имеет отчетливо практическую направленность, в его рамках реализуются следующие задачи: разработка электронной версии «Шугнанско-русского словаря» Д. Карамшоева; разработка программы морфологического анализа шугнанских словоформ; создание электронного корпуса шугнанских текстов; оцифровка шугнанских текстов (литература, фольклор, экспедиционные записи); выпуск шугнано-русско-английского разговорника. В рамках проекта уже запущен Telegram-бот с электронной версией «Шугнанско-русского словаря» Д. Карамшоева (разработчик бота — Ю. Макаров).

3.3. Онлайн-инструменты для создания контента на языке

Для сохранения и развития языков важно расширять сферы их функционирования, если в каких-то из них язык не используется. Использование языков в онлайн-пространстве уже давно стало привычным, но далеко не все языки широко в нем представлены или вообще используются. В то же время обеспечение такой возможности для всех языков является важной задачей в ревитализационных процессах.

3.3.1. Инфраструктура

Чтобы языки были представлены в цифровом формате и в онлайн среде, важно создавать необходимые для этого составляющие, например, шрифты и раскладки клавиатуры. Эти задачи все больше и больше становятся доступны пользователям компьютеров без специального образования благодаря разработкам удобных инструментов и распространению обучающих видео об их использовании. Таким образом, все больше возможностей получают представители языковых сообществ для реализации своих инициатив по созданию нужных элементов цифровой среды самостоятельно.

Так, SIL International разработал бесплатную программу с открытым исходным кодом Keyman Developer, которая позволяет создавать раскладку клавиатуры для своего языка. Keyman Developer доступен для многих платформ, включая Windows, macOS, iOS, Android, Linux и web. Инструмент обеспечивает возможность интегрирования пользовательских раскладок клавиатуры в веб-сайты и приложения и создания раскладки для всех основных операционных систем и устройств. Редактор пакетов Keyman Developer позволяет объединить в одном установщике все элементы клавиатуры, которые нужны любому пользователю для набора текста. Кроме того, в один пакет можно включить все необходимые шрифты, которые помогут пользователям печатать на языке на своем

компьютере. Раскладки, созданные в этом инструменте, можно интегрировать на мобильные устройства на основе Android с помощью приложения Keyboard App Builder. Приложение позволяет печатать на своем языке и в других приложениях, таких как Gmail, Facebook и WhatsApp. Keyboard App Builder устроен интуитивно понятно, снабжен подсказками на всем пути создания и установки клавиатуры. Весь интерфейс будущего приложения может быть переведен и локализован. Для распространения приложения не обязательно подключение к сети Интернет. Созданное приложение не требует доступа к интернету или к проводнику файлов.

Программное обеспечение Graphite, разработанное также командой SIL International, позволяет создавать так называемые «умные шрифты», способные отображать системы письменностей с разнообразными и сложными правилами построения. Инструмент был создан для решения задач в рамках документационных и ревитализационные проектов для отображения языковых данных в графике языков, которые не распространены широко в цифровой среде. «Умный шрифт» включает в себя не только собственно символы, но и дополнительные указания относительно того, как сочетать и располагать эти символы в пространстве. Примеры сложных правил построения шрифтов, которые возможны в Graphite: зависимость начертания от контекста; лигатуры; смена порядка; расщепление глифов; два направления письма; комбинирование диакритических знаков.

3.3.2. Контент и общение на языке в сети Интернет

Наличие контента и общения на языке в сети Интернет, является естественным для широко распространенных языков. В то же время, для небольших языковых сообществ эта сфера может быть не охвачена. Таким образом, требуются дополнительные целенаправленные усилия для того, чтобы увеличивать представленность языка в сети Интернет за счет создания оригинального контента или перевода уже существующего. Во многих случаях в языковых сообществах есть активный запрос на это. Современные технологии делают и этот процесс все более доступным для широкого круга людей, заинтересованных в поддержке и развитии своих языков. Остановимся на некоторых примерах доступных возможностей.

Платформа Bloom предоставляет пользователям возможность переводить на свой язык и озвучивать различные книги, а также читать опубликованные издания других пользователей. Сейчас на платформе есть книги на 540 языках мира. Кроме того, публикация книги на платформе требует заполнения шаблона об авторских правах и прочих данных исходной книги, для того чтобы после выхода книга была безупречно оформлена с правовой точки зрения. Интерфейс для чтения книг выполнен очень привлекательно и удобен для использования: при перелистывании страницы начинает проигрываться соответствующий аудиофайл, а на странице выделяется тот фрагмент, который в данный момент звучит. Кроме того, стоит отметить, что в этом интерфейсе визуализирована книга целиком, с иллюстрациями и с анимацией переворачивания страниц.

Другим инструментом, позволяющим расширить представленность языка в сети Интернет, является широко используемая платформа-видеохостинг YouTube (подробнее об этом см. Раздел 4.9). В рамках данной платформы разработан функционал, достаточно удобный для расшифровки и перевода текстов в аудио- или видеоформате с синхронизацией аннотаций с медиаконтентом [Rice 2021]. При этом не нужно устанавливать никакого программного обеспечения и разбираться в его функционале. Созданные таким образом расшифровку и перевод можно подключать к видео в качестве субтитров. Таким образом, это менее трудоемкая альтернатива озвучиванию видео, что часто недоступно пользователям. В то же время, в некоторых случаях, при наличии определенных ресурсов и договоренностей с автором, можно действительно переводить и озвучивать медиаконтент на другом языке. Об одной из таких возможностей, возможности перевода мультсериалов «Гора самоцветов» и «Лео и Тиг», подробно говорится в Разделе 4.8.

Большая доля общения в сети Интернет происходит в социальных сетях и мессенджерах. Дистанционное общение может играть ключевую роль для сохранения некоторых языков, на которых говорит небольшое количество людей, живущих удаленно друг от друга. При создании такой среды можно использовать те же популярные инструменты, что используются для распространенных языков: стикерпаки (наборы стикеров) в мессенджерах, мемы в социальных сетях, облака тегов, различные редакторы, работающие с текстами и изображениями. Например, в мессенджерах Telegram и WhatsApp любой пользователь может создать стикерпаки, содержащие короткие высказывания на языке, и это не требует вложения больших ресурсов[24]. Платформа Word Art позволяет создавать облака тегов разных форм, хотя есть ограничение по распознаваемым символами. Другой пример подобного инструмента — Funny Newspaper Generator with Your Own Picture — представляет собой редактор, в котором можно создать имитацию газетной страницы с новостями. Похожим образом iFake Text Message позволяет создать имитацию переписки в чате. А на платформе Meme Generator — Imgflip можно создавать мемы, в том числе на основе тех, что уже стали популярными. Во всех этих видах развлекательного контента в большей или меньшей мере может присутствовать поддерживаемый язык.

3.3.3. Пример платформы с коллекцией материалов на языке

Помимо инструментов для создания контента, существуют платформы и информационные ресурсы для хранения и распространения разного рода материалов на языке. Это могут быть коллекции медиаконтента или материалов в текстовом формате, как оцифрованных, так и изначально созданных в электронном виде. Для опубликованных материалов оцифровка и такое представление в сети Интернет могут обеспечить видимость и доступность. Таким образом, языковые данные получают «новую жизнь»: представители языковых сообществ и все заинтересованные получают возможность найти и использовать эти материалы, которые иначе могли бы быть просто забыты. Одним из примеров, когда такая онлайн-библиотека и сам процесс ее создания сыграли важную роль в сохранении и развитии языка, можно назвать опыт проекта Ulukau, посвященного гавайскому языку [Galla 2018]. Эта платформа создана Колледжем гавайского языка (Гавайский университет в Хило) и нацелена на обеспечение доступности гавайских материалов различных жанров: языковые материалы для использования в школах с гавайским языком обучения, материалы для создания печатных СМИ, радио, телевидения, социальных и цифровых медиа. Кроме того, этот ресурс стал площадкой для общения и объединения представителей гавайского сообщества. На ресурсе представлен и раздел «Словари» с возможностью поиска по нескольким словарям одновременно. Есть разделы с книгами по разным уровням владения языком, а также с газетами. Также есть коллекция аудио- и видеофайлов со звучащей речью на гавайском. Любой пользователь платформы может внести свой вклад и расшифровать какую-либо запись из коллекции. На платформе много и другой информации, связанной с культурой и историей Гавайских островов.

4. Методики создания ориентированных на языковые сообщества онлайн-инструментов поддержки языков

Ниже приводится методология создания онлайн-инструментов для поддержки языков. Мы связались с разработчиками, обсудили принципы устройства и работы созданных ими инструментов и уточнили, что нужно для того, чтобы этими инструментами воспользоваться. Читатели смогут получить ответ на вопрос, что можно сделать с языковыми данными, которые у них имеются, а также какие данные нужны, чтобы реализовать какую-то конкретную цель. Для всех инструментов указаны контактные лица, к которым можно обратиться за помощью. Мы специально выбрали онлайн-инструменты

[24] На эту тему готовится статья А. А. Воронковой «Стикеры на языках РФ. История возникновения и коммуникативные задачи».

разных типов (в соответствии с классификацией из Раздела 3): инструменты для документации (ориентированные как на лингвистов, так и на представителей сообщества), для овладения языком, для создания контента.

4.1. Лингвистические корпуса текстов

Ссылка	Корпуса. Малые языки России
Тип ресурса	Платформа для документации языков
Для кого	Лингвисты, языковые активисты, педагоги

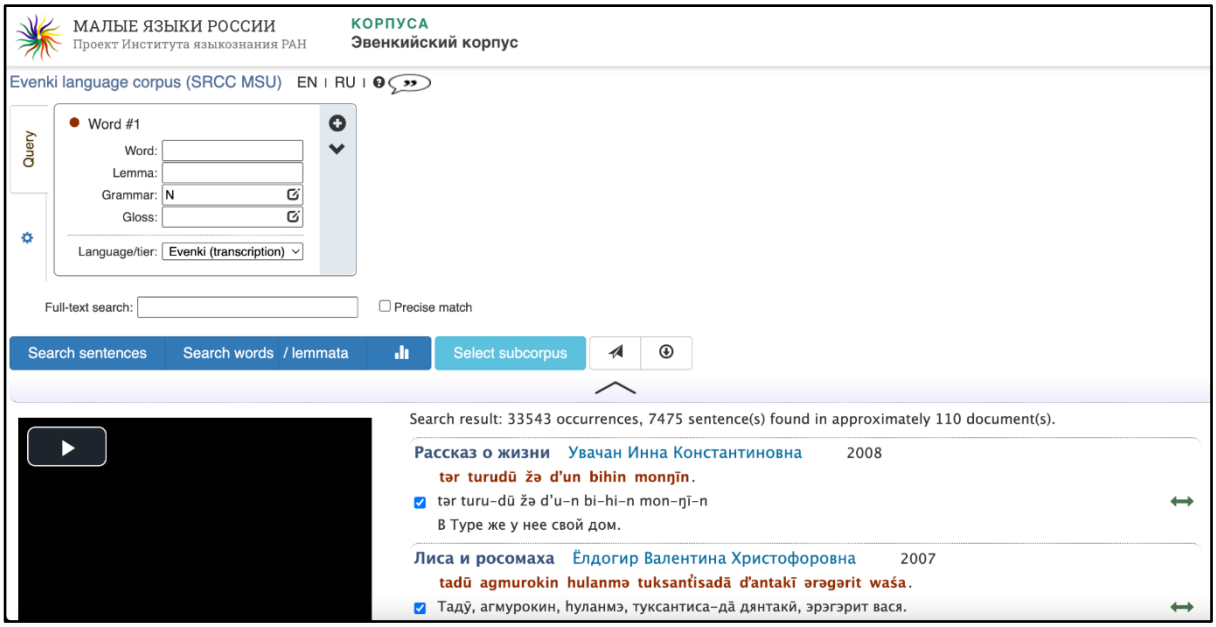


Рисунок 2. Поисковая выдача эвенкийского корпуса по запросу «существительное»

«Корпуса» — платформа, на которой размещаются языковые корпуса. Проект создается в Лаборатории исследования и сохранения малых языков Института языкознания РАН в рамках разработки сайта «Малые языки России».

Основная часть корпусов представлена на программном решении Tsakorpus, но некоторые корпуса интегрированы напрямую в среду платформы, поскольку существуют на самостоятельных программных решениях. Кроме того, на данный момент некоторые корпуса представлены исходными файлами программы SIL FieldWorks или pdf-файлами. Но авторы проекта работают над их интеграцией в среду и снабжением удобным и интуитивно понятным веб-интерфейсом для работы с ними.

На данный момент доступны корпуса для пяти миноритарных языков России и мира.

Корпус ительменского языка состоит из 15 архивных ительменских текстов, записанных В. И. Иохельсоном в 1910–1911 годах и А. П. Володиным в 1962–1973 годах. Морфологическое аннотирование выполнено К. О. Шейфер, С. К. Ганиевой и М. Р. Плугарёвым. Программная часть разработана Максимом Бажуковым.

В корпус кетского языка входят тексты на трех диалектах кетского языка из мультимедийного архива Лаборатории автоматических лексикографических систем НИВЦ МГУ им. М. В. Ломоносова и Лаборатории исследования и сохранения малых языков ИЯз РАН, записанные в 2002–2014 годах в ходе экспедиций по документации кетского языка под руководством О. А. Казакевич, а также архивные тексты, записанные Г. М. Корсаковым в 1937 году. Морфологическое аннотирование выполнено Ю. Е. Галяминой и Е. М. Будянской.

Материал корпуса горномарийского языка собран коллективом экспедиции под руководством Е. В. Кашкина по изучению горномарийского языка на кафедре Теоретической и прикладной лингвистики филологического факультета МГУ имени М. В. Ломоносова. В состав корпуса также входят тексты В. А. Акцорина [25] и Г. А. Саватеевой [26]. Общий объем текущей версии корпуса составляет 63522 словоупотребления.

