ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 504, с. 60–82

ИНФОРМАТИКА

УДК 004.428.4, 004.622, 004.811.51

ПРОГРАММНАЯ СИСТЕМА LingvoDoc И ВОЗМОЖНОСТИ, КОТОРЫЕ ОНА ПРЕДЛАГАЕТ ДЛЯ ДОКУМЕНТИРОВАНИЯ И АНАЛИЗА ОБСКО-УГОРСКИХ ЯЗЫКОВ

© 2022 г. Ю. В. Норманская[1,2], О. Д. Борисенко[1,*], И. Б. Белобородов[1], академик РАН А. И. Аветисян[1]

Поступило 05.02.2022 г.
После доработки 28.02.2022 г.
Принято к публикации 01.03.2022 г.

LingvoDoc (http://lingvodoc.ispras.ru) обеспечивает сервис для совместной языковой документации и расчетов по собранным данным. Эта программная система предоставляет GraphQL HTTP API для всех составляющих частей и позволяет пользователям создавать собственные расширения для анализа данных или даже интегрировать их со своим собственным программным обеспечением. Благодаря способу построения системы и схемы ее базы данных можно создавать автономные приложения, интегрированные с системой LingvoDoc: этим приложениям потребуется подключение к Интернету только один раз для синхронизации основных типов данных и для целей аутентификации. Сама система позволяет пользователям создавать многослойные словари, прикреплять их к географической карте, наполнять документы метаданными, делиться доступом к словарям с другими пользователями или со всеми. Система LingvoDoc также предоставляет детализированные списки контроля доступа для совместного использования, что позволяет разделить пользователей на группы редакторов словарей, корректоров и пользователей только для чтения. Система также предоставляет некоторые вычислительные алгоритмы для хранимых данных: фонологические вычисления, автоматическую и управляемую дедупликацию внутри словарей и т.д. Система позволяет пользователям выбирать структуру словаря. Она поддерживает следующие типы данных: текст, изображения, звуки (wav, mp3 и flac), разметки (форматы ELAN и Praat), направленные и не направленные связи между сохраняемыми объектами. Пользователь может выбрать наиболее подходящий формат для своего словаря. Также система обеспечивает хранение, просмотр и обработку корпусов ELAN. В системе создан ряд авторских программ, которые позволяют проводить обработку языкового материала с точки зрения фонетики и этимологии. Эти программы воспроизводят то, что ранее ученые делали вручную, увеличивая скорость анализа материала в десятки, а в ряде случаев даже в сотни раз. В данной статье представлен фрагмент возможностей документации и анализа материала обско-угорских языков с использованием системы LingvoDoc.

Ключевые слова: программное обеспечение, документирование и анализ языков, анализ данных.

DOI: 10.31857/S2686954322030055

1. ОБЗОР СИСТЕМ ЯЗЫКОВОЙ ДОКУМЕНТАЦИИ

LingvoDoc – это серверное программное обеспечение, предназначенное для документирования и анализа языков в совместной работе групп исследователей. На систему в значительной степени повлияли программные системы контроля версий, которые в основном используются программистами, но слишком сложны для повседневного использования лингвистами (Github) (Bitbucket). Основная цель проекта заключалась в разработке системы, которая предоставляла бы большинство своих функций через веб-интерфейс для наиболее частых задач документирования и обработки (ввода архивного и полевого аудио- и видеоматериала, его транскрибирования, этимологизации, глоссирования и последующего анализа и сравнения с другими языками) и в то же время предоставляла средства для внешней обработки данных через HTTP API для опытных пользователей, знакомых с программированием и обработкой естественных языков. Несколько существующих систем могут хранить аналогичные языковые данные; наиболее близкими к LingvoDoc являются проект Starling [1], проект LEGO [2], TypeCraft [3], Kielipankki [4] и

[1] *Институт системного программирования Российской академии наук им. В.П. Иванникова, Москва, Россия*

[2] *Институт языкознания Российской академии наук, Москва, Россия*

*****E-mail: al@somestuff.ru*

проект corpus-tools [5], но все эти системы имеют функциональные ограничения, которые мы пытались преодолеть.

Проект Starling – это настольное программное обеспечение, позволяющее создавать этимологически связанные словари. Отдельный словарь Starling выглядит как таблица с лексическими элементами; каждая лексическая запись имеет уникальный целочисленный идентификатор для конкретного словаря. Таблица может иметь неограниченное количество именованных столбцов двух типов: столбцы, содержащие текстовые данные, и столбцы, содержащие указатели на другие словари. Ячейки таблицы последнего типа содержат целочисленные значения, соответствующие идентификаторам внутри словаря, имя файла которого совпадает с именем столбца. Пользовательский интерфейс Starling помогает перемещаться по подключенным объектам и предлагает широкий спектр возможностей для ввода и анализа данных. Кроме того, Starling имеет некоторые функции экспорта, включая экспорт в HTML-представление, поэтому словари можно экспортировать только для чтения в Интернете.

Тем не менее у Starling есть определенные ограничения.

Эта система является проприетарной и не имеет публично открытых исходных кодов; таким образом, она не может быть изменена или расширена сторонними разработчиками. Впрочем, если бы исходники были открыты сообществу, это не сильно изменило бы ситуацию: программа разработана на языке программирования Harbour [6] (предок Clipper 5.3), который вряд ли можно назвать распространенным языком программирования, поскольку базы данных DBF не такие гибкие, как современные реляционные (и нереляционные) системы управления базами данных.

Вторая проблема заключается в невозможности синхронизировать все словари среди исследователей, работающих с программой: если какой-либо из словарей в какой-то момент “разветвится”, все добавленные связи для разветвленного словаря будут недействительны. Более того, поскольку словари представлены в виде списка простых файлов, общей проблемой является то, что все подключенные словари хранятся исследователями как архив файлов, который не может быть изменен никем, кроме их автора. Но многие исследователи пользуются относительно общими словарями (например, словарями протоформ), и ни один из них в данном случае не будет идентичным.

Третья проблема укрепила нашу мотивацию для разработки новой системы. Starling может хранить только текстовые (под “текстом” мы подразумеваем “переменные символьные данные”) ячейки данных без модификации и без какой-либо возможности хранить вложенные объекты, несмотря на то, что существует общая потребность в хранении медиафайлов (аудио, изображения , видео), аннотационные данные для медиафайлов (такие как формат EAF/ELAN для длинных текстов или формат TextGrid/Praat для фонологических данных (в частности, варианты произношения)) и различные типы ссылок, которые должны быть неизменяемыми даже при участии многих редакторов. LingvoDoc предназначен для работы со всеми перечисленными выше типами данных и предоставляет гибкую систему типов. Ее пользователи могут создавать типы столбцов, которые повторно используются во всей системе. Эта функция также полезна для определения различных типов отношений между словарями.

Проект **LEGO** кажется почти статичным без каких-либо возможностей для динамического изменения данных в системе пользователями программы. Однако у этой системы есть одна примечательная особенность, которая делает ее отличной межъязыковой онтологией понятий: каждый сохраненный объект словаря имеет связь с “Concepticon ID” в проекте GOLD. Таким образом, модель данных LingvoDoc дает возможность при необходимости легко использовать целые наборы данных LEGO и GOLD.

TypeCraft – это проект, который использует популярный движок MediaWiki (программное обеспечение для организации вики – веб-сайта, контент которого создают сами пользователи, используя браузер). Этот проект предоставляет отличные варианты метаданных и языковой документации, поскольку он наследует все функции движка Википедии. Кажется, что система предоставляет подстрочную текстовую аннотацию через язык разметки MediaWiki, и этот подход универсален, так как это очень популярная платформа. Но этот подход представляется менее гибким, чем формат EAF (ELAN), поскольку не позволяет выстраивать сложные взаимосвязи между слоями разметки. Также основное внимание в проекте уделяется языковой документации и созданию коллекции разговорных текстов; однако его трудно использовать для составления словарей или соединения лексических статей. Движок MediaWiki не предоставляет каких-либо сильных возможностей для типизации данных, поэтому эта функция вряд ли когда-нибудь появится.

Проект **Kielipankki** предоставляет множество лингвистических инструментов и отдельный вебпортал, посвященный расширенному поиску по корпусам (KORP) (своды документов). Весь проект ориентирован на корпуса и не предоставляет никаких инструментов унификации. Проект хранит массивные коллекции данных, которые, кажется, структурированы вручную: большинство коллекций организованы по-разному, поскольку

разные авторы могли отправить их в систему; администраторы системы помечают данные тегами. Мы не можем точно определить, какое подмножество хранимых корпусов индексирует KORP, но поисковая система мощная и охватывает большой объем данных. На самом деле, в данный момент система не ориентирована на словарные данные, а ее движок представлен закрытым исходным кодом, поэтому его нельзя повторно использовать или модифицировать.

Проект **Corpus-tools** предлагает несколько программных инструментов. **Salt** обеспечивает промежуточную модель отображения, которая может унифицировать форматы хранения сводов документов. Программный инструмент **Pepper** содержит набор конвертеров для некоторых форматов хранения корпусов. Отличительной особенностью **Pepper** является метод конвертации, который позволяет избежать преобразования каждого формата в каждый: каждый формат должен иметь конвертер внутрь и наружу из промежуточной модели Salt, после чего исходный корпус теоретически может быть преобразован в любой другой формат, представленный в **Pepper**. Проект **ANNIS** представляет собой средство визуализации и поиска в корпусах. **Electron** – это многоуровневый инструмент для аннотирования корпусов, который распространяется как программное обеспечение для настольных ПК. Он использует модель представления Salt в качестве формата хранения. У нас есть планы на будущее по интеграции с конвертерами Pepper, чтобы обеспечить возможности конвертации с помощью нашего веб-интерфейса, поскольку это единственная потенциально переиспользуемая часть проекта Corpus-tools для LingvoDoc.