В корпус эвенкийского языка входят тексты на северных, южных и восточных диалектах эвенкийского языка из мультимедийного архива Лаборатории автоматических лексикографических систем НИВЦ МГУ им. М. В. Ломоносова и Лаборатории исследования и сохранения малых языков ИЯз РАН, записанные в 1998–2021 годах в ходе экспедиций по документации эвенкийского языка под руководством О. А. Казакевич, а также архивные эвенкийские тексты, записанные Г. М. Василевич в 1930–1950 годы и Е. А. Лебедевой в 1950–1960 годы Морфологическое аннотирование выполнено в основном Е. Л. Клячко при участии Н. К. Митрофановой.

Корпус языка куллуи создан научным коллективом, занимающимся документацией индоарийского языка куллуи (штат Химачал-Прадеш, Индия), в составе Е. А. Ренковской (ИЯз РАН), Ю. В. Мазуровой (ИЯз РАН) и А. С. Крыловой (ИВ РАН). Программная часть корпуса разработана Е. В. Коровиной (ИЯз РАН). На данный момент в корпус входят спонтанно порожденные, а также элицитированные тексты на центральном диалекте куллуи, записанные в 2014–2017 годах в ходе экспедиций в округ Куллу (в деревни Наггар, Башинг, Тхава, Сума).

Планируется развитие общей среды для отдельных корпусов и интеграция их в эту среду, а также расширение функционала корпусов, в том числе синхронизация выдачи с аудиоформатом, возможность просмотра текста целиком и другие. Проект «Малые языки России» открыт к сотрудничеству и готов принимать материалы для размещения на своей платформе.

Если Вы хотите, чтобы Ваша коллекция текстов была представлена в проекте «Корпуса»:	
Контакты	Лаборатория исследования и сохранения малых языков ✉ minlanglab@iling-ran.ru
Что для этого нужно	● Коллекция текстов с поморфемной аннотацией в формате FLEx или ELAN (или уже готовый корпус на какой-либо программной платформе)

4.2. Приложение Language Forge

Ссылка	Language Forge
Тип ресурса	Приложение для документации языков
Для кого	Лингвисты

Language Forge (с англ. букв. ‘языковая кузница’) — это онлайн инструмент для создания и редактирования словарей, которые синхронизируются с словарями в проектах программного обеспечения FieldWorks Language Explorer (FLEx) [27]

[25] Акцорин В. А. Марийский фольклор. Мифы, легенды, предания. Йошкар-Ола, 1991.

[26] Саватеева Г. А. Лексические особенности правобережных говоров горномарийского языка. Дисс. … к. ф. н. Йошкар-Ола, 2005.

[27] Программное обеспечение FLEx позволяет создавать словарь и корпус текстов с поморфемной аннотацией для какого-либо языка. Кроме того, в проекте можно прописать основные грамматические характеристики языка и составить очерки грамматики. Эта среда активно развивается разработчиками. Налаживается

(с англ. букв. ‘полевой исследователь языков’). Словарные статьи могут включать аудиозаписи как на языке-источнике, так и на языке перевода. Приложение предлагает широкий функционал для совместной работы над одним проектом. Отлажен простой механизм синхронизации версий, а также возможность видеть изменения, вносимые другими участниками, практически в реальном времени. Кроме того, доступна функция комментирования, которая позволяет пользователям оставлять заметки и предлагать изменения в отдельных полях словарной статьи, при этом менеджер словаря может просмотреть каждый комментарий и принять его или добавить в список задач на будущее (функция «TODO»). Существует три категории доступа к словарю для зарегистрированных пользователей: наблюдатель, комментатор или соавтор. Language Forge доступен как через браузер, так и в виде приложения для смартфона (работает в операционных системах Windows, Mac, Linux, iOS и Android).

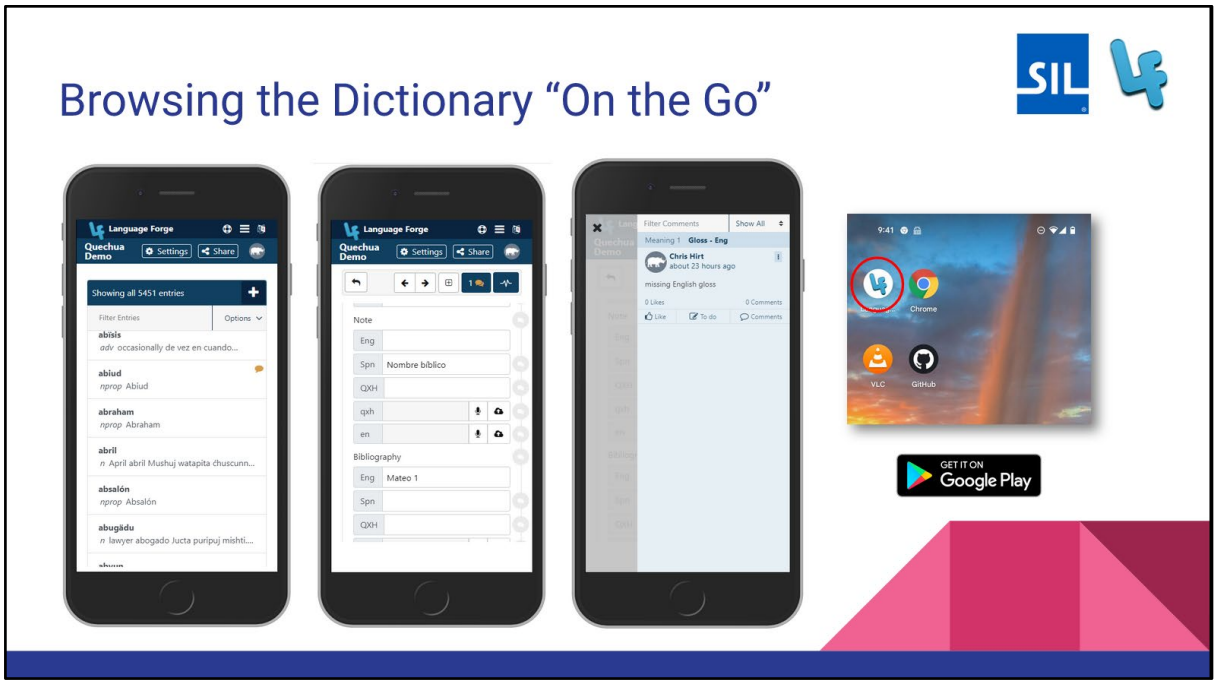


Рисунок 3. Некоторые страницы приложения Language Forge.

Language Forge разрабатывается лингвистами и программистами из Университета Пайап (Тайланд, Payap University) и Международного летнего института лингвистики (SIL International) под руководством Кристофера Хёрта.

Интерфейс переведен на шесть языков: английский, французский, индонезийский, кхмерский, киньяруанда и испанский. Для участия в переводе интерфейса можно воспользоваться следующей инструкцией.

Приложение регулярно обновляется, последний на момент написания данного текста релиз вышел 4 октября 2022 года. Все нововведения из обновлений перечисляются в соответствующей рубрике на сайте сообщества. Кроме того, несколько видео на эту же тематику доступны в плейлисте Language Forge на YouTube. Различные технические пояснения и ответы на многие вопросы можно найти в блоге сообщества приложения на сайте SIL International. Известные проблемы и ограничения по использованию ресурса перечислены здесь. На YouTube доступно также видео 2021 года с пошаговым объяснением того, как устроен Language Forge.

возможность совместной работы над проектами, онлайн доступность, возможность подгрузить аудио- и видеоматериалы. Подробнее о FLEx можно узнать на соответствующей странице SIL International по адресу: https://software.sil.org/fieldworks/.

Если Вы хотите создавать словарь и работать с коллегами в приложении Language Forge:	
Контакты	Форум сообщества Language Forge
Что для этого нужно	● Техническое оборудование с возможностью аудиозаписи и с подключением к сети Интернет (ноутбук, смартфон, планшет, компьютер) ● Языковые материалы в аудио и текстовом формате или в формате FLEx

4.3. Платформа ЛингвоДок

Ссылка	Lingvodoc
Тип ресурса	Платформа для хранения языковых материалов и работы с ними
Для кого	Лингвисты, языковые активисты, педагоги

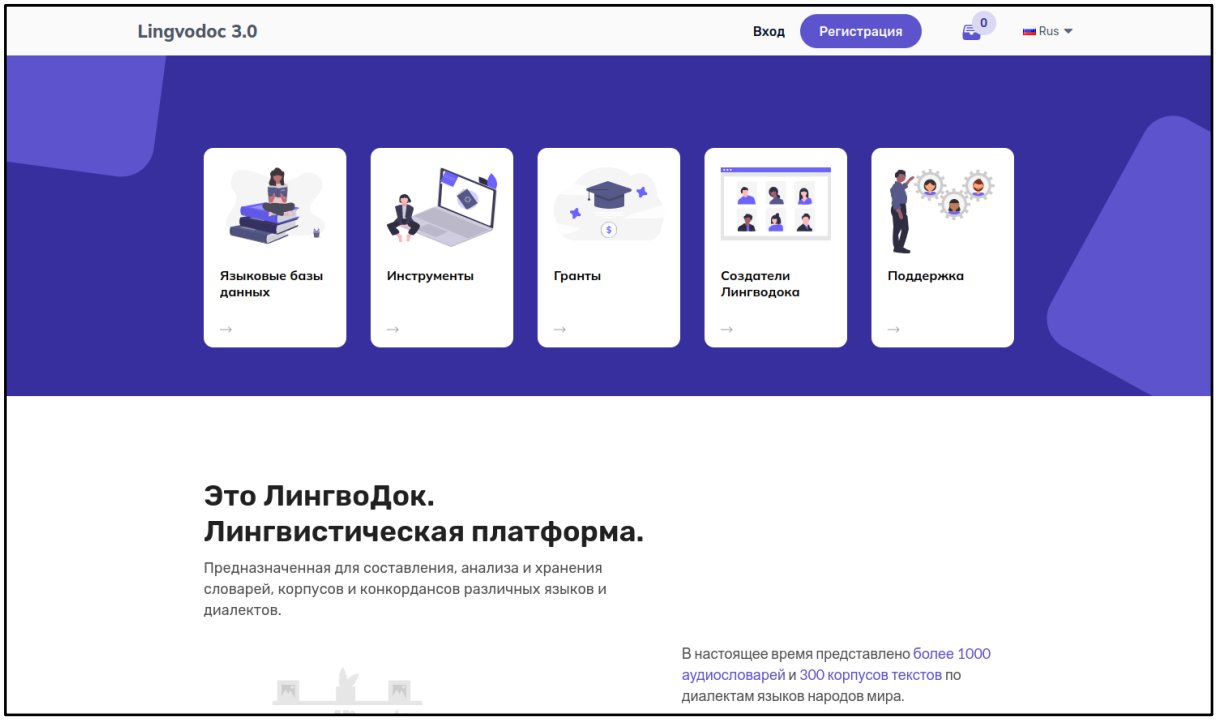


Рисунок 4. Главная страница платформы ЛингвоДок

Это платформа для хранения и анализа языковых материалов, которая позволяет работать над ними нескольким пользователям одновременно. Кроме того, в ее функционал входит создание учебных материалов из текстовых корпусов. На платформе выложены словари, в том числе озвученные, и корпуса текстов на многих языках России, в особенности уральских и алтайских. ЛингвоДок постоянно развивается и расширяет свой функционал, а языковые базы данных постоянно пополняются новыми материалами.

Платформа разрабатывается коллективом Лаборатории «Лингвистические платформы» ИСП РАН им. В. П. Иванникова под руководством д.ф.н. Ю. В. Норманской.

Ключевыми особенностями платформы ЛингвоДок являются возможности совместной работы с архивами и лингвистического анализа языковых данных [Дыбо, Норманская 2021]. В ЛингвоДок создано 13 программ для анализа языковых данных

с фонетической, морфологической и этимологической точек зрения (хотя только 4 из них доступны для всех пользователей системы). Подобный анализ ранее выполнялся вручную лингвистами, а система позволяет сделать это гораздо быстрее.

Система поддерживает и позволяет просматривать и обрабатывать следующие типы данных: текст, изображение, звук (wav, mp3 и flac), разметка (форматы ELAN и Praat). На платформе пользователи могут создавать многослойные словари, прикреплять их к географической карте, снабжать их метаданными, делиться доступом к словарям с другими пользователями. Система позволяет пользователям задавать структуру словаря, включая в словарную статью необходимые в каждом случае типы информации: аудиофайл с произнесением слова, транскрипцию, написание в орфографии языка, диалектные пометы, этимологические сведения и другие. Кроме того, каждый аудиофайл можно соотнести с его анализом средствами программного обеспечения Praat. Инструменты работы со словарями позволяют формировать по загруженным данным инвентарь фонем, получать результаты анализа спектрограмм в различных форматах, измерять степень сходства фонетических систем по разным языковым материалам [Kosheliuk 2021b, 2021d; Normanskaja 2018; Normanskaja et al. 2022]. Словарные статьи в разных словарях при необходимости можно связывать друг с другом, через функцию «когнаты» [Kosheliuk 2021e]. Такие связи между словарями позволяют с помощью автоматических алгоритмов выявлять в онлайн режиме фонетическое сходство языков, употребление тех или иных морфологических параметров в определенном значении. Кроме того, загружая и пополняя словари, можно автоматически проверять и удалять возможные дубликаты словарных статей.

Что касается текстов, система обеспечивает хранение, просмотр и обработку корпусов в формате программного обеспечения ELAN [Kosheliuk 2021c; Норманская 2022]. На материале словарей и корпусов можно строить карты фонетических, морфологических или лексических изоглосс в синхронии и диахронии [Kosheliuk 2021a].

В последнее время в рамках сотрудничества с различными республиканскими ВУЗами на платформе ЛингвоДок создаются глоссированные корпуса со снятой омонимией на башкирском, якутском, мордовских и удмуртском языках, которые в будущем лягут в основу обучающих платформ. В планах у создателей сформировать обучающую среду, в которой будет возможно создание большого количества упражнений для любого уровня владения языком (в том числе, что немаловажно, для уровня владения выше среднего) [Норманская 2021]. Упражнения не будут повторяться и будут предлагаться пользователю с учетом его индивидуального прогресса в изучении языка. Учителя также смогут зарегистрироваться в системе, чтобы отслеживать статистику по своим ученикам.

На платформе ЛингвоДок реализована система с различными уровнями доступа к данным [Kosheliuk 2021f]. Размещенные материалы могут быть доступны только определенным зарегистрированным пользователям или любому посетителю ресурса, в том числе с возможностью скачивания данных. При этом создатель словаря или корпуса может давать коллегам различный набор возможностей редактировать данные и проект.

Авторы проекта регулярно проводят курсы повышения квалификации «Возможности платформы ЛингвоДок для описания языков народов России». Участники курсов имеют возможность на основе функционала ЛингвоДока проанализировать свой языковой материал и получить готовые результаты. Кроме того, на сайте ЛингвоДока есть целый раздел с большим количеством обучающих материалов (https://github.com/ispras/lingvodoc-react/wiki). Пользователи также могут получить оперативную техподдержку через мессенджер Telegram.

Если Вы хотите, чтобы Ваши документационные материалы были представлены на платформе ЛингвоДок:	
Контакты	al@somestuff.ru группа в Telegram
Что для этого нужно	• Языковые материалы в каком-либо цифровом формате (docx, xlsx, eaf, textgrid, wav, mp3, flac и др.)

4.4. Платформа Living Dictionaries

Ссылка	Living Dictionaries
Тип ресурса	Платформа для документации лексикона
Для кого	Представители языкового сообщества, языковые активисты, лингвисты

Рисунок 5. Главная страница платформы Living Dictionaries

Платформа «Живые словари» (англ. *Living Dictionaries*) — это бесплатный ресурс, на котором представлены словари языков, находящихся под угрозой исчезновения. На этой платформе любой желающий может создать словарь, в котором словарные статьи могут включать транскрипцию слова, написание в разных орфографических системах, переводы на несколько языков, примеры употребления, указание на часть речи, семантическое поле, морфосинтаксический анализ и другие метаданные, а также аудиофайл и изображение [Anderson, Daigneault 2021a, 2021b].

Платформа была запущена в 2006 году лингвистами Института поддержки исчезающих языков «Живые языки» (англ. *Living Tongues Institute for Endangered Languages*) Грегори Д. С. Андерсоном и К. Дэвидом Харрисоном и называлась изначально «Говорящий Словарь» (англ. *Talking Dictionary*), см. [Anderson et al. 2012; Harrison et al. 2019]. В 2020 году институт изменил название платформы на «Живые словари».

«Живые словари» созданы как свободный и доступный инструмент, позволяющий заниматься документацией исчезающих языков не только лингвистам, но и языковым активистам, поддерживая, таким образом, инициативы по сохранению и развитию языков, возникающие в языковых сообществах. Словари на этой платформе представляют собой совместные мультимедийные проекты, возглавляемые общественными активистами

из разных уголков планеты, которые ориентированы на сохранение и передачу языков последующим поколениям.

Платформа функционирует в любом интернет-браузере на любом компьютере или мобильном устройстве с операционной системой Android или iOS. Менеджеры и участники проекта могут создавать словари, управлять ими, редактировать их, используя только смартфоны или планшеты, которые могут функционировать как полноценные рабочие инструменты для создания аудиозаписи, ввода лингвистических данных и других мультимедиа. Каждый словарь может быть общедоступным или частным, в зависимости от того, какой уровень доступа предоставит создатель словаря.

Сайт переведен на 13 языков: английский, ассамский, бенгальский, французский, иврит, суахили, хинди, португальский, русский, испанский, малайский, китайский, индонезийский.

В настоящее время для словарей реализованы следующие особенности: управление словарем (просмотр и редактирование словарных статей в режиме реального времени); запись и воспроизведение аудио; загрузка фотографий; оффлайн-доступ к данным; поиск; каждой словарной статье приписан номер; для словарных статей введены семантические поля; есть возможность настроить частный и общий доступ к словарям; можно воспользоваться экранной клавиатурой при необходимости; функция импорта данных (в форматах CSV, JSON).

Среди перспектив развития разработчики выделяют следующие: дополнительные возможности импорта и экспорта (в стандартном формате FLEx); соотнесение словарных записей с геотегами; просмотр карточек для изучающих язык; экспертная проверка правописания; возможность внесения предложений от пользователей; Push-уведомления для создателей словарей; загрузка видео.

«Живые словари» работают при подключении к сети Интернет, автономный режим находится в стадии разработки. Платформа запрограммирована с использованием HTML, CSS, Javascript и ReactJS с интеграцией Svelte и использует Google Firebase на сервере в качестве базы данных, размещенной в облаке. Платформа работает под некоммерческой лицензией с доступным исходным кодом.

«Живые словари» созданы в первую очередь для представителей языкового сообщества, которые заинтересованы в документировании своих языков. Платформа сознательно сделана доступной широкому кругу пользователей. У нее интуитивно понятный интерфейс и управление. Она проста в использовании и подходит любому человеку с базовой компьютерной грамотностью. Члены сообщества, создающие словарь, могут не владеть своим родным языком, но должны иметь доверительные отношения со свободно говорящими носителями, которые смогут поучаствовать в записи материалов для словаря. В то же время платформа популярна и в среде профессиональных лингвистов, занимающихся документацией, поскольку она позволяет удобно и в короткие сроки сделать доступными для сообщества записанные лингвистом материалы.

Обучающие видео о том, как работать с платформой, можно найти как в разделе Tutorials на сайте платформы, так и в плейлисте Tools for Endangered Languages (Living Dictionaries Tutorials) на YouTube. Кроме того, доступна подробная инструкция в текстовом виде в разделе FAQ на сайте платформы.

Если Вы хотите, чтобы Ваш язык был представлен на платформе «Живые словари»:

Контакты	Форма обратной связи на платформе
Что для этого нужно	● Техническое оборудование с возможностью аудиозаписи и с подключением к сети Интернет (ноутбук, смартфон, планшет, компьютер)

4.5. Приложение LIG-Aikuma

Ссылка	LIG-Aikuma
Тип ресурса	Приложение для документации
Для кого	Языковые активисты, представители сообщества, лингвисты

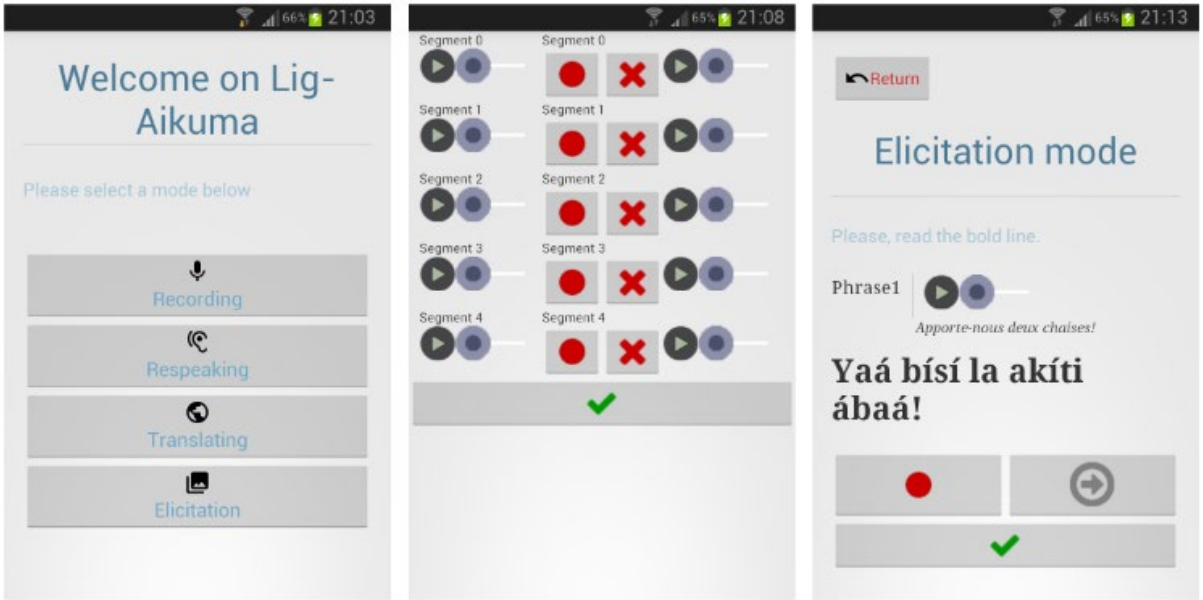


Рисунок 6. Внешний вид некоторых разделов приложения LIG-Aikuma [Adda et al. 2016]

«LIG-Aikuma» (букв. с языка усаруфа (папуасские языки) ‘место встречи’) — это бесплатное приложение с открытым кодом для устройств на Android. Приложение делает удобным процесс записи звучащей речи разными методами, а также предоставляет возможность создания параллельного корпуса звучащей речи [Godard et al. 2018].

Данный вариант приложения построен Лоран Безасье (Laurent Besacier) и Элоди Готье (Elodie Gauthier) из Гренобльской лаборатории компьютерных наук (Laboratoire d’Informatique de Grenoble, LIG) на базе приложения «Aikuma», разработанного Стивеном Бердом и Флорианом Ханке (больше информации доступно по ссылке https://en.wikipedia.org/wiki/Aikuma) [Besacier et al. 2019].

Приложение является в первую очередь инструментом для проведения аудиозаписи звучащей речи. Все записанные файлы хранятся в памяти того устройства, на котором установлено приложение. Таким образом, приложение выполняет свои основные функции оффлайн. Подключение к сети Интернет может понадобиться, если пользователь захочет перенести материалы, созданные в приложении, в облачное хранилище. Приложение не является «открытым архивом», когда записанные данные становятся доступными посетителям какой-либо платформы и пользователям приложения, как, например, «Живые словари».

В приложении реализованы следующие методы записи языковых данных: запись (в том числе спонтанной речи), перезапись (при необходимости исправить исходный файл или заменить его файлом лучшего качества), перевод (внесение перевода уже записанного фрагмента), элицитация (с письменным стимулом, изображением или видео, которые подгружаются в приложение пользователем) [Adda et al. 2016]. В режиме записи приложение работает фактически, как диктофон, однако, в отличие от встроенных диктофонов, создает файлы в формате wav. И позволяет потом работать с этими файлами. В режиме перезаписи можно повторно записать как весь фрагмент, если он вышел неудачно, так и часть фрагмента. При этом, если с файлом уже была проделана работа

по переводу и добавлению метаданных, все это сохраняется. В режиме элицитации можно предлагать носителю различные стимулы: текст, изображение, видео. В первом случае на экран устройства выводится фраза, записанная на языке-посреднике, которую необходимо произнести на языке-объекте. Фразы по одной подгружаются из текстового файла, содержащего все необходимые фразы, разбитые построчно. Во втором случае носителю языка предлагается изображение, которое нужно описать, или назвать то, что изображено. В третьем случае носителю предлагается прокомментировать видео. Изображение и видео также предоставляет пользователь, они должны храниться на устройстве с приложением. Таким образом, можно легко наладить запись звучащей речи разных дикторов (в том числе носителей разных языков) в ответ на одни и те же стимулы. Записанные таким образом языковые данные можно связывать друг с другом, формируя параллельный корпус звучащей речи [Blachon et al. 2016; Gauthier et al. 2016]. В режиме перевода фрагменты звучащей речи можно снабжать письменным переводом на язык-посредник.

В функционал приложения входит перенос файлов в облако, передача по Bluetooth и другие способы выведения данных с устройства, на котором производилась запись.

Если Вы хотите использовать приложение «LIG-Aikuma» в Вашей деятельности:	
Контакты	Форма обратной связи на платформе
Что для этого нужно	● Техническое оборудование на Android с возможностью аудиозаписи (смартфон, планшет) ● Подключение к сети Интернет необходимо, чтобы получить установочный файл и при желании выгружать файлы в облако

4.6. Платформа Ревита

Ссылка	Revita
Тип ресурса	Обучающая платформа
Для кого	Педагоги, студенты, изучающие язык самостоятельно

Рисунок 7. Стартовая страница профиля пользователя на платформе Ревита.

Это обучающая платформа, позволяющая работать с текстами (как из библиотеки, так и своими). Она предназначена для тех, кто уже владеет языком на достаточном уровне (чтение текстов со словарем, базовое знание грамматики). На данный момент открыты

полные версии для финского и русского языков [Катинская и др. 2020] и бета-версии для каталанского, китайского, французского, немецкого, итальянского, казахского, португальского, испанского, шведского, турецкого, северносаамского, а также для языков России: эрзянского, коми-зырянского, лугового марийского, якутского [Ivanova et al. 2019], татарского и удмуртского. Ревита задумывалась как проект по развитию и поддержке миноритарных языков, в особенности финно-угорских [Katinskaia, Yangarber 2018]. В настоящее время платформа развивается как инструмент для овладения иностранными языками в целом. Однако, наработанные технологии и механизмы впоследствии можно будет применить и к малоресурсным языкам.

Платформа разрабатывается в Лаборатории изучения языков Хельсинкского университета. Руководитель проекта — Роман Янгарбер, доцент кафедры цифровых гуманитарных наук Хельсинкского университета.

Ревита представляет собой уникальную платформу. С одной стороны, она закрывает лакуну в обучающих инструментах, предназначенных для уже владеющих языком на уровне CEFR A1 и выше. С другой стороны, Ревита позволяет работать с подгружаемыми пользователями текстами, а не с ограниченной библиотекой, собранной разработчиками. Таким образом, овладевающий языком может выбирать интересные и актуальные для него материалы. Одной из ключевых особенностей платформы является глубокая персонализация: новые упражнения с текстами генерируются с учетом предыдущих результатов учащегося. Упражнения адаптируются таким образом, чтобы учащийся мог отрабатывать темы, вызывающие сложности лично у него, но при этом уровень трудности не должен быть слишком высоким. Кроме того, учащийся и педагог могут отслеживать прогресс в обучении. В профиле пользователь может отслеживать на какие грамматические темы он выполнил задания и насколько успешно. Платформа позволяет преподавателям объединять зарегистрированных учеников в группы, для доступа к общему набору текстов и отслеживания прогресса [Stoyanova et al. 2021].

Платформа предоставляет следующие возможности работы с текстом [Katinskaia et al. 2018]:

- заполнение пропусков в тексте (поставить слово в нужную форму, записать по произнесению, выбрать из вариантов, в форме кроссворда);
- работа с лексическими карточками (слово — перевод).