Все вышеперечисленные системы не покрывают потребности нашей исследовательской группы по ряду причин.

1. Ни одна из существующих систем не обеспечивает расширяемой системы для типов данных и возможности одновременного создания пользовательской структуры для словарей и корпусов, что строго необходимо для межсловарных связей с различной семантикой (это важно для возможности связи родственных слов в разных языках, имеющих разное значение).

2. Большинство систем не предоставляют API для внешней обработки данных или интеграции.

3. Большинство систем не предоставляют исходные коды с разрешительной лицензией, поэтому их нельзя повторно использовать или адаптировать для конкретных нужд.

4. Ни одна из систем, кроме TypeCraft, не поддерживает совместную интерактивную работу над одним и тем же языковым ресурсом.

5. Ни одна из систем не поддерживает децентрализованный режим с отложенной синхронизацией между соавторами.

Эти причины послужили основным мотивом для разработки новой системы.

2. ОБЗОР СИСТЕМЫ LingvoDoc

LingvoDoc – система, ориентированная на создание словарей и корпусов текстов и их анализа. На главной странице (lingvodoc.ispras.ru) пользователю предоставляется список опубликованных словарей и корпусов текстов, структурированный в соответствии с генетической классификацией языков (представленной в виде дерева языков, их диалектов и говоров), а также с учетом ссылок на гранты и организаций, в которых проводилась работа. Вверху страницы расположено меню быстрой навигации. Число в квадратных скобках показывает, сколько словарей/корпусов содержит язык или языковая семья (рис. 1).

Словарь представлен как многослойная структура, и каждый из слоев (в стандартном случае – это лексические и парадигматические входы) содержит свои типизированные столбцы. Как показано на рис. 2, число указывает, содержит ли словарь как лексические, так и парадигматические входы, или еще дополнительные слои, а нажатие кнопки “Просмотр” показывает список доступных слоев. Щелчок по имени выбранной перспективы приведет пользователя внутрь выбранного слоя словаря или коллекции корпусов.

Структура языкового дерева, представляющего генетическую классификацию языков, является гибкой, поэтому в нее можно добавлять новые языки и диалекты, но она по-прежнему имеет некоторые подмножества “защищенных” языков и диалектов, которые могут быть изменены только системными администраторами. Система предоставляет функции изменения месторасположения языка или диалекта на языковом дереве. Каждый “незащищенный” язык может быть помещен в нужный узел дерева в качестве дочернего или близкородственного языка, а можно также добавить дочерний узел к любому языку с помощью кнопки “Создать” (рис. 3).

Языковое дерево используется в процессе создания словаря или корпуса текстов (рис. 4, 5). Словарь может быть создан в двух режимах: с нуля или в виде словаря, импортированного из некоторых источников, поддерживаемых системой. В настоящее время поддерживаются следующие источники импорта: формат ранних версий LingvoDoc и CSV со специальным символом-разделителем. Файлы CSV можно создавать из файлов Excel, используя диалоговое окно экспорта Starling или любым другим способом. Модуль импорта LingvoDoc для CSV поддерживает массовый импорт нескольких файлов с сохранением взаимосвязей между импортированными слова-

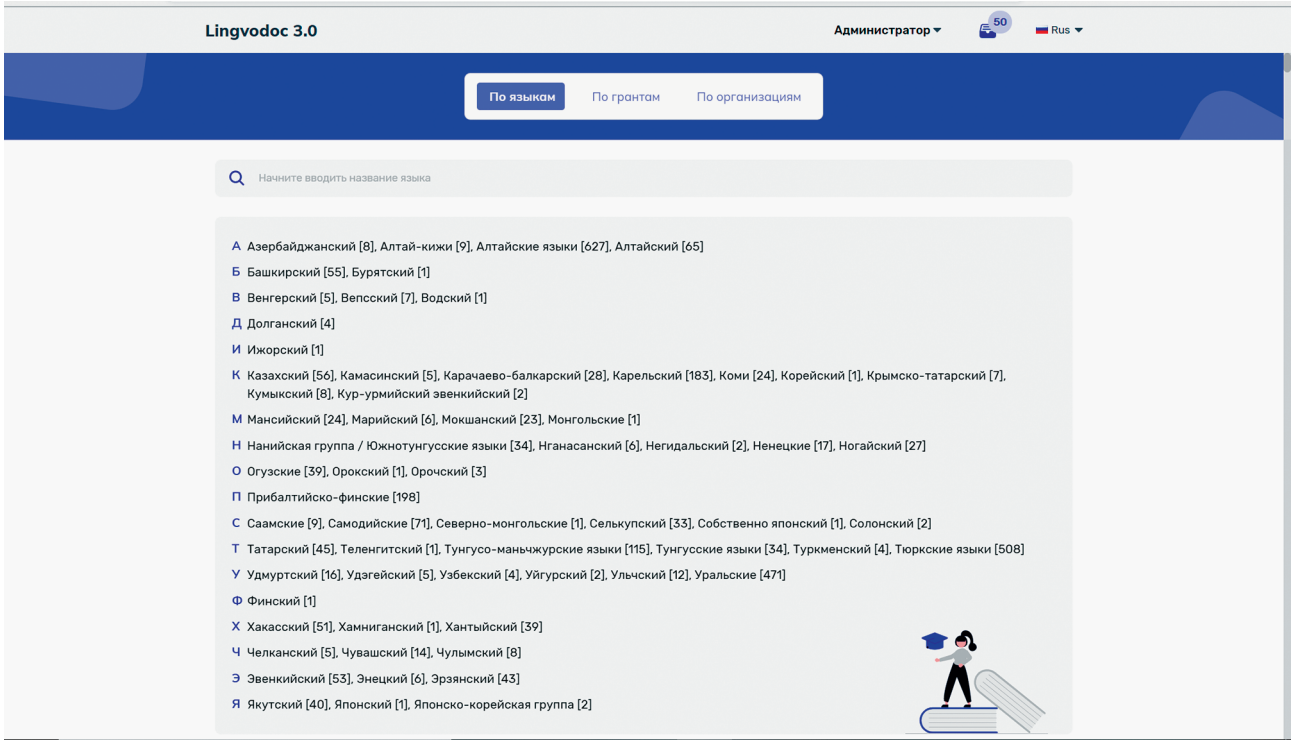

Рис. 1. Навигация по языку.

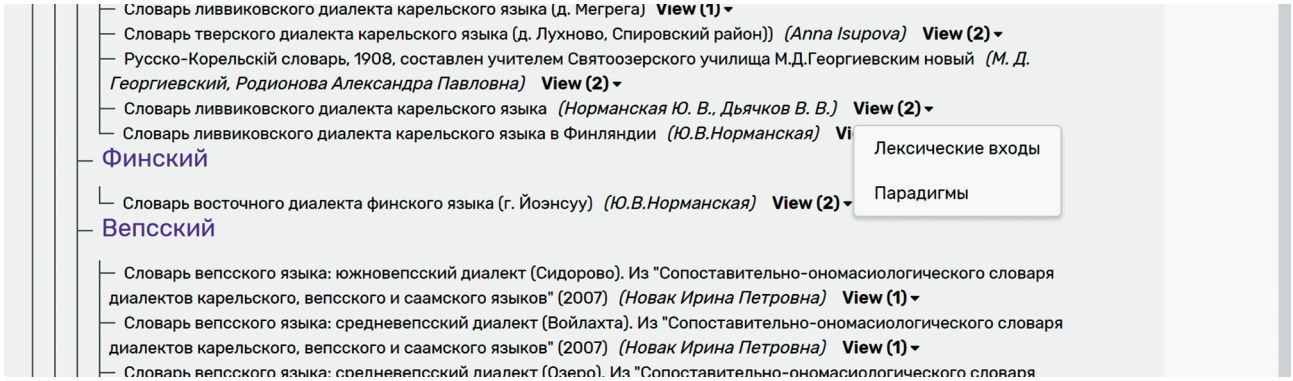

Рис. 2. Перспективы словаря.

рями, если источником CSV-файлов является Starling.

Каждый словарь имеет гибкую настраиваемую структуру. Пользователь может создавать лексические и парадигматические входы словаря и любые дополнительные перспективы с любым набором столбцов (полей), самостоятельно выбирая, какой тип данных в них будет размещен. Если в существующем списке нет подходящих столбцов, можно создать новый тип столбца, но в настоящее время в системе уже создано несколько десятков столбцов, например, звук, разметка, когнаты, парадигматические формы и контексты и другие (рис. 6). Могут быть следующие типы столбцов: 1) текстовые: пользователь видит возможность ввода текста; 2) аудио: пользователь видит селектор файлов для загрузки и кнопки для воспроизведения уже загруженных файлов. Возможно также создавать “вложенные” поля. Это полезно для зависимых данных, таких как файлы разметки спектрограмм аудиофайлов: к одному аудиофайлу можно присоединить иметь более одной версии разметки спектрограмм на отдельные звуки.