Учебный материал создается автоматически с использованием искусственного интеллекта, языковых технологий, прикладной лингвистики и науки о данных в образовании. Для каждого языка прописываются грамматические правила, разбитые по категориям, а также составляется лексикон. Подгружаемые пользователями тексты разбираются с помощью автоматического морфоанализатора [Копотев и др. 2019]. На основе разметки и грамматики генерируются упражнения. Искусственный интеллект подключается на этапе отслеживания прогресса обучения и адаптации предлагаемых ученику упражнений. Кроме того, к платформе подключен внешний сервис Яндекса SpeechKit, преобразующий текст в речь, так что пользователь может прослушать любое слово из текста. Ведется работа над тем, чтобы читались фразы и далее весь текст целиком. Для разных языков эти возможности сейчас представлены в разной степени. Например, для русского можно выбрать перевод на один из 16 языков, есть произнесение не только лексического входа, но и словоформы из текста. А для коми-зырянского есть только перевод на русский язык. Дело в том, что некоторые из сложных функций платформы работают, только если имеется большое количество языковых данных, которые нужны для разработки более надежных языковых моделей. Поиск достаточного количества данных может быть проблематичным для малоресурсных языков, что ограничивает возможности обучающей платформы.

Ревита не представлена сейчас в формате приложения, но платформа адаптирована для мобильных устройств. Например, система может предлагать упражнения с выбором из нескольких вариантов ответа, если ввод текста с клавиатуры невозможен.

Подробную и актуальную информацию о платформе можно найти на странице проекта на сайте Хельсинского университета: Revita: Language learning and AI | University of Helsinki.

Если Вы хотите, чтобы Ваш язык был представлен на обучающей платформе Ревита:	
Контактное лицо	Роман Янгарбер roman.yangarber@helsinki.fi
Что для этого нужно	• Синтаксический парсер (можно разработать с командой Ревиты) • Морфологический анализатор (можно разработать с командой Ревиты) • Изначальная библиотека текстов (файлы в текстовом формате) • Грамматические правила (в определенной форме, разработанной для автоматических механизмов Ревиты) • Лексикон (лемма, часть речи и другая информация в формате xml, json или csv) • Специалист (носитель-консультант, или эксперт по языку высокого уровня) • Программист (может быть из команды Ревиты) • Финансирование

4.7. Приложения Динары Стёпиной

Ссылка	Приложения Динары Степиной на Google Play
Тип ресурса	Обучающее приложение
Для кого	Изучающие язык, лингвисты, языковые активисты, педагоги

Динара Стёпина — языковая активистка и разработчица приложений, ориентированных на людей, которые изучают языки России, находящиеся под угрозой исчезновения. Это, чаще всего, разговорники, которые иногда содержат словарную информацию. Все приложения созданы для операционной системы Android. Ниже мы расскажем о каждом из них. К Динаре можно обращаться, если есть желание сделать нечто подобное.

«Selkup Keyboard» — кириллическая раскладка для владеющих селькупским языком.

Приложение «Ӌвэ́лай! (Selkup)» создано для изучения селькупского языка, включает в себя разговорник и словарь.

«Itelmen Keyboard» — виртуальная клавиатура, содержащая все символы и буквы актуального ительменского алфавита. После установки и настройки печатать текст на ительменском можно будет в любом приложении, установленном на телефоне пользователя.

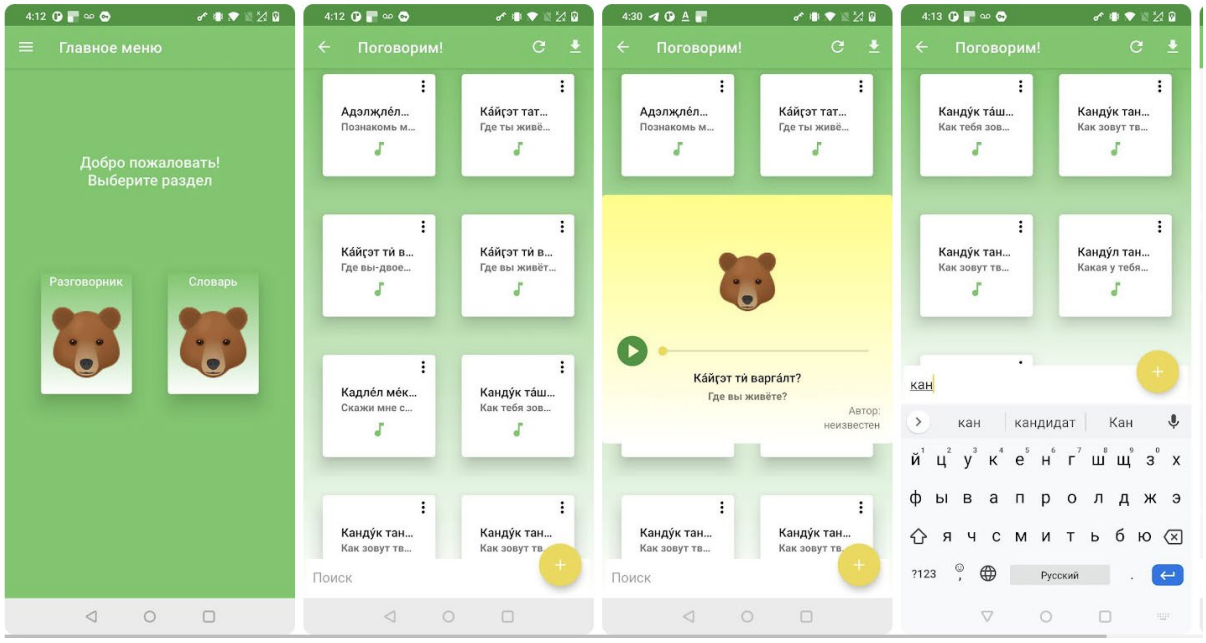


Рисунок 8. Скриншоты из приложения «Ҕвэ́лай! (Selkup)» для изучения селькупского языка.

Интерактивное приложение «Ҕҳынэсхч!» предназначено для изучающих ительменский язык, которое включает в себя аудиоразговорник и необходимые словари. Для удобного поиска по словарю рекомендуем установить виртуальную ительменскую клавиатуру.

«Learn Talysh!» (‘Учи талышский’) — приложение для изучения талышского языка. В красочном приложении поместились более 250 слов и базовых грамматических конструкций, с помощью которых можно научиться составлять множество полезных фраз. Приложение ориентировано на тренировку понимания на слух, поэтому аудио записаны с участием носителей разных говоров. Чтобы пользоваться «Learn Talysh!», вам не нужен интернет: слова и фразы можно учить прямо в приложении.

«Хэсэди (Разговорник)» поможет всем изучающим ульчский не только запоминать полезные фразы, но и прослушивать полноценные тексты на ульчском с переводом на русский. Кроме того, можно добавлять свои собственные фразы и записывать к ним аудио.

«Хэсэди (Словарь)» включает в себя как русско-ульчский, так и ульчско-русский словари на основе словарей О. П. Суника. Пользователям доступны не только переводы слов, но и информация о грамматике ульчских существительных и глаголов.

«Хэсэку!» — красочное приложение для изучения нанайского языка. Включает в себя русско-нанайский и нанайско-русский словари, а также разговорник с множеством озвученных носителями фраз на различные темы от «Семьи» до «Похода в горы».

Эти приложения очень востребованы среди представителей сообществ. И некоторые, например, посвященные ительменскому языку, созданы по их запросу. Динара планирует расширять функционал уже созданных и будущих приложений, не ограничивая его только возможностью прослушивать аудиозаписи слов и фраз. Одной из ключевых возможностей, которые планируется добавить и отладить, является возможность для пользователей пополнять базы словаря и разговорника новыми примерами, с предварительной модерацией менеджерами проектов. Эту опцию можно реализовывать двумя способами, которые зависят от условий подключения к сети Интернет. Если приложение скачивается пользователем и может быть использовано без подключения, то добавленные новые примеры будут доступны только после обновления приложения. Если же приложение работает в онлайн режиме, то новые примеры будут сразу появляться у всех пользователей. Другим важным направлением работы является расширение интерактивных возможностей

приложения и жанрового разнообразия представленных языковых данных, а также подключение видеоформата. Таким образом, приложение из справочного станет полноценным инструментом для изучения языка: будут добавлены интерактивные упражнения, в том числе на материале текстов и диалогов, а также интерактивный справочник по грамматике.

Кроме того, Динара Стёпина планирует перейти к разработке кросс-платформенных приложений, чтобы они были доступны пользователям не только операционной системы Android, но и iOS.

Для того, чтобы можно было приступать к разработке, необходимо определенным образом подготовить материалы: нарезать аудиофайлы на произнесения слов или фраз, если они были записаны единым массивом, назвать эти файлы, используя только латиницу и нижнее подчеркивание (без цифр и знаков препинания) и занести названия файлов в таблицу, сопоставив им ту информацию, которую вы хотите отразить в приложении (транскрипцию, запись в графике языка, переводы, какие-либо пометы). Разработка самого простого приложения-словаря или приложения-разговорника в случае, когда все материалы уже подготовлены, занимает от двух недель. В каждом конкретном случае следует сначала проконсультироваться с Динарой Стёпиной.

Контакты	Динара Стёпина ✉ setare497@gmail.com
Что для этого нужно	● Подготовленные языковые материалы (аудиофайлы нарезаны, данные занесены в таблицу) ● Финансирование

4.8. Перевод мультфильмов «Гора самоцветов» и «Лео и Тиг»

Ссылка	«Гора самоцветов» «Лео и Тиг»
Тип ресурса	Создание контента на языке
Для кого	Языковые активисты, лингвисты

Перевод мультфильмов на языки России очень востребован в языковых сообществах, поскольку на многих языках России очень мало контента вообще и детского в частности. При этом создание мультфильма с нуля — это очень дорогостоящий проект, а озвучить уже снятый мультфильм можно силами небольшой инициативной группы с посильными финансовыми затратами. По вопросам создания своей озвучки для представленных здесь мультсериалов необходимо обращаться к Василию Харитонову (языковому активисту и сотруднику Научного центра), который получил от правообладателей согласие на размещение в открытом доступе своих переводов и может помочь с организацией новых проектов.

«Гора самоцветов» — мультипликационный сериал по мотивам сказок народов России. Состоит из 87 серий продолжительностью примерно по 13 минут каждая. Серии выполнены в разных жанрах анимации. Посмотреть все мультфильмы можно в плейлисте YouTube. Мультфильмы были сняты студией «Пилот» (подробнее о студии можно прочитать в [Малюкова 2013: 32–39]), ранее под руководством Александра Татарского (подробнее о нем можно прочитать в [Капков (сост.) 2016]) при поддержке Федерального агентства по культуре и кинематографии. Перевод одного из эпизодов на нанайский язык, выполненный командой Василия Харитонова, можно посмотреть на YouTube.

«Лео и Тиг» — российский анимационный детский мультсериал (на данный момент снято 78 серий продолжительностью 11 минут каждая), созданный студией «Паровоз», о приключениях маленького леопарда Лео и рыжего Тигра. Посмотреть все мультфильмы можно в плейлисте YouTube. Уже несколько лет идет проект по переводу этого мультсериала на нанайский язык.

В общих чертах процесс перевода выглядит следующим образом. Выбирается серия для перевода. Далее производится расшифровка русского текста мультфильма (для некоторых серий «Горы самоцветов» такие расшифровки уже есть) и делается его письменный перевод на нужный язык. Лучше, чтобы над переводом работал не один человек, а целая группа знатоков языка. Следующий этап — аудиозапись полученного перевода. Для этого часто бывает достаточно одного диктора / актера. Желательно, чтобы это был носитель языка. К записи нужно подготовиться. Можно делать это в студии или с помощью профессионального диктофона, но можно и подручными средствами. Необходимо выставить на звукозаписывающем устройстве настройки максимального качества, а запись производить в помещении без эха (хорошо убирают эхо ковры, развешанные по стенам). Во время записи необходимо включить мультфильм (без звука), чтобы озвучивать фразы с учетом визуального контекста и длительности эпизода. Можно записывать по частям и желательно сделать несколько дублей. Последний этап — самый сложный с технической точки зрения. Необходимо на звуковую дорожку мультфильма без русскоязычных реплик (так называемые минуса́, они есть у Василия Харитонова) наложить ваш озвученный перевод. При этом дорожку желательно приглушать во время звучания реплик. Можно обратиться на этом этапе к профессиональному звукорежиссеру. Если есть желание самостоятельно разбираться в работе аудиоредактора, то можно взять консультацию у Василия Харитонова. Немаловажные требования со стороны правообладателей, о которых нужно помнить: о готовом к выкладыванию в открытый доступ мультфильме нужно сообщать правообладателю через Василия Харитонова, при выкладывании видео на платформе YouTube нельзя ставить свою рекламу (соответствующая галочка при публикации), устраивать показы мультфильмов можно исключительно в некоммерческих целях.

Контакты	Василий Харитонов vasiliysan@list.ru
Что для этого нужно	● Звукозаписывающее оборудование ● Персональный компьютер или ноутбук ● Финансирование

4.9. Субтитры в YouTube

Ссылка	Платформа YouTube
Тип ресурса	Создание контента на языке
Для кого	Языковые активисты, педагоги

Платформа YouTube доступна и популярна среди всех пользователей сети Интернет и, таким образом, является удачным инструментом для сохранения и развития языков, которым могут пользоваться как лингвисты, так и языковые активисты и представители языковых сообществ без специального образования. С одной стороны, это доступная возможность увеличивать представленность языка в сети Интернет, создавать контент на языке, расширять сферы функционирования языка. Вместе с тем, это удобный

инструмент для удаленной документации, с помощью которого можно расшифровывать и переводить тексты, что является этапом подготовки лингвистического корпуса.

В сентябре 2021 года Программа по документации исчезающих языков (англ. Endangered Languages Documentation Programme, сокр. ELDP) организовала серию воркшопов[28], посвященных различным аспектам дистанционной работы, на которых лингвисты делились своим опытом проведения удаленных проектов в 2021 году. Один из воркшопов был посвящен использованию возможностей платформы YouTube для создания перевода и субтитров представителями сообщества («YouTube as a transcription and translation tool», ведущий Александр Райс). На этом воркшопе участники могли научиться создавать субтитры и перевод, а также интегрировать эту практику в рабочий процесс удаленной документации.