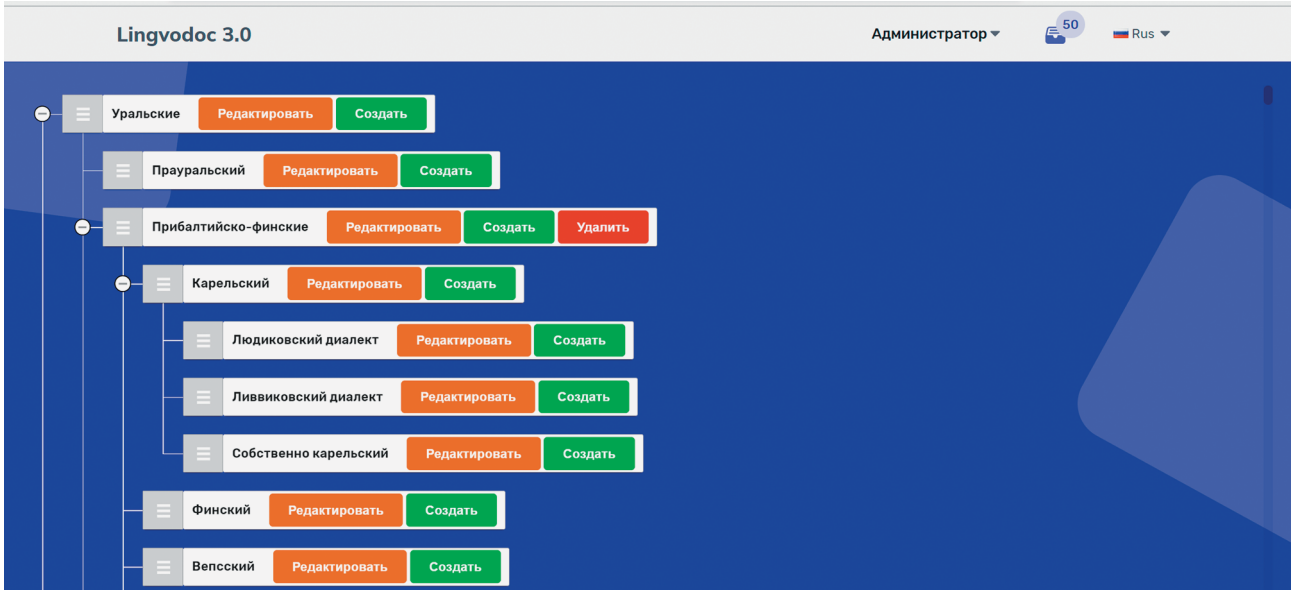

Рис. 3. Интерфейс структурирования языкового дерева.

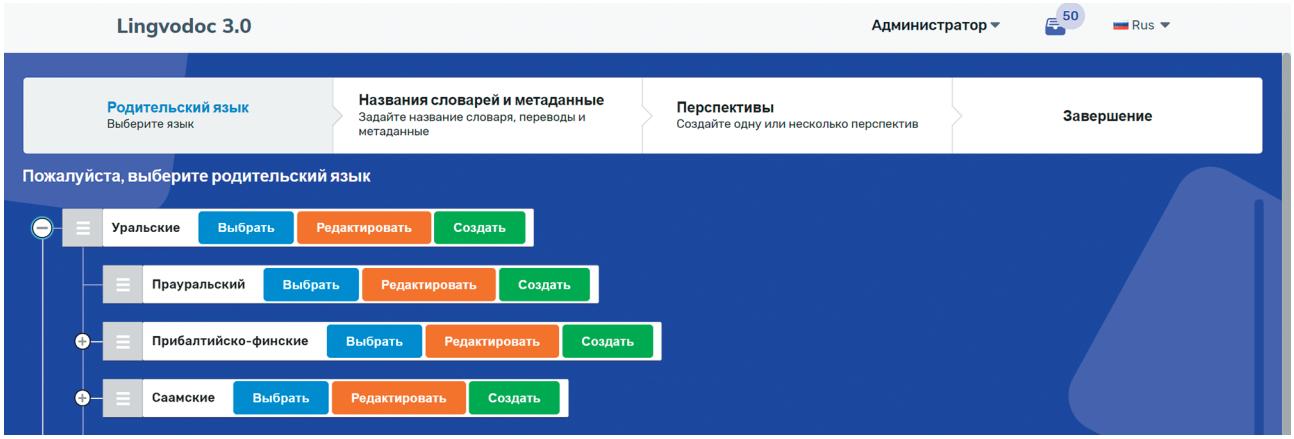

Рис. 4. Выбор языка в диалоге создания словаря.

После создания словаря пользователь становится единственным владельцем нового словаря, может редактировать его и давать другим пользователям право одновременной и независимой работы в этом же словаре. Если работа выполнена по гранту, пользователь передает часть своих прав, в основном, связанных с процессом публикации, зарегистрированному в системе гранту, фактически, его руководителю (гранты могут создавать системные администраторы по просьбе руководителей грантов).

Процесс редактирования очень похож на таблицы Excel с рядом отличий (рис. 7). Каждая ячейка в таблице может иметь множество отдельных значений (используемых для отдельных вариантов данных от разных авторов, например, звуков). В ячейки можно вставлять данные, соответствующие типу столбца: текст, звук, размеченные спектрограммы, ссылки, направленные и двунаправленные ссылки, позволяющие связать форму слова с другими словарями родственных языков. Каждая новая ячейка создается и заполняется в “неопубликованном” состоянии, поэтому ее могут видеть только те, у кого есть права на работу со словарем. Кроме того, редактирование данных выполняется в транзакционном режиме “пометить как удаленное и создать новое”, поэтому любое редактирование данных ячейки снова помечает их как “неопубликованные”.

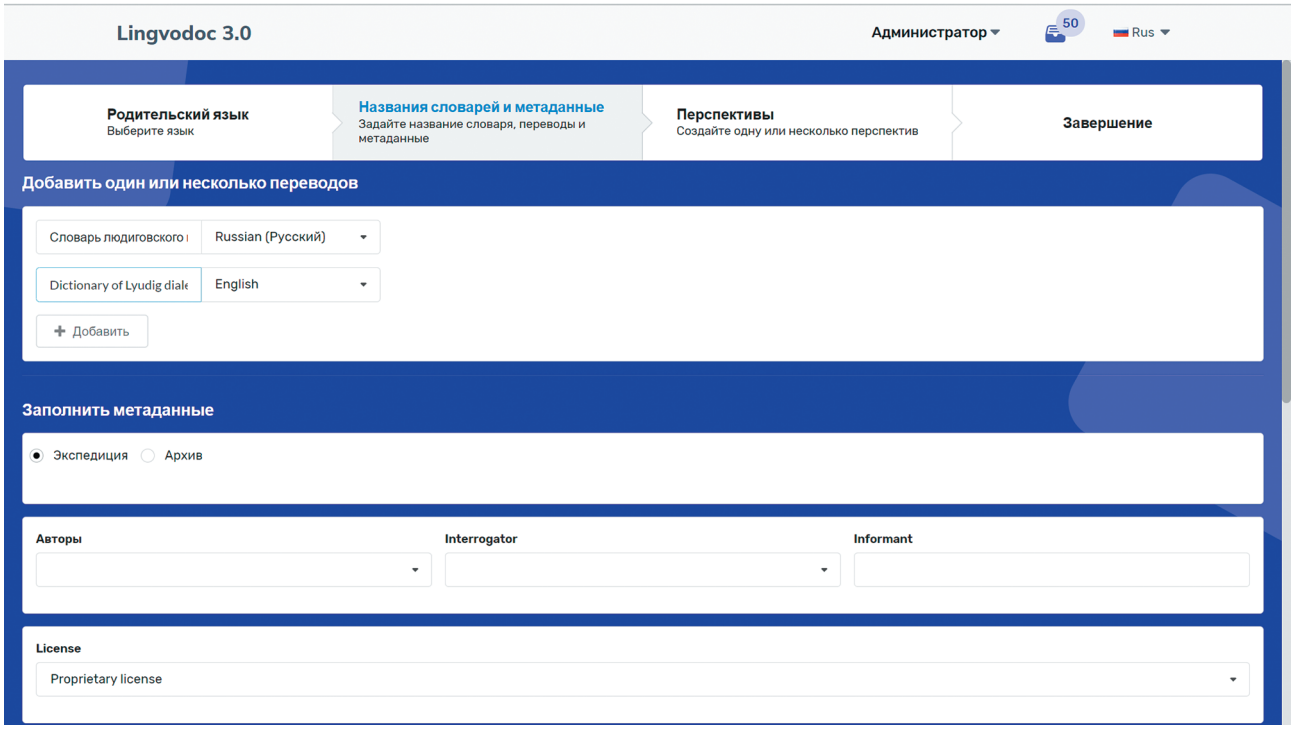

Рис. 5. Ввод описания словаря, который на предыдущем этапе был привязан к конкретному языку или диалекту на языковом дереве.

Каждый словарь и перспектива (лексические или парадигматические входы) имеют подробный список возможностей делиться правами разрешенных действий. Пользователи, имеющие разрешения на публикацию для лексических и/или парадигматических входов, могут изменить состояние каждой ячейки с неопубликованного на опубликованное и наоборот. Также существуют специальные разрешения, связанные с состоянием всего словаря. Каждый словарь/корпус, лексические и парадигматические входы в них могут находиться в следующем списке состояний: “Опубликовано” (доступно в полном объеме всем пользователям ЛингвоДока), “Скрыто” (доступно только тем пользователям, у которых есть права на редактирование словаря), “WiP” (доступно только тем пользователям, у которых есть права на редактирование словаря, но может быть задействовано в Поиске), “Ограниченный доступ” (пользователям ЛингвоДока открыта только первая страница, полный доступ только для тех пользователей, которые имеют права на редактирование). Опубликованные или словари/корпуса с ограниченным доступом становятся видны на странице “Языковые базы данных”: словари/корпуса.

Как показано на рис. 8, для каждого словаря/корпуса можно ввести метаданные: имена авторов обработки записанных данных, сборщиков материала, информантов, год сбора данных, название населенного пункта, где был записан материал, его географические координаты, проставив локализацию на карте, добавить файлы, связанные тематически со словарем в форматах pdf, docx и xlsx.

В LingvoDoc в меню “Статистика” у каждого словаря/корпуса хранятся данные о всех действиях его авторов. Можно просмотреть статистику изменений, сделанных пользователями, имеющими права на редактирование, в выбранном временном диапазоне. В словарях есть опция “Объединить лексические предложения”, в которой реализованы автоматический поиск и последующее объединение дублетных форм с разной степенью сходства во всех или в выбранном столбце, которую могут задавать авторы словаря.