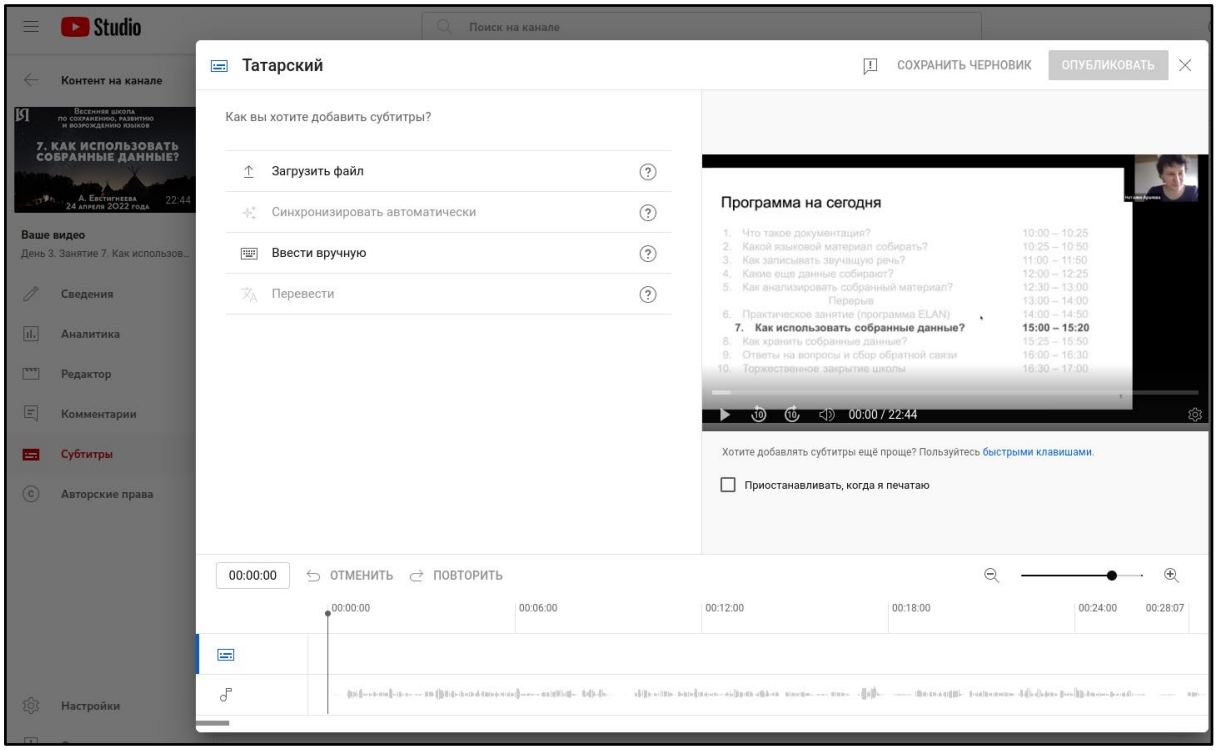


Рисунок 9. Окно Творческой студии YouTube, в котором создаются субтитры.

Использование платформы YouTube для расшифровки или перевода текстов в аудио- или видеоформате имеет много достоинств. С одной стороны, не нужно устанавливать никакого программного обеспечения и разбираться в его функционале. Кроме того, возможность работать на YouTube не зависит от операционной системы компьютера или ноутбука. При этом достаточно не самого стабильного и скоростного соединения с сетью Интернет. Если есть возможность просматривать видеосюжеты на платформе YouTube, то получится и создавать субтитры. Есть возможность совместной работы над одним проектом, хоть и не одновременно. Вдобавок созданные таким образом субтитры можно легко сделать доступными сообществу.

Однако есть и ряд ограничений, которые частично можно «обойти». Средствами платформы YouTube сложно сделать точную сегментацию аудиопотока. Если есть такая необходимость, то можно это поправить после, например, в программе ELAN. Или там же подготовить разбивку на сегменты заранее. Кроме того, платформа YouTube не дает возможности создавать несколько слоев аннотации, чтобы различать разных говорящих, если их в тексте несколько. В то же время, можно создать отдельные файлы с субтитрами для каждого из говорящих. Некоторым техническим ограничением можно назвать недостаточность смартфона для работы в Творческой студии YouTube. Кроме того, стоит

[28] Все пять воркшопов доступны в записи по ссылке: https://vimeo.com/showcase/8872469.

соблюдать некоторые меры безопасности, если вы не планировали распространять ваши материалы в открытом доступе. Важно выставить правильный уровень доступа при загрузке файла («Ограниченный доступ»), выходить из аккаунта при завершении рабочей сессии с общественного компьютера и удалить видео после завершения проекта.

В общих чертах процесс расшифровки и перевода текста выглядит следующим образом. Сначала необходимо загрузить видео на свой YouTube-канал (не забыв ограничить к нему доступ). Разбираться в настройках загрузки видео не обязательно, можно оставлять выставленные по умолчанию. После завершения загрузки нужно перейти в раздел редактирования видео и выбрать в меню слева опцию «Субтитры». Если YouTube создал автоматические субтитры, попытавшись распознать язык текста, их можно удалить. Добавляя субтитры, необходимо выбрать язык, при этом нужного языка в списке может не оказаться, тогда лучше выбрать какой-то редкий язык, чтобы YouTube не создал снова автоматические субтитры. В открывшемся окне для создания субтитров нужно выбрать опцию «Ввести вручную» и снизу появится таймлайн, а слева возможность добавлять аннотации. YouTube автоматически формирует сегмент длительностью в 2 секунды, что можно менять. В появившемся сверху слева пространстве нужно напечатать расшифровку выделенного фрагмента. Далее можно проигрывать видео до следующего момента с речью и, остановив, создать новую аннотацию. И так далее. Чтобы завершить рабочую сессию и сохранить работу, нужно нажать «Сохранить черновик». А если работа над субтитрами завершена, можно нажать «Опубликовать». После этого можно приступать к созданию второго слоя субтитров, например, с переводом. При этом в окне редактирования откроются уже созданные субтитры с сегментацией и расшифровкой, что сделает процесс перевода более удобным. После публикации созданных субтитров появляется возможность их скачать в формате srt. Файлы этого формата при необходимости можно использовать для создания слоев аннотации в программном обеспечении ELAN.

Весь описанный выше процесс пошагово продемонстрирован на видеозаписи упомянутого воркшопа ELDP на английском языке. Кроме того, за консультацией можно обратиться к сотруднице Научного центра Елене Будянской.

Контакты	Елена Будянская ✉ budyanskaya.lena@gmail.com
Что для этого нужно	● Техника с подключением к сети Интернет ● Тексты на языке в видео- или аудиоформате ● Аккаунт на платформе YouTube у всех участников проекта ● Программное обеспечение: ELAN (при необходимости)

4.10. Википедия на языке

Ссылка	Платформа Википедия
Тип ресурса	Создание контента на языке
Для кого	Языковые активисты, лингвисты, представители языковых сообществ

Фонд Викимедиа устанавливает ряд критериев и требований для создания новой Википедии. Прежде чем подавать заявку, необходимо ознакомиться со страницей «Правила подачи заявок на новый языковой раздел», где говорится о критериях, которым фонд следует, чтобы считать предложение приемлемым, а также о процессе его утверждения или отклонения.

Одним из условий окончательного утверждения является перевод интерфейса Википедии на новый язык. «Статистика локализации» описывает текущую доступность переводов интерфейса MediaWiki на разные языки. Перевод программного обеспечения, лежащего в основе Википедии, выполняется на веб-сайте https://translatewiki.net (см. инструкцию «Перевод программного обеспечения, лежащего в основе Википедии» на английском языке).

С другой стороны, на странице «Запросы на новые языки» можно проверить, было ли уже представлено предложение для того или иного языка и на каком этапе оно находится.

Как только было принято решение запросить создание новой википедии, необходимо следовать инструкциям, приведенным в «инструкциях по добавлению нового запроса».

Если требования для возможности участия соблюдены, Языковой комитет должен подтвердить, что это «правомочно». Следующим шагом является начало написания тестового проекта на сайте Вики-Инкубатора, где потенциальные проекты Викимедиа в версиях на новых языках могут быть упорядочены, написаны, протестированы и проверены на пригодность для размещения Фондом Викимедиа. Полный список вики-проектов см. в разделе «Инкубатор: Вики-проекты».

Наконец, если все требования соблюдены и детальная проверка не выявила проблем, языковой комитет уведомит все сообщество о предстоящем утверждении, добавив уведомление в раздел «Обсуждение: Языковой комитет». Если больше не возникнет проблем, запрос будет одобрен, и разработчикам будет предложено создать вики.

Что касается создания в Википедии контента на языках, на Весенней школе по сохранению, развитию и возрождению языков, проведенной Научным центров весной 2022 года А. Э. Аарони, один из программистов Википедии, в режиме реального времени показал, как работает специальный интерфейс с автоматическим переводом, которым можно пользоваться при создании новых статей в Википедии. Запись мастер-класса доступна по ссылке.

Контакты	Амир Аарони amir.aharoni@mail.huji.ac.il
Что для этого нужно	● Техника с подключением к сети Интернет

5. Пилотная онлайн-платформа по ревитализации и документации

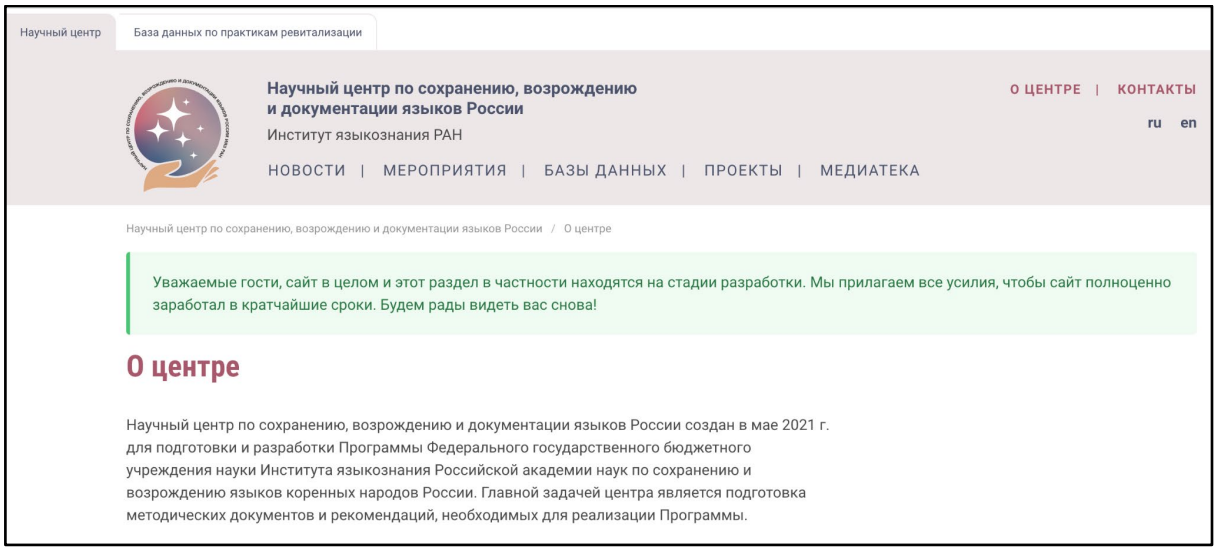


Рисунок 10. Стартовая страница сайта Научного центра.

Осенью 2022 года Научным центром была запущена пилотная версия онлайн-платформы по ревитализации и документации, реализованная К. К. Поливановым. Мы также продумали ее общую структуру и идеи, которые планируется реализовывать в последующие годы. Такой ресурс представляется актуальным как для лингвистов, так и для языковых активистов и представителей сообществ, заинтересованных в сохранении и развитии своих языков. Данная платформа будет, с одной стороны, отражать деятельность Научного центра, а с другой стороны, станет интерактивной справочной системой по различным аспектам документации и ревитализации языков, с возможностью пополнения данных пользователями ресурса. Таким образом, планируется формирование своего рода сообщества вокруг платформы, благодаря действиям которого представленная информация будет отвечать требованиям полноты и актуальности.

Для разработки и запуска пилотной версии платформы с учетом дальнейших перспектив ее развития были решены следующие технические задачи.

- Создание модели данных на основе технического задания как для сайта центра, так и для связанных с ним баз данных.
- Развертка и настройка на сервере программной среды, необходимой для функционирования сайтов.
- Развертка фреймворка управления содержимым Drupal 9 для проектирования системы управления контентом сайта и баз данных, а также подключение и настройка необходимых программных модулей.
- Получение и привязка SSL сертификатов для обеспечения безопасного https соединения с сайтами (шифрование соединения в целях повышения безопасности).
- Создание системы управления содержимым на платформе Drupal 9 согласно разработанному техническому заданию.
- Разработка дизайна базовых элементов конструкции страниц, а также дизайна интерфейсов и отдельных элементов.
- Создание макетов базового графического решения.
- Программирование вывода данных как для отдельных страниц, так и для представлений списков данных.
- Верстка страниц сайта и баз данных согласно макетам и разработанному техническому заданию, в частности, создание каскадных таблиц стилей (css), программирование шаблонов вывода данных (twig), создание программных каркасов html страниц и блоков.
- Написание JavaSript кодов для работы динамических и интерактивных элементов сайта. Отдельно верстка мобильной версии сайтов.
- Подключение и настройка google analytics и яндекс-метрики для отслеживания посещаемости ресурсов.
- Создание многопользовательской системы взаимодействия с содержимым сайтов, с соответствующими ролями.
- Настройка мультиязычности сайтов и программирование необходимых для этого элементов.
- Разработка пользовательского интерфейса перехода между сайтом и базами данных.
- Регулярное обслуживание серверной программной среды, а также системы сайта.