В каждом словаре есть меню “Инструменты”, в котором в настоящее время авторам словаря доступно 13 программ (из них только 4 доступны всем пользователям системы) анализа языковых данных, которые воспроизводят фонетический, морфологический и этимологический анализ, ранее выполняемый лингвистами вручную, в десятки, а иногда в сотни раз быстрее. Подробнее о некоторых из них см. ниже в описании методов анализа обско-угорских языков.

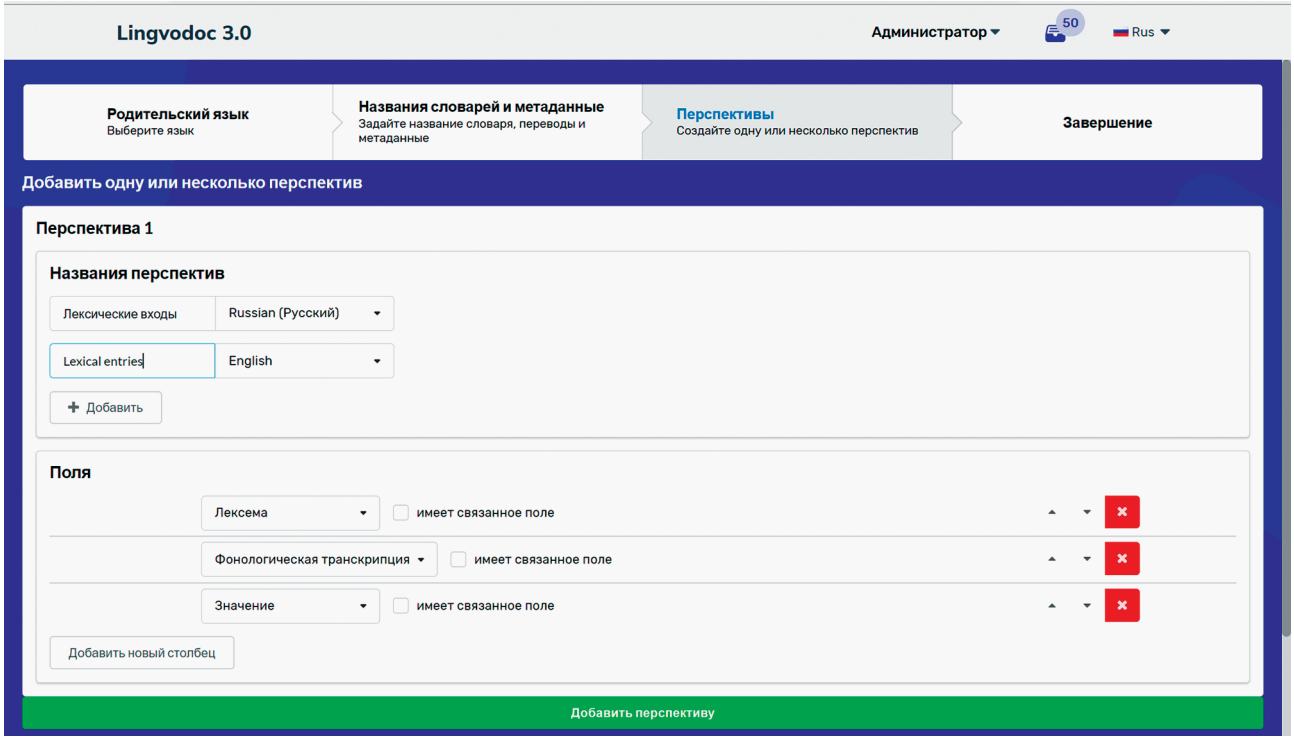

Рис. 6. Режим настройки структуры словаря: добавление столбцов.

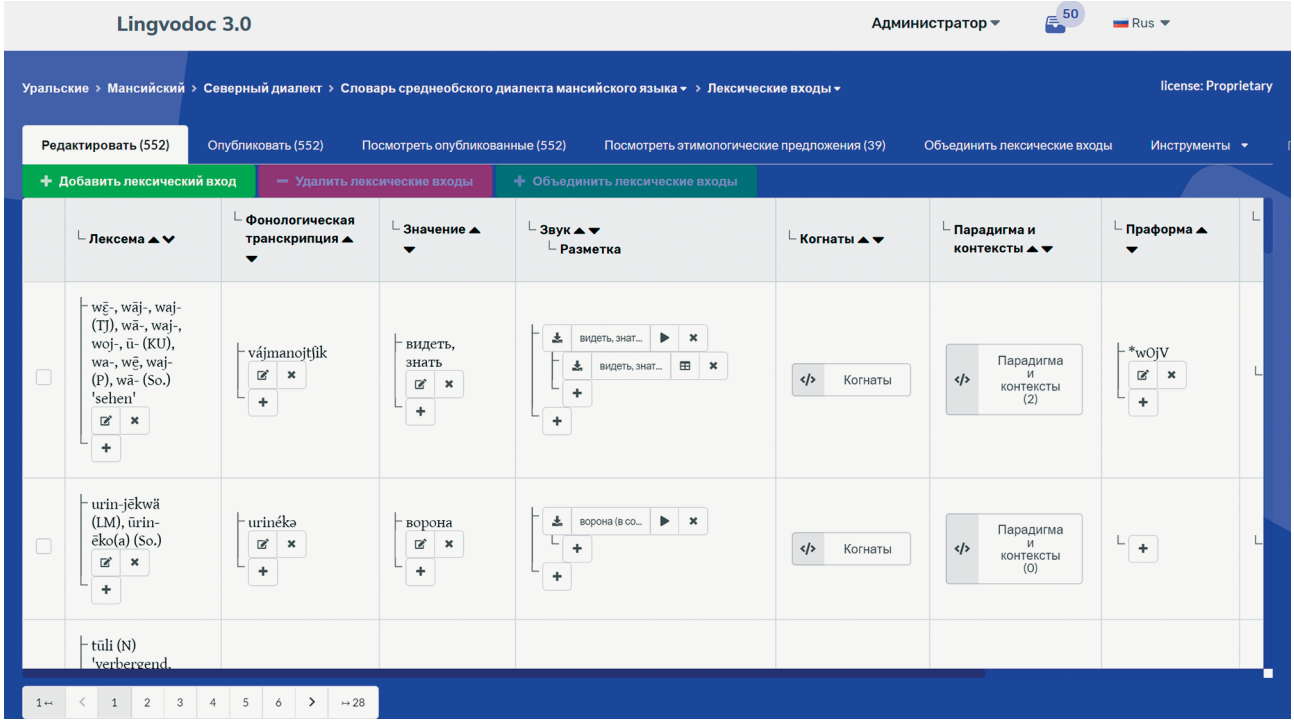

Рис. 7. Фрагмент перспективы Лексические входы в словаре обского диалекта мансийского языка.

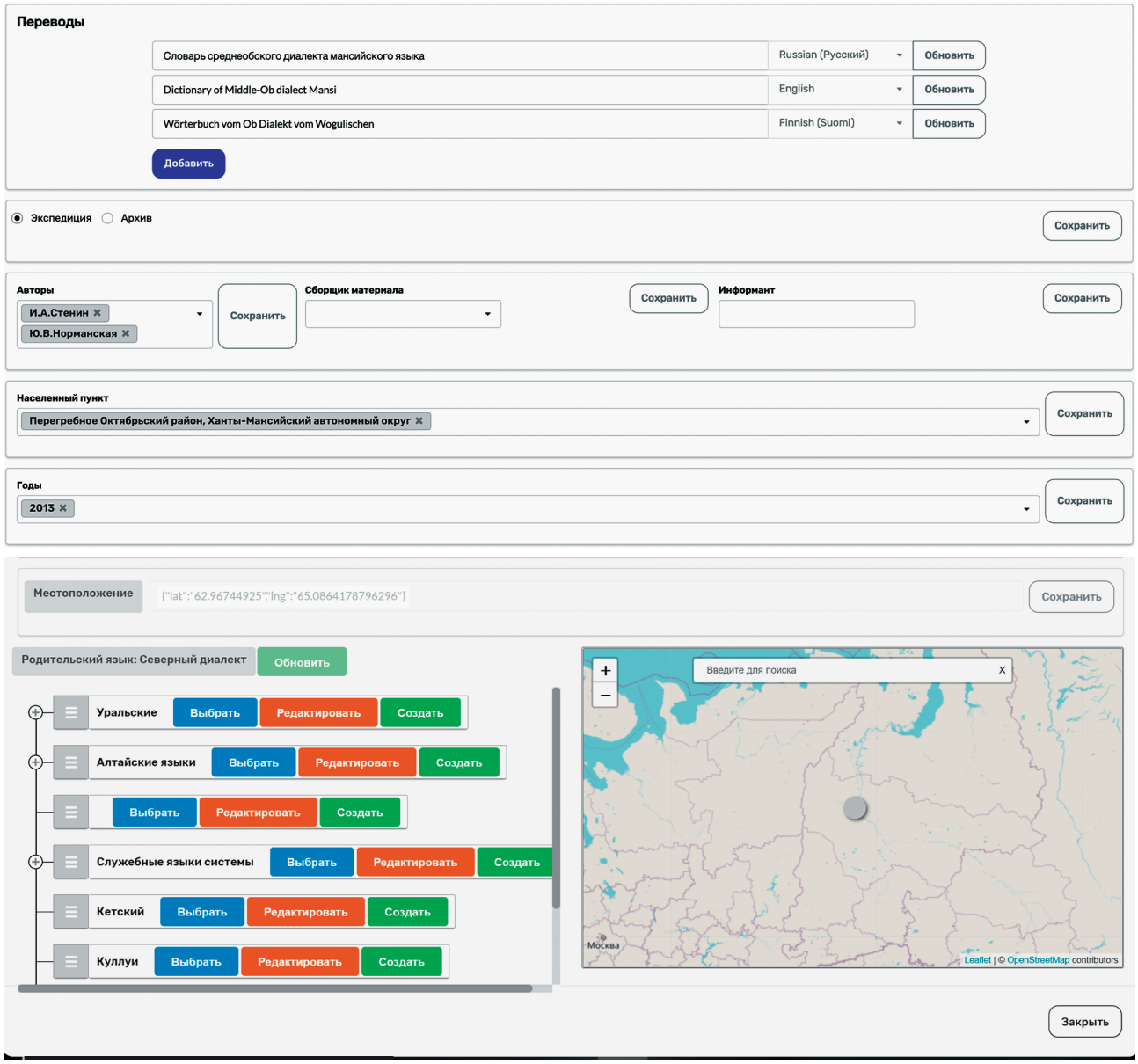

Рис. 8. Меню для ввода метаданных словаря.