5.1. Структура онлайн-платформы

На онлайн-платформе Научного центра содержится информация разного рода о различных аспектах документации и ревитализации языков. При этом платформа

задумана не только как информационный ресурс, но и как инструмент для тех, кто ведет деятельность в этих областях. Содержательное наполнение и функциональная нагрузка определяют структуру платформы. Из стандартных разделов, которые присутствуют на всех онлайн-ресурсах, независимо от их тематики и направленности, на онлайн-платформе представлены следующие: «О центре», «Новости» и «Контакты». Эти разделы являются чисто информационными и призваны ввести пользователя в контекст. Так, в разделе «О центре» кратко представляется Научный центр: время и изначальные цели его формирования, результаты работы за прошлые годы (на данный момент 2021 год), а также сотрудники центра. Раздел «Новости» посвящен актуальным событиям в жизни Научного центра и в развитии онлайн-платформы. В разделе «Контакты» указаны способы связи с сотрудниками Научного центра, а также ссылки на другие ресурсы, которые поддерживаются центром. Например, ссылка на канал на платформе YouTube. В дальнейшем планируется сделать в этом разделе форму обратной связи.

Своего рода копилкой всевозможных информационных ресурсов о языках и медиаконтента на языках России призван стать раздел «Медиатека», в котором будет содержаться коллекция опубликованных материалов, аудио- и видеоконтента по теме платформы, а также планируются разделы с контентом на языках России. Каждый материал, указанный в разделе, будет снабжен подробными метаданными, что, с одной стороны, позволит создать удобную систему навигации, а с другой стороны, обеспечит возможность тонко настраиваемого поиска в соответствии с запросом пользователей платформы.

Одной из важных составляющих деятельности Научного центра являются открытые мероприятия, посвященные разным аспектам сохранения и поддержки языков. При организации таких мероприятий необходимо публиковать организационную информацию в сети Интернет. Кроме того, после таких мероприятий остаются полезные видеозаписи, презентации и другие материалы, к которым уместно оставлять возможность доступа не только для участников этих мероприятий, но и для всех желающих. Этому посвящен раздел платформы «Мероприятия». На данный момент там содержится информация о двух онлайн-мероприятиях, в организации которых Научный центр принимал участие, при этом подробная информация и материалы находятся на отдельных веб-сайтах: Весенняя школа по сохранению, развитию и возрождению языков (22–24 апреля 2022 года, онлайн) и Семинар «Языковое гнездо. Создаем и работаем» (15 апреля 2022 года, онлайн). В будущем планируется интегрировать организационные и содержательные страницы, посвященные мероприятиям, в общую структуру платформы, обеспечивая таким образом централизацию информации, единообразное оформление и упрощенную навигацию по данным разделам. Кроме того, у пользователей будет возможность запросить повтор или развитие того или иного мероприятия через обратную связь.

В разделе «Проекты» содержится информация о длительных проектах, в которых принимают участие и которые организуют сотрудники Научного центра. Это могут быть проекты как по документации, так и направленные на сохранение и развитие языков России. Помимо справочной информации о каждом проекте (время проведения, участники, цели и задачи, результаты и другое), в этом же разделе будут представлены и сами результаты проектов. Планируется, что представление результатов будет ориентировано на широкий круг пользователей и, в первую очередь, на представителей языковых сообществ и языковых активистов. Так, например, по итогам документационных проектов будут сформированы разделы платформы, на которых будут опубликованы тексты на языках в видеоформате, снабженные субтитрами и переводом, а также поморфемной аннотацией в удобном формате. Эти же материалы будут отражены и в «Медиатеке», что отвечает современным требования к построению платформ и обеспечению различных вариантов доступа к одной и той же информации.

Этот же принцип применяется и к разделу с рабочим названием «Базы данных». Переход к этому разделу может быть осуществлен через главное меню в шапке веб-сайта. Одновременно можно перейти к каждой из представленных баз данных через закладки,

размещенные над шапкой. Такое выделение данных ресурсов в том числе отражает их особую роль в структуре и функционале онлайн-платформы: именно в этом разделе будет реализован инструментарий для работы с данными, позволяющий проводить их анализ, пополнять и поддерживать их актуальность. Подробнее каждая база данных будет описана в Разделе 5.2.

Важно отметить, что техническая верстка платформы изначально учитывала возможность реализации мультиязычного перевода для всех разделов. В том числе и для баз данных, что представляет отдельную техническую задачу. На данный момент начата работа по переводу на английский язык, что делает платформу доступной для не владеющих русским языком пользователей. Поскольку опыт по документации и ревитализации на российской почве играет значительную роль в мировом контексте в данной сфере, такой выход на международную арену крайне важен. Тем более, что средства обратной связи, заложенные в платформе, позволят вести взаимный обмен опытом между специалистами и языковыми активистами по всему миру. Кроме того, планируется перевод платформы и на языки России, что полностью соответствует подходу к работе с языками, которого придерживается Научный центр. Таким образом, расширяются сферы функционирования языков России, увеличивается их представленность и видимость в сети Интернет, создается контент на этих языках, повышается их престиж, что должно положительно влиять на их сохранение и развитие.

5.2. Раздел онлайн-платформы «Базы данных»

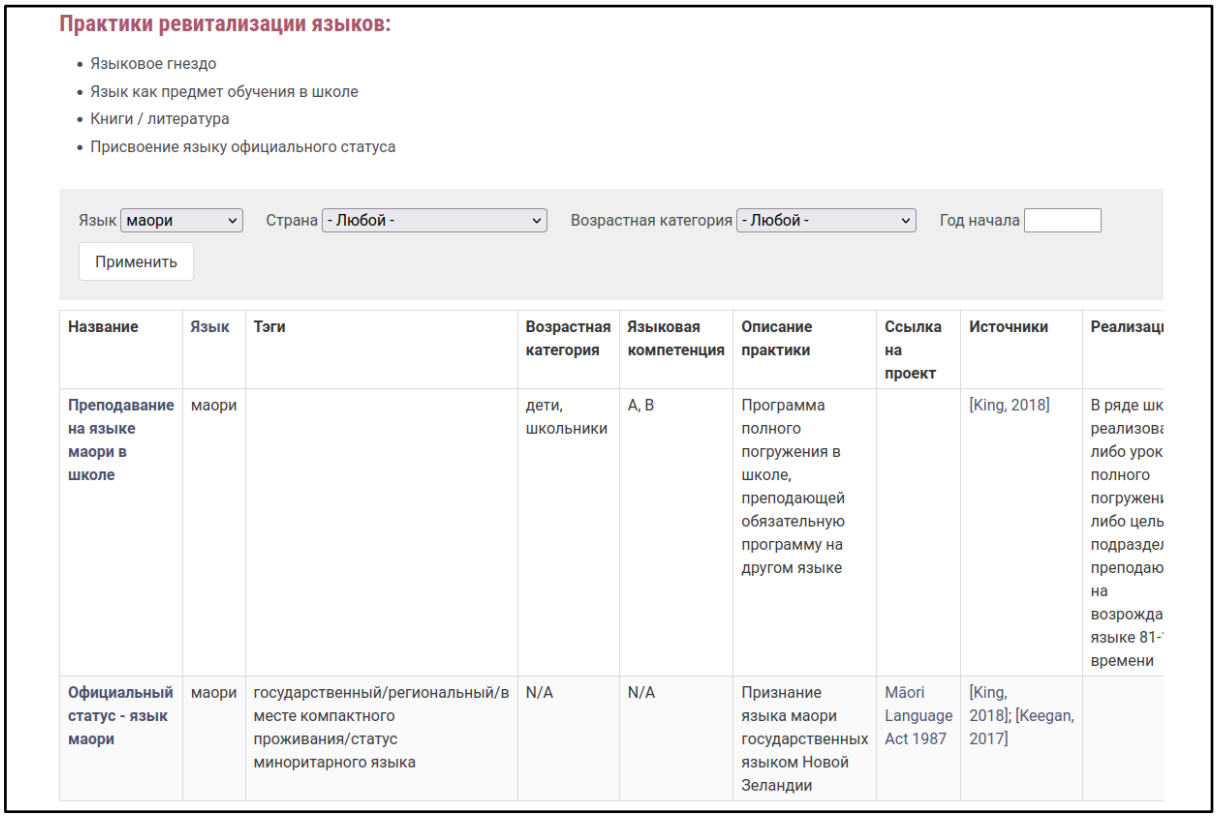
В данном разделе онлайн-платформы будут функционировать следующие базы данных: «Лучшие российские и зарубежные практики сохранения и развития языков», «Состояние языков России», «Задокументированность языков России», работа над которыми ведется сотрудниками Научного центра с 2021 года, а также «Онлайн-инструменты для поддержки, ревитализации и документации языков», отражающая результаты работы в 2022 году, представленные в данном документе.

Базы данных разрабатывались различными рабочими группами сотрудников Научного центра, каждая из которых выбрала подходящую для своей области структуру. На этапе разработки платформы стало понятно, что пока придется представить эти базы данных раздельно, поскольку изначально не было заложено единообразия, и информация, содержащаяся в базах, оказалась по-разному структурированной. Было выработано техническое решение, позволяющее все же связать данные из разных баз, что особенно важно для поставленных перед платформой задач. Переход от одной базы к другой будет осуществляться за счет перекрестных ссылок на дополнительные материалы по той или иной теме.

Как и для всей платформы, для баз данных реализована техническая возможность перевода на несколько языков. Начат перевод базы по практикам ревитализации на английский язык. Данная база была представлена авторами проекта международному сообществу в ходе доклада «The Database on Revitalization Practices: understanding the processes of language preservation and restoration» (с англ. ‘База данных по практикам ревитализации: осмысливая процессы сохранения и развития языков’) на III Международной конференции по ревитализации коренных и миноритарных языков (III International Conference on Revitalisation of Indigenous and Minoritized Languages, 13–16 16 сентября 2022 года, г. Жирона (Испания) и г. Перпиньян (Франция)).

База данных «Лучшие российские и зарубежные практики сохранения и развития языков» состоит, в первую очередь, из собрания практик по возрождению, сохранению, поддержке и развитию того или иного языка. Для удобства использования и исследования, все зафиксированные практики разделены на категории, а в рамках категорий — в подкатегории. Помимо этого, каждой практике присваивается ряд значений из закрытого списка, дополнительно классифицирующих их по таким параметрам, как целевая аудитория (возрастная группа) и языковая компетенция. Каждой практике присвоен список тегов —

дополнительных значений, опционально классифицирующих практики по какому-либо второстепенному параметру. Наконец, каждая практика дополнительно описывается при помощи уточняющей информации (в свободной форме): описание практики, ссылка на проект, источники, год начала и завершения практики, финансирование, количество участников проекта, условия реализации, автор или организатор проекта и другие. В базе реализованы два способа презентации практик: в рамках сводной таблицы (см. Рисунок 11) и в виде отдельной веб-страницы с подробным структурированным описанием (см. Рисунок 12).



Практики ревитализации языков:

- Языковое гнездо
- Язык как предмет обучения в школе
- Книги / литература
- Присвоение языку официального статуса

Язык маори | Страна - Любой - | Возрастная категория - Любой - | Год начала

Применить

Название	Язык	Тэги	Возрастная категория	Языковая компетенция	Описание практики	Ссылка на проект	Источники	Реализац
Преподавание на языке маори в школе	маори		дети, школьники	A, B	Программа полного погружения в школе, преподающей обязательную программу на другом языке		[King, 2018]	В ряде шк реализова либо урок полного погружен либо цель подразде преподаю на возрожда языке 81- времени
Официальный статус - язык маори	маори	государственный/региональный/в месте компактного проживания/статус миноритарного языка	N/A	N/A	Признание языка маори государственных языком Новой Зеландии	Māori Language Act 1987	[King, 2018]; [Keegan, 2017]	

Рисунок 11. Пример выдачи при поиске по языку маори в сводной таблице практик ревитализации.

Для удобства пользования сводной таблицей практик реализована возможность сортировки записей по параметру «язык». Кроме того, встроена возможность поиска нужных практик по параметрам «язык», «страна», «возрастная категория» и «год начала». Некоторые данные в сводной таблице и на отдельных страницах практик снабжены ссылками на внешние ресурсы, в которых содержится более подробная информация по той или иной теме, относящейся к данной записи.

Разработан и находится на стадии внедрения механизм, позволяющий пользователям сайта добавлять в базу практики ревитализации, известные им, но отсутствующие в базе. Для внесения данных авторами проекта была разработана специальная форма, закрытая от внешнего пользователя. Но использовать ее же для задач краудсорсинга нецелесообразно ввиду сложности ее организации. Для внешних пользователей была разработана более интуитивно понятная форма, помогающая описать ту или иную практику по необходимым параметрам. Затем заполненная форма проходит стадию модерации и почти автоматически переносится в базу данных с некоторыми дополнениями, которые администраторам придется сделать вручную. Кроме того, через обычную форму обратной связи пользователи платформы смогут оставлять свои комментарии к уже существующим записям в базе, что позволит дорабатывать информацию в случае обнаружения каких-либо неточностей.

Преподавание на языке маори в школе

Язык	маори
Страна	Новая Зеландия
Классификация	Образование » Учебный/воспитательный процесс на языке » Язык как средство обучения в школе
Возрастная категория	дети, школьники
Языковая компетенция	A, B
Описание практики	Программа полного погружения в школе, преподающей обязательную программу на другом языке
Источники	[King, 2018]
Год начала	1985
Реализация	В ряде школ реализованы либо уроки полного погружения, либо целые подразделения, преподающие на возрождамом языке 81-100% времени

Рисунок 12. Пример страницы, посвященной практике преподавания языка в школе для языка маори.