В качестве корпусов текстов возможна загрузка текстов в формате .odt, звуковых или видеофайлов, и видео- или аудиокорпусов аннотированных/глоссированных текстов, подготовленных в программе Elan. Коллекция корпусов выглядит так же, как и обычный словарь, но имеет несколько заранее запрограммированных столбцов, которые при необходимости можно изменить: звук, звук с аннотацией, текстовый файл, комментарии. В LingvoDoc есть собственная веб-программа для просмотра корпусов в формате Elan (рис. 9). Также нами создана программа конвертации корпусов текстов Elan в конкордансы, где в лексическом входе представлена основа слова, а в парадигматических входах собраны все словоизменительные формы слова и контексты их употребления.

В системе также имеется диалоговое окно поиска, которое поддерживает запросы как по словарям, так и по корпусам в режимах ИЛИ/И (рис. 10). Это позволяет создавать сложные запросы с множеством опций. Пользователь может объединить несколько результатов поиска на карте мира, чтобы увидеть их пересечения.

LingvoDoc предоставляет бесплатную регистрацию в системе для новых пользователей (рис. 11). Для регистрации новому гостю системы достаточно нажать кнопку “Регистрация”, предоставить минимальную информацию о себе и до-

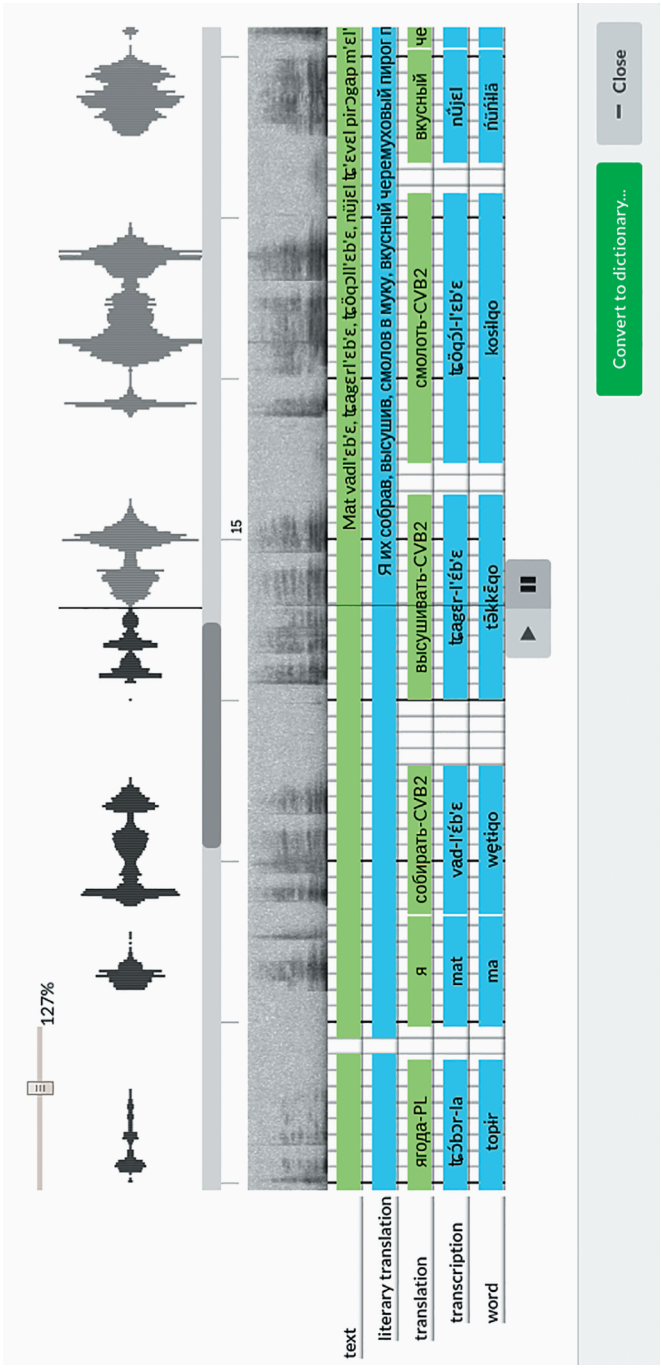

Рис. 9. Средство просмотра LingvoDoc корпусов текстов в формате Elan.

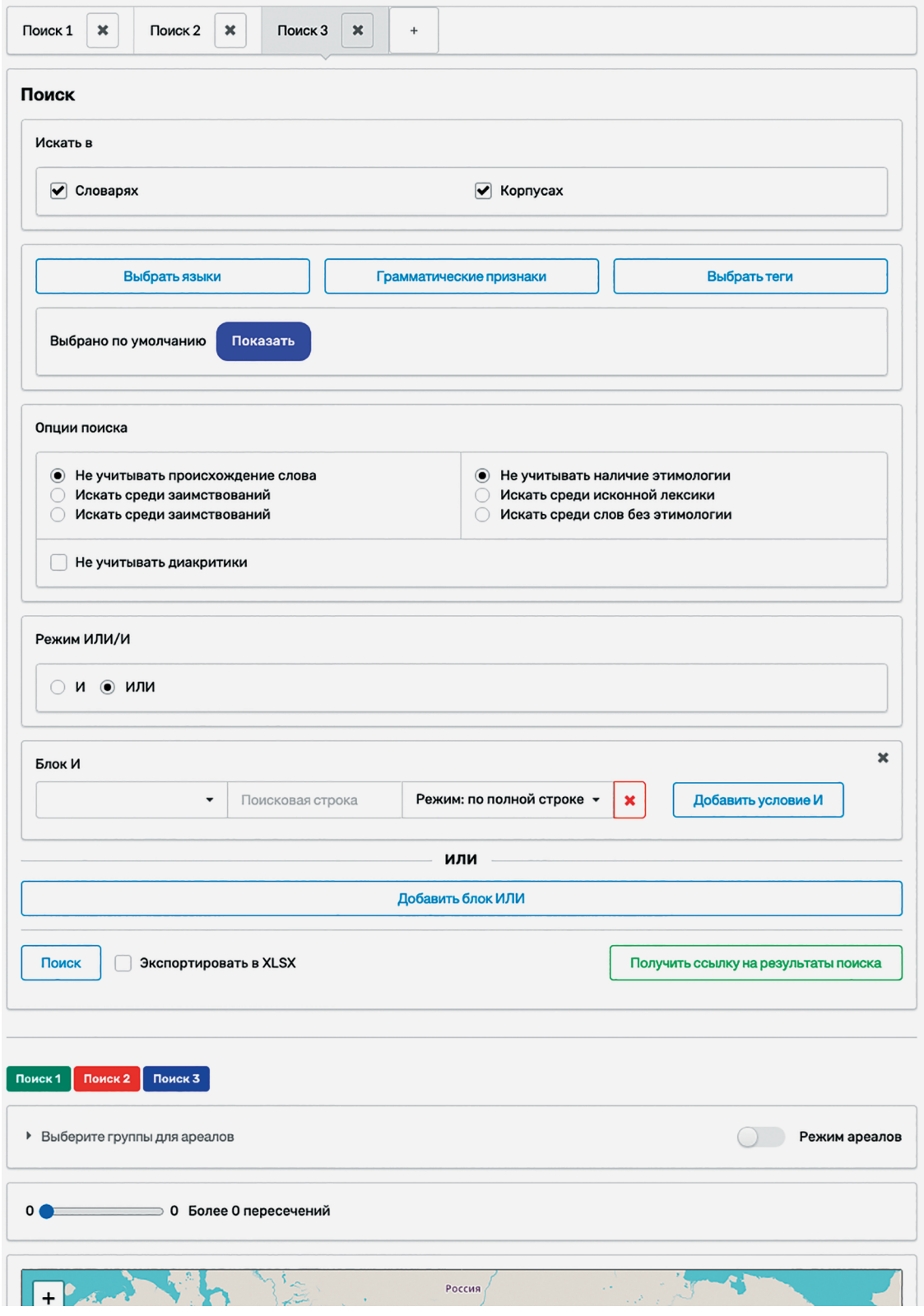

Рис. 10. Диалог поиска LingvoDoc.

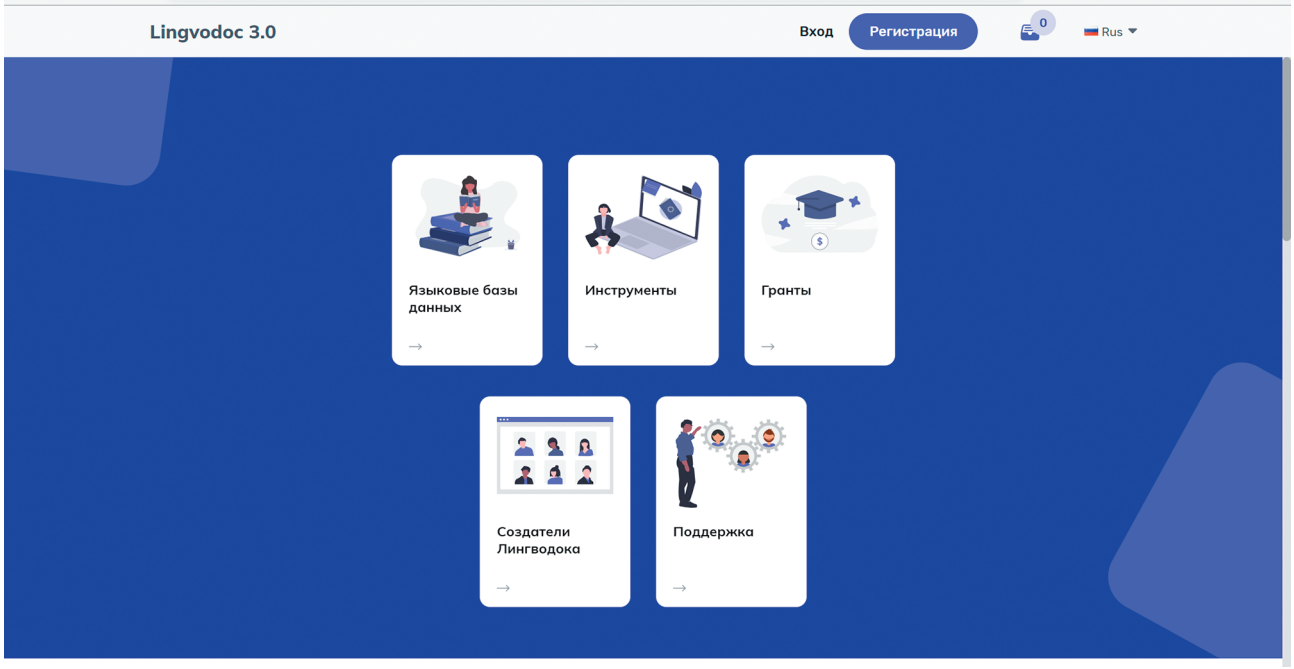

Рис. 11. Главная страница и кнопка регистрации.

ждаться, пока один из модераторов системы подтвердит, что это реальный человек, ранее не забаненный в системе.

3. ВНУТРЕННЕЕ УСТРОЙСТВО LingvoDoc

Программный комплекс LingvoDoc предоставляет следующие возможности:

1. Совместная работа над словарями (по аналогии с Google Docs или Github).

2. HTTP GraphQL API для интеграции с любым другим программным обеспечением.

3. Веб-интерфейс (эталонное клиентское приложение), использующий GraphQL API.

4. Гибкие списки контроля доступа (ACL) для совместного редактирования, просмотра и публикации. Каждый словарь в системе может использоваться совместно с любым другим пользователем системы, только для чтения, чтения-записи и публикации. Любой пользователь системы, не имеющий прямого доступа к конкретному словарю, может предлагать правки, которые могут быть рассмотрены редакторами словаря.

5. Многоязычные переводы словарей на основе одних и тех же данных. Все названия могут быть переведены в системе на любой язык.

6. Масштабируемая архитектура (предназначена для использования облачных ресурсов для масштабирования).

7. Полуавтономные клиенты с двусторонней синхронизацией. Пользователь может находиться практически где угодно и при этом синхронизировать свои данные, если он этого хочет и имеет подключение к Интернету.

8. Возможность создания **собственных порталов** с данными, принадлежащими группе пользователей или организации; система имеет двустороннюю синхронизацию с возможностями центральной системы.

9. Мультиарендность. Система изначально поддерживает полную изоляцию доступа среди участников словаря: один пользователь может получить доступ к отдельным словарям для личного использования, совместной работы и внутреннего использования, при этом каждый словарь размещается в его собственном учреждении и по желанию используется другими пользователями или учреждениями.

10. Безопасность. Мы не знаем паролей пользователей; система предназначена для хранения данных с использованием самых современных методов, чтобы обеспечить безопасность данных пользователей.

Исходные коды системы доступны под лицензией Apache 2.0.

Главной особенностью системы является поддержка полуавтономной синхронизации пользовательских данных. Эта функция уникальна для систем такого типа и основана на концепции составного первичного ключа [9]. Основная идея

такого типа синхронизации заключается в том, что каждый пользователь при каждом входе в систему (включая автономные установки клиента) получает уникальный для клиента составной ключ большого целого числа. После этого каждый объект, создаваемый конкретным клиентом, нумеруется на основе специального идентификатора последовательности. Во время каждого процесса синхронизации автономный клиент получает новый уникальный составной ключ персональной идентификации. Таким образом, каждый объект в системе имеет уникальную для объекта комбинацию ключа идентификатора клиента и ключа идентификатора объекта. Этот метод позволяет нам использовать концепцию "синхронизации в любое время": для каждого конкретного автономного приложения или процесса онлайн-входа клиента генерируется уникальный ключ идентификации объекта.

Несомненно, такие системы существуют уже давно, хотя почти все они требуют систем ручного разрешения конфликтов. Одними из самых известных примеров являются проекты GitHub и GitLab, основанные на системе контроля версий git. Эти проекты довольно просты в использовании, если вы один программист без необходимости синхронизировать свои проекты с чьими-либо еще. Но даже небольшой конфликт версий требует ручного разрешения конфликтов, что является сложным процессом и работает только с недвоичными данными. Наш подход очень похож на разрешение конфликтов в системе управления базами данных CouchBase [7], но используется для классических реляционных СУБД, и вполне вероятно, что мы изобрели его немного раньше, поскольку мы не можем найти никаких случаев применения этого метода до 2015 г.

Второй основной концепцией является виртуальный объект, который не содержит никаких данных, но служит якорем, на который ссылаются другие объекты. В системе LingvoDoc почти каждый объект базы данных имеет уникальную комбинацию идентификаторов в виде составных ключей. Словари, перспективы, столбцы, ячейки и строки в таблице являются основными примерами таких объектов. Строка в таблице – интуитивно понятная иллюстрация такого якоря. Данные, хранящиеся в ячейках таблицы, имеют ссылку на свой составной идентификатор строки в качестве привязки и организованы в таблицу с помощью внешнего веб-приложения. Внутренне данные организованы в более древовидной, чем табличной форме. У каждого автора может быть столько версий таких привязанных данных, сколько он хочет; система не ставит никаких ограничений. Несколько версий одного или многих авторов представлены в виде списка версий для объекта, и издатель несет ответственность за выбор того, какие из них являются правильными. Правильных версий иногда может быть много: например, если диктор трижды повторяет одно и то же слово, для одного лексического входа будет три правильных звуковых файла, и система отобразит их все.

В терминах реляционной модели данных это означает, что мы храним данные "версии" в денормализованной форме и объединяем данные на стороне сервера.

Представим, например, что у M авторов разные мнения по какому-то объекту (например, переводу определенного термина). Система не ограничивает количество перечисленных переводов при условии, что каждый из M авторов имеет соответствующие права на редактирование словаря.

Однако для этого в системе предусмотрены специальные режимы просмотра (см. обзор в первой части статьи). Основное представление, конечно же, – это представление редактора: оттуда редакторы данных могут делать все, что захотят. Вторая точка зрения – точка зрения издателя. Используя это представление, люди, ответственные за словарь, могут одобрить один или несколько правильных объектов, используемых в виртуальном объекте привязки. Например, если лексическая статья имеет 5 вариантов транскрипции и 10 вариантов перевода, и владелец этого словаря считает, что только одна из транскрипций и три перевода верны, он может выбрать только их и опубликовать свой выбор для остальных исследователей.

Последняя проекция – это представление просмотра/гостя/исследователя данных. Здесь вы можете увидеть данные, которые были загружены и проверены авторами и издателями.

4. API-ИНТЕРФЕЙС LingvoDoc GraphQL

LingvoDoc, безусловно, предлагает стандартный доступ через веб-интерфейс, что не представляет особого интереса с точки зрения параллельных технологий. **Полный доступ к системе LingvoDoc можно получить через систему GraphQL API (на основе HTTP).** GraphQL – это способ построения API для веб-приложений, предложенный Facebook в 2012 г. и широко используемый в настоящее время.

Основные особенности подхода GraphQL:

1. Пользователь API не видит внутренностей системы, а только упрощенные абстракции, которые намного проще и интуитивно понятны.

2. Все API системы можно исследовать и интроспектировать с помощью стандартных клиентов GraphQL, таких как расширение браузера Altair или любое другое.

3. Пользователь API может получить именно те данные, которые ему нужны в данный момент. Классические приложения REST API возвраща-

ют все данные для каждого метода или используют сложные селекторы данных для ограничения вывода; например, для метода /персона вы получите имя, дату рождения, фамилию, контактный номер и т.д. Тот же метод, реализованный при помощи GraphQL, позволяет выбрать, нужны ли им только фамилии и ничего больше.

Каждый объект в системе имеет четкий метод доступа; наш веб-интерфейс – это просто эталонный клиент, реализованный на JavaScript. Все уровни доступа доступны с помощью нашего простого API на основе HTTP. Все данные в системе доступны через API и возвращаются в формате JSON.

Большая часть данных хранится в СУБД PostgreSQL. Примечательной частью схемы базы данных является использование денормализованной схемы с составными первичными ключами. С помощью этого трюка мы добавляем в систему собственные типы данных. Таблицы в словарях не хранятся в виде таблиц внутри базы данных, а формируются backend API динамически с использованием таблиц, перечисленных на рис. 12.