В базе данных «Состояние языков России» содержится информация о различных параметрах, описывающих языковую ситуацию разных языков России: количество носителей и уровни владения, наличие исследований по языку, функционирование в различных сферах жизни, наличие контента на языке и другие. На данный момент в базе подробно описаны 104 языка России. На настоящий момент разработаны принципы представления информации, содержащейся в базе, а также внутренняя форма для внесения этих данных, позволяющая реализовать необходимую визуализацию на техническом уровне. С одной стороны, к описанию состояния того или иного языка России можно будет перейти из раздела «Языки России», в котором языки будут представлены на интерактивной карте и в сводной таблице. Метки языков на интерактивной карте будут различаться соответственно приписанным языкам статусам витальности. Сводная таблица, помимо названия языков, будет содержать информацию о языковой семье и о статусе витальности. При нажатии на название языка или на метку на карте пользователь будет попадать на страницу с подробным структурированным описанием языковой ситуации для данного языка. С другой стороны, разработанный механизм ввода данных позволит визуализировать значения этих параметров в виде диаграммы, наглядно отражающей языковую ситуацию. Этот формат будет реализован в разделе «Анализ данных». Опция поиска будет позволять визуализировать информацию по комбинации полей: язык, год, конкретные параметры. Кроме того, заложена возможность создания на одной странице нескольких диаграмм по разным языкам для сравнительного анализа. Помимо собственно данных о состоянии языков, в базе будут представлены два больших теоретических раздела, посвященных методологии оценки состояния и сбора данных. Так же, как и первой базе, будет реализован механизм добавления информации внешним пользователем с предварительной модерацией.

В базе «Задокументированность языков России» отражена информация о том, какие документационные материалы собраны по разным языкам России, на которых сейчас говорит очень мало людей. На данный момент в базе представлено 50 языков и диалектов России. Изначально, как и с остальными проектами, информация накапливалась в формате таблицы с достаточно большим количеством полей и листов, что, естественно, требует переработки для визуализации и использования на платформе. Сотрудниками Научного

центра был разработан механизм внесения данных таблицы в базу платформы через специальную форму, поля которой связаны друг с другом так, чтобы можно было выводить данные базы для внешнего пользователя в разных форматах и сочетаниях. Данные могут быть представлены в виде отдельной страницы, посвященной конкретному языку, в виде всплывающего окна и в виде краткой таблицы. Доступ к данным предоставляется в разделе с рабочим названием «Поиск по базе». К странице, посвященной конкретному языку, можно перейти через интерактивную карту нажатием на маркер соответствующего языка, а также осуществив поиск по базе и нажав на название выбранного языка в выдаче. Поиск по базе можно осуществлять по двум параметрам: «язык» и «область документации» (грамматика, словарь, текст и некоторые другие). При этом при выборе нескольких языков будет выдана краткая таблица, по которой можно будет сравнить задокументированность этих языков. Кроме того, будет реализована возможность быстрого доступа к подробным данным: при выборе одного языка и одной области документации соответствующие данные будут представлены во всплывающем окне. Отдельная страница языка будет содержать краткую информацию о социолингвистических аспектах языковой ситуации этого языка, релевантную для документации. Например, количество носителей выражается в нескольких цифрах: количество владеющих языком в той или иной степени, в том числе пассивно, и отдельно количество людей, готовых работать с лингвистами в документационных проектах. Далее на странице будет представлена наглядная визуализация степени задокументированности языка с возможностью получения более детализированных данных по каждой области документации во всплывающем окне. Для этой базы также будет реализован механизм добавления информации внешним пользователем с модерацией.

6. Перспективы дальнейшей работы

Среди перспектив дальнейшего исследования можно отметить следующее. Новые онлайн-инструменты постоянно появляются, поэтому следует продолжать их мониторинг и регулярно пополнять базу данных новыми позициями, поддерживая таким образом ее актуальность. Сам перечень инструментов с описаниями планируется представить на сайте Научного центра в виде базы данных с удобным и понятным интерфейсом, возможностью поиска по нескольким параметрам и формой обратной связи, через которую любой человек сможет поучаствовать в наполнении базы и предложить внести данные по новым онлайн-инструментам.

Кроме того, в планах есть проведение анализа существующих цифровых инструментов, которые уже реализованы для языков России. Необходимо понять, для каких языков России какие инструменты функционируют, и сформулировать, чего не хватает и какие инструменты целесообразно реализовывать в первую очередь. Так, в 2023 году нам предстоит сформировать приоритеты в области разработки цифровых инструментов для 5 языков России. В более далекой перспективе (2026–2032 годы) планируется предоставлять научно-методическое сопровождение разработки приоритетных цифровых инструментов для различных миноритарных языков России.

Мы также планируем сами пользоваться описанными инструментами, в частности, в рамках работы по документации языков России. Будут в электронном виде создаваться озвученные разговорники, озвученные переводные и диалектологические словари исследуемых языков. Будут разрабатываться, публиковаться и пополняться аннотированные корпуса, в том числе мультимедийные, для исследуемых языков. Многие найденные инструменты также можно использовать в процессе дистанционной работы с носителями в рамках удаленной документации.

В то же время планируется развитие и поддержка информационного ресурса Научного центра, разработанного в этом году. Ресурс содержит научно-методические рекомендации и инструменты для сохранения и развития языков России, и эта информация будет регулярно пополняться в процессе работы нашего Научного центра. Нам предстоит

оформить базы данных по состоянию языков России и по задокументированности языков России в соответствии с разработками, подготовленными в этом году. Планируется также реализовать многоязычный интерфейс сайта, переведя его на различные языки России. Это небыстрый процесс, требующий вовлечения знатоков языка и дополнительного финансирования на их работу. Кроме того, на основе сайта в 2023 году также запланирована разработка онлайн-платформы по документации языков России. Эта платформа будет содержать результаты наших документационных проектов и будет ориентирована в первую очередь на представителей языковых сообществ и языковых активистов.

7. Отражение результатов работы в проекте Федеральной программы по сохранению и развитию языков России

В данном разделе мы приводим список мероприятий, касающихся использования онлайн-технологий в сфере поддержки языков России. Эти мероприятия основаны на исследовании онлайн-инструментов, описанном в этом документе, и были включены в проект Федеральной программы по сохранению и развитию языков России, разработанный осенью 2022 года в Институте языкознания РАН. Номера в первом столбце отражают нумерацию мероприятий в проекте Программы.

2.30	Выделение финансирования на создание настольных, интерактивных, компьютерных игр, мультфильмов и фильмов, книг и других обучающих и развлекательных материалов на целевых языках
3.4	Проведение анализа существующих IT-инструментов для целевых языков, формирование приоритетов в области разработки IT-инструментов для каждого целевого языка
3.5	Проведение аналитики новых IT-инструментов на российском и зарубежном рынке и возможности их применения для целевых языков
3.6	Разработка раскладок клавиатуры для целевых языков, подходящих для различных операционных систем, в том числе мобильных
3.7	Обеспечение поддержки системных шрифтов для целевых языков в различных операционных системах, в том числе мобильных
3.8	Разработка морфологических анализаторов для целевых языков
3.9	Разработка систем проверки орфографии для целевых языков для различных операционных систем
3.10	Разработка автоматических машинных переводчиков для целевых языков
3.11	Разработка систем автоматического распознавания и синтеза звучащей речи для целевых языков
3.13	Разработка мобильных и онлайн-словарей для целевых языков
3.14	Разработка мобильных и онлайн-разговорников для целевых языков
3.15	Разработка платформ для интерактивного обучения целевым языкам, в том числе ресурсных баз

3.16	Разработка инструментов для автоматического перевода, генерации субтитров, воспроизведения автоматически переведенной речи в видео- и аудиопотоках на целевых языках
3.17	Разработка системы хранения мультимедийных материалов на целевых языках для их последующей разметки, использования для машинного обучения и других практических задач
3.18	Проведение хакатонов по разработке цифровых решений для целевых языков
5.3, 5.4	Создание, публикация и распространение оригинального контента на целевых языках разных жанров и форматов, для разной целевой аудитории как в самом языковом сообществе, так и за его пределами: печатные СМИ, радио, телевидение, книги и журналы, фильмы и мультфильмы, интернет-издания, аудиокниги, подкасты и т. п.
14.7	Создание и интеграция субтитров на целевых языках для информационных, научно-популярных и развлекательных телевизионных передач

Заключение

Данный документ посвящен исследованию существующих лингвистических онлайн-инструментов с точки зрения их возможного использования для поддержки, ревитализации и документации языков. Мы собрали коллекцию релевантных инструментов в табличном виде, выделили параметры их описания и описали инструменты в соответствии с выработанным параметрам (см. Приложение 1). При этом для удобства описания и поиска получилось разделить инструменты на несколько общих групп в зависимости от сферы их использования.

Кроме того, были описаны базовые методики создания ориентированных на языковые сообщества онлайн-инструментов. Мы связались с разработчиками, обсудили принципы устройства и работы этих инструментов, уточнили необходимые для их создания навыки, оборудование, языковые материалы и финансирование. Полученные результаты мы отразили в виде справочных материалов в Разделе 4.

В течение года мы вместе с программистом разрабатывали и реализовывали интерфейс пилотной онлайн-платформы по документации и ревитализации. Осенью 2022 года эта платформа была запущена. Была также разработана и реализована база данных по лучшим российским и зарубежным практикам ревитализации языков, сейчас активно ведется ее наполнение. Кроме того, были разработаны спецификации для базы данных по состоянию языков России и базы данных по задокументированности языков России, подробно описанные в Разделе 5.2.

Еще одним результатом стал подготовленный список мероприятий, касающихся использования онлайн-технологий в сфере поддержки языков России. Эти мероприятия были включены в проект Федеральной программы по сохранению и развитию языков России, разработанный осенью 2022 года в Институте языкознания РАН.

Нет сомнений в актуальности темы онлайн-инструментов в свете тренда на цифровизацию и общего развития технологий. Стоит продолжать мониторинг подобных приложений и онлайн-платформ, а также информировать языковые сообщества России о существовании таких инструментов и помогать внедрять их в жизнь.

Список литературы

Антопольский А. Б. Лингвистические ресурсы и технологии в России: Состояние и перспективы (Обзор) // Социальные новации и социальные науки. М.: ИНИОН РАН, 2021. № 2. С. 114–131.

Винклер М. Э. А., Павлова О. М. Итоговые документы, подготовленные в ходе научно-исследовательской работы «Разработка плана и методик сохранения и возрождения языков России»: 3. Российские практики сохранения, возрождения и развития языков [электронный ресурс]. Институт языкознания РАН, 2021. URL: https://iling-ran.ru/languages_of_russia/2021_stage1/doc3.pdf (дата обращения: 01.03.2023).

Дыбо А. В., Норманская Ю. В. Система «Лингводок»: Опыт работы с тюркскими материалами [устный доклад]. Доклад на конференции «The IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”» (Кызыл, 21–23 сентября 2021 г.).

Капков С. (сост.). Александр «Pilot» Татарский. М.: ИД Мещерякова, 2016.

Катинская А., Жуэ Х., Янгарбер Р. Упражнения на основе текста: как компьютер помогает преподавателю // Русский язык за рубежом. 2020. № 6(283). С. 50–57.

Копотев М., Катинская А., Иванова С., Янгарбер Р. REVITA: изучение языка на основе корпусных подходов // Корпусная лингвистика. 2019. С. 30–39.

Корнев Т. В. Итоговые документы, подготовленные в ходе научно-исследовательской работы «Разработка плана и методик сохранения и возрождения языков России»: 4. Зарубежные практики сохранения, возрождения и развития языков [электронный ресурс]. Институт языкознания РАН, 2021. URL: https://iling-ran.ru/languages_of_russia/2021_stage1/doc4.pdf (дата обращения: 01.03.2023).

Малюкова Л. Сверхкино. Современная российская анимация. Девяностые/нулевые. СПб.: Умная Маша, 2013.

Норманская Ю. В. Планы на развитие платформы ЛингвоДок в рамках консорциума организаций из 4 Республик [устный доклад]. Доклад на конференции «Открытая конференция ИСП РАН им. В. П. Иванникова» (Москва, 2–3 декабря 2021 г.).

Норманская Ю. В. Первые тюркские кириллические книги на платформе ЛингвоДок // Родной язык. 2022. № 1. С. 43–57.

Сюрюн А. А., Будянская Е. М., Евстигнеева А. П., Жорник Д. О. Итоговые документы, подготовленные в ходе научно-исследовательской работы «Разработка теоретических принципов и практических аспектов сохранения и возрождения языков России»: 5. Годовой отчет по проведенной работе в области документации [электронный ресурс]. Институт языкознания РАН, 2022.

Харитонов В. С., Иванов В. В., Каде М. А. Итоговые документы, подготовленные в ходе научно-исследовательской работы «Разработка плана и методик сохранения и возрождения языков России»: 2. Состояние языков России [электронный ресурс]. Институт языкознания РАН, 2021. URL: https://iling-ran.ru/languages_of_russia/2021_stage1/doc2.pdf (дата обращения: 01.03.2023).

Adda G., Adda-Decker M., Ambouroue O., Besacier L., Blachon D., Bonneau-Maynard H. E., Gauthier E., Godard P., Hamlaoui F., Idiatov D., Kouarata G.-N., Lamel L., Makasso E.-M., Rialland A., Stuker S., van de Velde M., Yvon F., Zerbian S. Innovative technologies for under-resourced language documentation: The BULB Project // LREC 2016 Workshop — CCURL 2016. Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (23 May 2016) — Proceedings / C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, E. Wandl-Vogt (eds.). Portoroz, 2016. P. 59–66.

Adams O., Cohn T., Neubig G., Cruz H., Bird S., Michaud A. Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation [электронный ресурс] // Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan) / H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, T. Tokunaga (eds.). European Language Resources Association (ELRA), 2018. P. 3356–3365.
URL: https://halshs.archives-ouvertes.fr/halshs-01709648v4/document (дата обращения: 01.03.2023).

Anderson G. D. S., Daigneault A. L. Living Dictionaries: An Electronic Lexicography Tool for Community Activists // Electronic lexicography in the 21st century: Proceedings of the eLex 2021 conference (virtual, 5–7 July 2021) / I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (eds.). Brno: Lexical Computing CZ s.r.o., 2021a. P. 339–360.

Anderson G. D. S., Daigneault A. L. Living Dictionaries: A Web Platform to Protect Endangered Languages and Maintain Global Linguistic Diversity [устный доклад]. Доклад на конференции «ContribuLing 2021» (Online, 3–4 June 2021b).

Anderson G. D. S., Fahringer J., Harrison D., Forepiso K. R., Barth D. Matukar Online Talking Dictionary [материалы устного доклада]. Доклад на конференции «12th International Conference on Austronesian Linguistics» (Denpasar, 2–6 July 2012).

Babinski S., Haakman K., Lake A., Yi I., Cram C., Ananthanarayan S., Bowern C. How can endangered language collections be made more discoverable? [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October 2022).