Все данные, загружаемые пользователями в виде файлов, хранятся в файловой системе, совместимой с POSIX.

Бэкенд-часть LingvoDoc использует Pyramid в качестве веб-фреймворка, Graphene для построения GraphQL API, Celery для распределения задач, SQLAlchemy для связи с системой управления базами данных PostgreSQL.

Фронтенд-часть LingvoDoc – это классическое веб-приложение, построенное на платформе React и клиенте Apollo GraphQL.

На рис. 13 представлена схема взаимодействия.

5. АНАЛИЗ ОБСКО-УГОРСКИХ ЯЗЫКОВ

Обско-угорские языки: хантыйский и мансийский в ЛингвоДоке одни из наиболее описанных: в настоящее время в ЛингвоДоке доступны 18 словарей мансийских и 32 словаря хантыйских диалектов. Многие материалы по ним, например, аудиозаписи восточно-мансийского диалекта (с. Шугур), или говора п. Хулимсунт сосьвинского мансийского, говора с. Корлики ваховского хантыйского и многие другие – это единственные доступные ученым аудиозаписи этих говоров, которые значительно отличаются от остальных диалектов. Многие из говоров обско-угорских языков практически не описаны. Поэтому очень важно сохранить эти записи и провести их анализ на высоком научном уровне. Ниже будет показано, какие возможности есть для этого в ЛингвоДоке.

Еще в начале XX века обские угры: носители хантыйского и мансийского языков занимали огромную территорию от верховьев Печоры на севере Уральских гор до Томской области рек Юган, Васюган, Вах (около трех тысяч километров от северо-западного до юго-восточного регионов их расселения). Конечно, для языков, распространенных на такой огромной территории, была характерна значительная диалектная раздробленность. Хантыйский и мансийский языки разделены на четыре диалектных группы каждый, между носителями которых взаимопонимание отсутствует. В начале XX века в каждой диалектной группе было еще несколько диалектов, которые также значительно отличались друг от друга морфологическими и фонетическими системами.

К сожалению, в настоящее время ситуация катастрофически меняется. Часть диалектных групп уже исчезла: последние носители южных и западных диалектов мансийского языка умерли уже в середине XX века. Часть диалектов хантыйского языка: низямский (промежуточная группа между северными и южно-хантыйскими диалектами), салымский (промежуточная группа между западными и восточными хантыйскими диалектами) считались уже исчезнувшими, но полевые экспедиции, проведенные в рамках наших проектов, позволили найти нескольких последних носителей этих диалектов.

Большинство диалектов обско-угорских языков не имеют полных грамматик или словарей. Существующие описания не соответствуют современным стандартам или труднодоступны для ученых из России. Например, в описаниях, сделанных европейскими учеными в XIX–начале XX в., используются знаки финно-угорской транскрипции с множеством дополнительных символов, значение которых сейчас не всегда ясно. Также следует отметить, что описания, созданные в XIX в., могут значительно отличаться друг от друга по степени точности, ср., например, в Норманская (2015) анализ транскрипции мансийских говоров из одних и тех же населенных пунктов, сделанный Б. Мункачи и А. Каннисто, которые достаточно часто противоречат друг другу. В русских и советских кириллических транскрипциях мансийского языка, наоборот, практически не используются дополнительные символы. Перевод из одной системы транскрипций в другую невозможен.

В последние годы лингвисты из разных стран мира, понимая критическое состояние обско-угорских языков, активно занимаются их фиксацией и изучением. Упомянем лишь наиболее масштабные проекты:

• EuroBABEL Ob-Ugric Languages, **Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects** под руководством Е.К. Скрибник (http://www.babel.gwi.uni-muenchen.de/, см. на этом сайте также подробные ссылки на

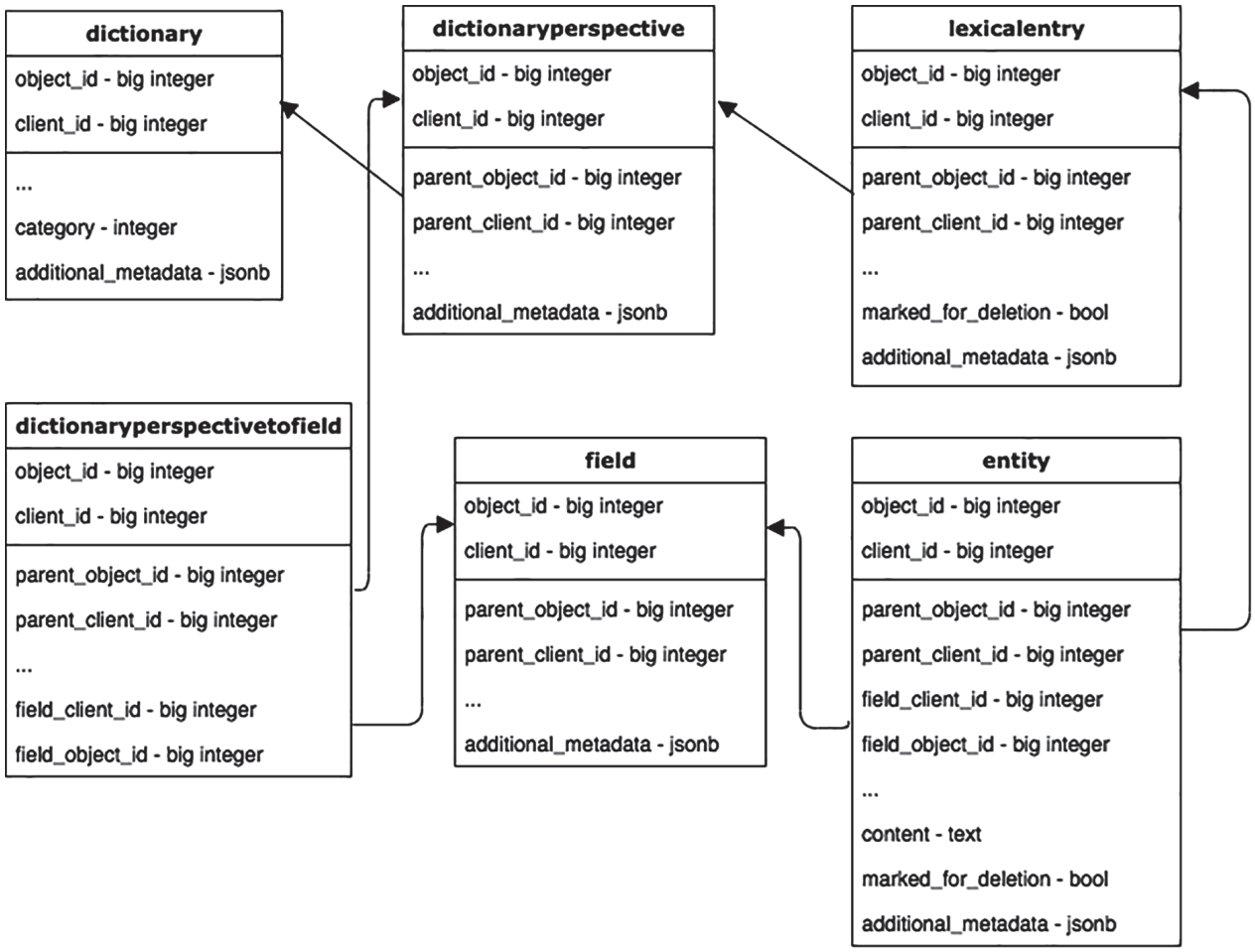

Рис. 12. Значимая часть схемы базы данных LingvoDoc.

различные ресурсы по обско-угорским языкам и этнографии). В этом проекте были собраны полевые материалы по казымскому, сургутскому и юганскому диалектам хантыйского и сосьвинскому диалекту мансийского, а также глоссированные тексты по северным, западным и восточным диалектам мансийского и северо-западным и восточным диалектам хантыйского.

• Multimedia documentation of the endangered Vasyugan and Alexandrovo Khanty dialects of Tomsk region in Siberia http://www.policy.hu/filtchenko/FTG%20ELDP%20project/audio.html под руководством А.Ю. Фильченко, в рамках которого были собраны и проанализированы восточно-хантыйские тексты.

Но, как видно, до начала нашей работы отсутствовали в открытом доступе звуковые данные по восточному диалекту мансийского языка, обскому говору северного диалекта мансийского, низямскому (промежуточного между северными и южными), салымскому (промежуточного между западным и восточными) диалектами хантыйского языка. Ранее фактически не привлекались к анализу многочисленные тексты, подготовленные в России на хантыйском и мансийском языках в XIX в. Мы нашли в архивах и библиотеках Санкт-Петербурга и Финляндии несколько десятков книг: Евангелия, богослужебная литература, словари на различных, частично уже исчезнувших диалектах хантыйского и мансийского языков. Начиная с 2012 г. наша группа провела многочисленные исследования различных диалектов обско-угорских языков: были организованы экспедиции в удаленные районы Западной Сибири, где с помощью представителей местных администраций были найдены последние носители мансийских и хантыйских диалектов. В 2012 г. С.В. Онина организовала экспедицию к последним носителям хантыйских диалектов, которые живут на реке Назым, и во время полевой работы были найдены последние носители низямского диалекта хантыйского языка, который ранее считался исчезнувшим. В 2013 г. М.К. Амелина организовала экспедицию к последним носителям юкондинского диалекта восточно-мансийского