Beck D., Shannon P. SLEXIL: User-centred software for community language documentation // Language Documentation & Conservation. Vol. 14. 2020. P. 488–502.

Begay W. R. Mobile apps and indigenous language learning: New developments in the field of indigenous language revitalization. Master's Thesis. Tucson: University of Arizona, 2013.

Besacier L., Gauthier E., Voisin S. Lessons learned after development and use of a data collection app for language documentation (lig-aikuma) // Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019 / S. Calhoun, P. Escudero, M. Tabain, P. Warren (eds.). Canberra: Australasian Speech Science and Technology Association Inc., 2019. P. 3897–3901.

Bettinson M., Bird S. Collaborative Fieldwork with Custom Mobile Apps // Language Documentation & Conservation. Vol. 15. 2021. P. 411–432.

Bigi B. SPPAS: Multilingual approaches to the automatic annotation of speech // The Phonetician. Vol. 111–112. 2015. P. 54–69.

Blachon D., Gauthier E., Besacier L., Kouarata G.-N., Adda-Decker M., Rialland A. Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App [электронный ресурс] // 5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9–12 May 2016, Yogyakarta, Indonesia (Procedia Computer Science, Vol. 81) / S. Sakti, M. Adriani, A. Purwarianti, L. Besacier, E. Castelli, P. Nocera (eds.). Elsevier Science, 2016. P. 61–66.
URL: https://www.sciencedirect.com/science/article/pii/S1877050916300448 (дата обращения: 01.03.2023).

Burke M., Zavalina O. L., Chelliah S. L., Phillips M. E. User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users // Language Documentation & Conservation. Vol. 16. 2022. P. 1–24.

Cassels M., Farr C. Mobile applications for Indigenous language learning: Literature review and app survey // Working Papers of the Linguistics Circle. Vol. 29, № 1. 2019. P. 1–24.

Chase B., Borland K. Networks of support: How online resources are built, maintained, and adapted for community language revitalization needs at FirstVoices // Language Documentation & Conservation. Vol. 16. 2022. P. 209–227.

Cordin P., Rabanus S., Alber B., Mattei A., Casalicchio J., Tomaselli A., Bidese E., Padovan A. VinKo, Versione 2 (20.12.2018, 09:20) [электронный ресурс] // Lo spazio comunicativo dell'Italia e delle varietà italiane. Korpus im Text / R. Bauer, T. Krefeld (eds.). München: Ludwig-Maximilians-Universität, 2018. P. 1–15.
URL: http://www.kit.gwi.uni-muenchen.de/?p=13739&v=2 (дата обращения: 01.03.2023).

van Esch D., Foley B., San N. Future Directions in Technological Support for Language Documentation // Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages. Vol. 1 (Papers) / A. Arppe, J. Good, M. Hulden, J. Lachler, A. Palmer, L. Schwartz, M. Silfverberg (eds.). Honolulu: Association for Computational Linguistics, 2019. P. 14–22.
URL: https://aclanthology.org/W19-6003 (дата обращения: 01.03.2023).

Everson R., Honoré W., Grimm S. An Online Platform for Community-Based Language Description and Documentation // Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages. Vol. 1: Papers. Honolulu: Association for Computational Linguistics, 2019. P. 1–5.

Foley B., Arnold J., Coto R., Durantin G., Ellison T. M., van Esch D., Heath S., Kratochvíl F., Maxwell-Smith Z., Nash D., Olsson O., Richards M., San N., Stoakes H., Thieberger N., Wiles J. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS) [электронный ресурс] // Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018). 2018. P. 205–209.
URL: https://www.isca-speech.org/archive/pdfs/sltu_2018/foley18_sltu.pdf (дата обращения: 01.03.2023).

Galla C. K. Digital Realities of Indigenous Language Revitalization: A Look at Hawaiian Language Technology in the Modern World // Language and Literacy. Vol. 20, № 3. 2018. P. 100–120.

Garnica B., Lindsey K. Archiving the Ende Language Corpus [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October 2022).

Gauthier E., Blachon D., Besacier L., Kouarata G.-N., Adda-Decker M., Rialland A., Adda G., Bachman G. LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies [электронный ресурс] // INTERSPEECH 2016 (Proceedings of the conference): Show & Tell Contribution (September 8–12, 2016, San Francisco, USA). International Speech Communication Association (ISCA), 2016. P. 381–382.
URL: https://www.isca-speech.org/archive/pdfs/interspeech_2016/gauthier16_interspeech.pdf (дата обращения: 01.03.2023).

Genee I., Junker M.-O. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization // Language Documentation & Conservation. Vol. 12. 2018. P. 274–314.

Godard P., Adda G., Adda-Decker M., Benjumea J., Besacier L., Cooper-Leavitt J., Kouarata G.-N., Lamel L., Maynard H., Mueller M., Rialland A., Stueker S., Yvon F., Zanon-Boito M. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments [электронный ресурс] // Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan) / H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, T. Tokunaga (eds.). European Language Resources Association (ELRA), 2018. P. 3366–3370.
URL: https://aclanthology.org/L18-1531.pdf (дата обращения: 01.03.2023).

Griscom R. T. Supporting linguistic data collection from afar: A mobile metadata system // Language Documentation & Conservation. Vol. 16. 2022. P. 98–119.

Harrison K. D., Lillehaugen B. D., Fahringer J., Lopez F. H. Zapotec Language Activism and Talking Dictionaries // Electronic lexicography in the 21st century: Proceedings of the eLex 2019 conference (Sintra, 1–3 October 2019) / I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (eds.). Brno: Lexical Computing CZ s.r.o., 2019. P. 31–50.

Heaton R., Hooshiar K. Engaging Communities in digital development for the Native American Languages Collection Presentation [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of the Indigenous Languages» (Berlin & Online, 3–7 October 2022).

Huilcán Herrera M. I. The Use of Technologies in Language Revitalisation Projects // Journal of Global Indigeneity. 2022. Vol. 6, № 1. P. 1–17.

Ivanova S., Katinskaia A., Yangarber R. Tools for supporting language learning for Sakha // 22nd Nordic Conference on Computational Linguistics (NoDaLiDa): Proceedings of the Conference (Linköping Electronic Conference Proceedings, № 167; NEALT Proceedings Series, № 42) / M. Hartmann, B. Plank (eds). Linköping: Linköping University Electronic Press, 2019. P. 155–163.

Katinskaia A., Nouri J., Yangarber R. Revita: A Language-learning Platform at the Intersection of ITS and CALL [электронный ресурс] // Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan) / H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, T. Tokunaga (eds.). European Language Resources Association (ELRA), 2018. P. 4084–4093.
URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/841.pdf (дата обращения: 01.03.2023).

Katinskaia A., Yangarber R. Digital cultural heritage and revitalization of endangered Finno-Ugric languages [электронный ресурс] // Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (Helsinki, Finland, March 7–9, 2018) / E. Mäkelä, M. Tolonen, J. Tuominen (eds.). Vol. 2084. 2018. P. 111–121.
URL: https://ceur-ws.org/Vol-2084/paper9.pdf (дата обращения: 01.03.2023).

Koole M., Lewis K. w. Mobile Learning as a Tool for Indigenous Language Revitalization and Sustainability in Canada: Framing the Challenge // Indigenous Studies: Breakthroughs in Research and Practice / Information Resources Management Association (ed.). Hershey, PA: IGI Global, 2020. P. 520–532.

Kosheliuk N. LingvoDoc: extralinguistic information and maps [электронный ресурс] // EasyChair Preprint. № 6766. 2021a.
URL: https://easychair.org/publications/preprint/Mj13 (дата обращения: 01.03.2023).

Kosheliuk N. LingvoDoc: New Opportunities for Graphic and Phonetic Analysis of Endangered Language [электронный ресурс] // EasyChair Preprint. № 5387. 2021b.
URL: https://easychair.org/publications/preprint/CPbr (дата обращения: 01.03.2023).

Kosheliuk N. LingvoDoc: Working with Text Corpora [электронный ресурс] // EasyChair Preprint. № 6242. 2021c.
URL: https://easychair.org/publications/preprint/nSP6 (дата обращения: 01.03.2023).

Kosheliuk N. LingvoDoc: Working with Phonology [электронный ресурс] // EasyChair Preprint. № 5387. 2021d.
URL: https://easychair.org/publications/preprint/NGSp (дата обращения: 01.03.2023).

Kosheliuk N. LingvoDoc: Working with Cognate Analysis [электронный ресурс] // EasyChair Preprint. № 5387. 2021e.
URL: https://easychair.org/publications/preprint/BcXt (дата обращения: 01.03.2023).

Kosheliuk N. «LingvoDoc» as a Solution to the Problems of Conservation of Endangered Languages (Using Example of the Mansi Language) [электронный ресурс] // EasyChair Preprint. № 8383. 2021f.
URL: https://easychair.org/publications/preprint/17Xh (дата обращения: 01.03.2023).

Kruijt A. VinKo: crowdsourcing language documentation of minoritized language varieties in Italy [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October, 2022).

Li J., Brar A., Roihan N. The use of digital technology to enhance language and literacy skills for Indigenous people: A systematic literature review [электронный ресурс] // Computers and Education Open. Vol. 2. 2021.
URL: https://www.sciencedirect.com/science/article/pii/S2666557321000069?via%3Dihub (дата обращения: 01.03.2023).

Miangah T. M., Nezarat A. Mobile-Assisted Language Learning // International Journal of Distributed and Parallel systems. Vol. 3, № 1. 2012. P. 309–319.

Michaud A., Adams O., Cohn T., Neubig G., Guillaume S. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit // Language Documentation & Conservation. Vol. 12. 2018. P. 393–429.

Miller J. C. Digital preservation for the masses: Creating and publishing workflows and guidelines for institutions and smaller organisations interested in digital preservation, data management, and archiving [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of the Indigenous Languages» (Berlin & Online, 3–7 October 2022).

Nathan D. Digital archives: essential elements in the workflow for endangered languages documentation and revitalisation // Language Documentation and Description. Vol. 5. 2008. P. 103–119.

Nathan D. Archives 2.0 for Endangered Languages: From Disk Space to MySpace // International Journal of Humanities and Arts Computing. Vol. 4. 2010. P. 111–124.

Nathan D. Digital Archiving // The Cambridge Handbook of Endangered Languages / P. K. Austin, J. Sallabank (eds.). Cambridge: Cambridge University Press, 2011. P. 255–273.

Nathan D. Access and Accessibility at ELAR, an Archive for Endangered Languages Documentation // Language Documentation and Description. Vol. 12 (Special Issue on Language Documentation and Archiving) / D. Nathan, P. K. Austin (eds.). London: School of Oriental and African Studies, 2014. P. 187–208.

Normanskaja J. V. Dictionaries on Samoyedic languages and LingvoDoc software system for collaborative work on dictionaries and online publishing [электронный ресурс] // NORDSCI Conference Proceedings. Vol. 1 (Book 1). 2018. P. 251–259.
URL: https://www.nordsci.org/_files/ugd/32dbef_7c5c05b1526e44159333eaac749299eb.pdf (дата обращения: 01.03.2023).

Normanskaja Y. V., Borisenko O. D., Beloborodov I. B., Avetisyan A. I. The Software System LingvoDoc and the Possibilities It Offers for Documentation and Analysis of Ob-Ugric Languages // Doklady Mathematics. Vol. 105. 2022. P. 187–206.

Ovide E., García-Peñalvo F. J. Internet technologies as a tool in indigenous education: The case of the wichi people in ‘the impenetrable’ area in Argentina // Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM’16) / F. J. García-Peñalvo (ed.). New York: Association for Computing Machinery, 2016. P. 441–445.

Panigrahi S., Seyfeddinipur M., Kung S. Bridging the Divide: Connecting Language Activist Efforts and Language Archives [устный доклад]. Доклад на конференции «Language

Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October, 2022).

Paterson III H., Barker P., Lee M. R. Metadata for Language Learning Resources: Advancing OLAC for Language Development [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October, 2022).

Pride K., Tomlin N., AnderBois S. LingView: A Web Interface for Viewing FLEx and ELAN Files // Language Documentation & Conservation. Vol. 14. 2020. P. 87–107.

Rabanus S., Kruijt A., Tagliani M., Tomaselli A., Padovan A., Alber B., Cordin P., Zamparelli R., Vogt B. M. VinKo (Varieties in Contact) Corpus v1.1 [электронный ресурс]. Eurac Research CLARIN Centre, 2022.
URL: http://hdl.handle.net/20.500.12124/46 (дата обращения: 01.03.2023).

Rice A. Using YouTube as the Primary Transcription and Translation Platform for Remote Corpus Work // Language Documentation & Conservation. Vol. 15. 2021. P. 514–550.

Stoyanova N., Hou J., Kopotev M., Yangarber R. Integration of computer-aided language learning into formal university-level L2 instruction // L'analisi linguistica e letteraria. Vol. 29, № 3. 2021. P. 117–126.

Sullivant R. Archival description for language documentation collections // Language Documentation & Conservation. Vol. 14. 2020. P. 520–578.

Thieberger N., Harris A., Barwick L. PARADISEC: Its History and Future // Research, Records and Responsibility: Ten Years of PARADISEC / A. Harris, N. Thieberger, L. Barwick (eds.). Sydney: Sydney University Press, 2015. P. 1–15.

Thieberger N., La Rosa M., Lewincamp S. Language manuscripts in archives [устный доклад]. Доклад на конференции «Language Documentation and Archiving during the Decade of Indigenous Languages» (Berlin & Online, 3–7 October, 2022).

Ungsitipoonporn S., Watyam B., Ferreira V., Seyfeddinipur M. Community Archiving of Ethnic Groups in Thailand // Language Documentation & Conservation. Vol. 15. 2021. P. 267–284.

Yi I., Lake A., Kim J., Haakman K., Jewell J., Babinski S., Bowern C. Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives [электронный ресурс] // Journal of Open Humanities Data. Vol. 8. 2022.
URL: https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.59 (дата обращения: 01.03.2023).

ПРИЛОЖЕНИЕ

Ориентированные на языковые сообщества онлайн-инструменты поддержки языков

Приложение можно посмотреть по ссылке.