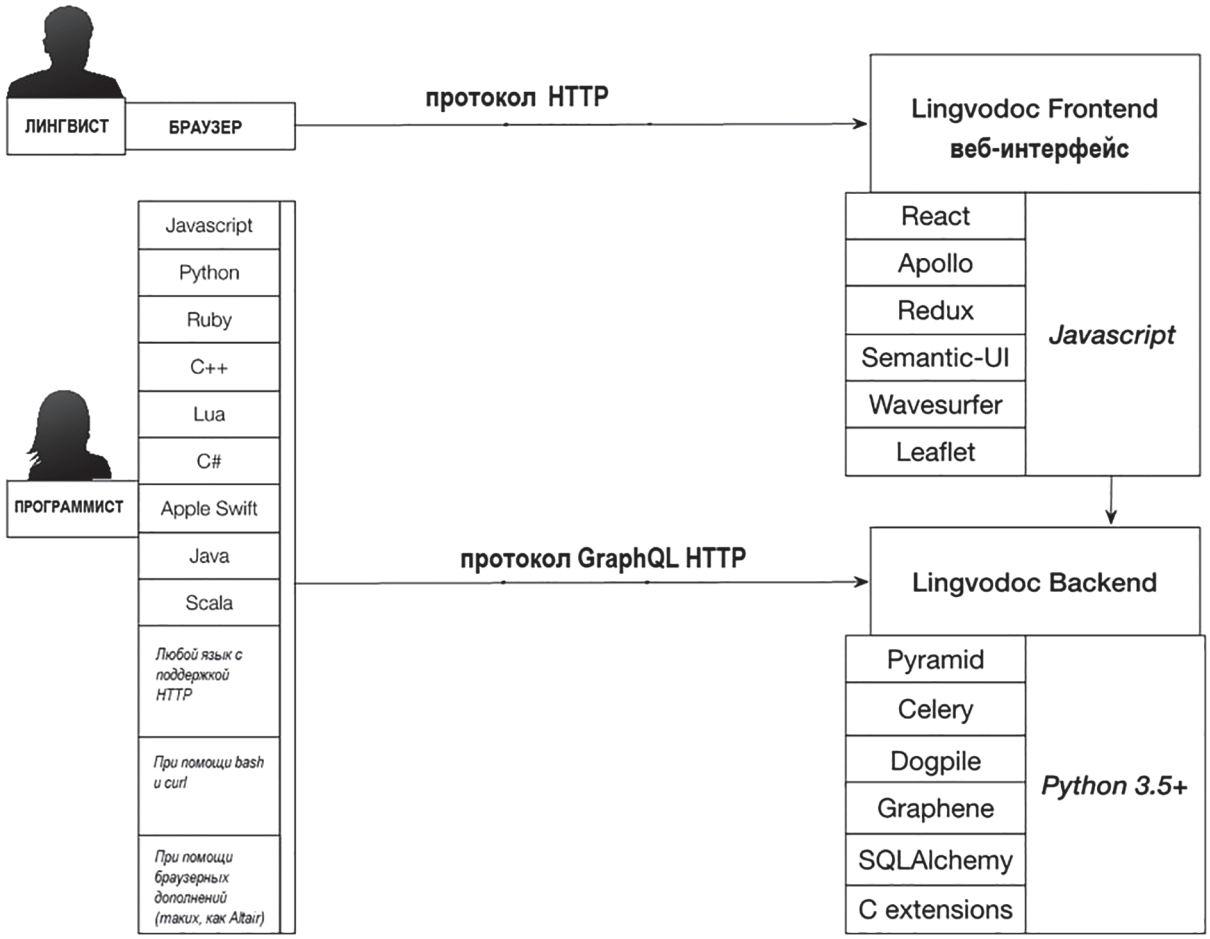

Рис. 13. Схема взаимодействия с LingvoDoc.

языка в д. Шугур Кондинского района ХМАО[1], И.А. Стенин собрал полевой материал от носителей обского мансийского диалекта, живущих в двух деревнях Октябрьского района ХМАО: Нижние Нарынкары и Перегребное.

Так, LingvoDoc позволяет каждому исследователю, имеющему полевые аудиозаписи, создать собственные мультимедийные словари, соединив в системе звук со спектрограммами Праата. Далее возможен анализ собранных данных с точки зрения экспериментальной фонетики, диалектной принадлежности, этимологии. В настоящее время это возможно делать он-лайн на сайте http://lingvodoc.ispras.ru/ (подробнее техническую информацию об этих возможностях см. выше в первой части настоящей статьи).

Эти возможности делают LingvoDoc необходимым специалистам по исчезающим языкам и диалектам, поскольку обычно в словарях этих языков приводится только транскрипция слов, которую в большинстве случаев невозможно проверить. Но, как показали наши исследования, даже в наиболее известных словарях уральских языков очень часто встречаются ошибки в транскрипциях, например, в словаре Munkácsi, Kálmán (1986), несопоставимость которого со словарем Kannisto (2015) была упомянута выше; в диалектологическом словаре селькупского языка Быконя (2005), где одно и то же слово в любом говоре может быть записано 2–3 способами на разных или даже на одной странице словаря. Часто уже нет возможности проверить, какой из вариантов транскрипции правильный, потому что носителей восточных и западных диалектов мансийского языка, южных и центральных диалектов селькупского языка уже практически нет. Программы, которые мы используем в ЛингвоДоке, позволят как будущим исследователям, так и носителям обско-угорских языков услышать, как произносилось слово на тех диалектах, которые через 10–20 лет уже исчезнут, и сравнить транскрипцию словарей с результатами, полученными после анализа звука в экспериментально-фо-

[1] Два последних носителя умерли в 2013 и 2017 г.

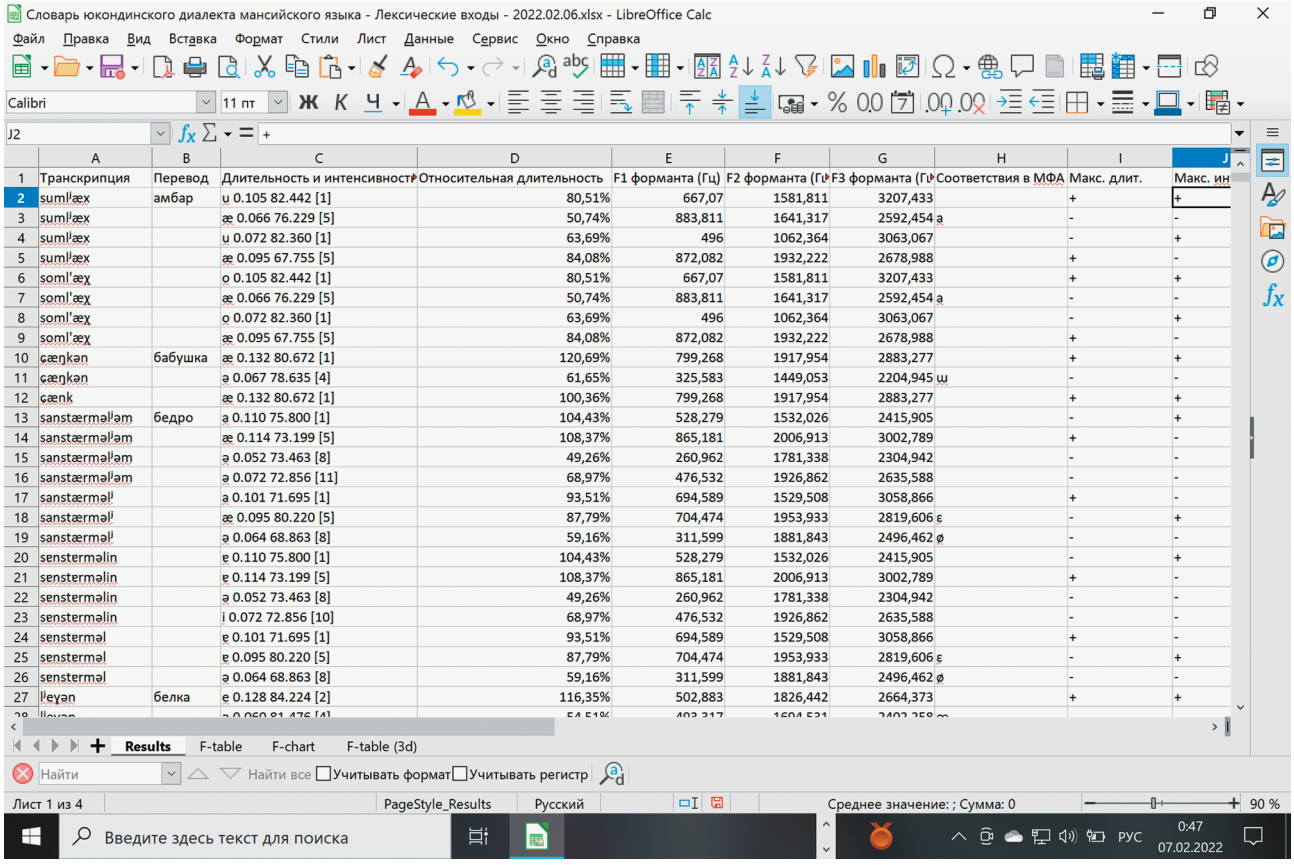

Рис. 14. Пример таблицы с фонетическими параметрами гласных в юкондинском диалекте мансийского языка после обработки словаря в программе “Результаты анализа спектрограмм”.

нетической программе Праат. Тот факт, что каждый пользователь словаря сможет не только посмотреть картинки уже размеченных спектрограмм, но и, скачав, проверить правильность их обработки, значительно повысит надежность обработки данных и позволит создать среду, в которой возможно достижение взаимопонимания между специалистами всего мира, которые используют разные системы транскрипций. Поскольку как словари, так и программы для их обработки доступны он-лайн, становится возможным любому ученому получить доступ к материалам, и в он-лайн режиме обсудить правильность той или иной транскрипции.

Как было сказано выше, в настоящее время в ЛингвоДоке доступны 18 словарей мансийских и 32 словаря хантыйских диалектов. Каждый из доступных он-лайн словарей обско-угорских языков насчитывает 600–1000 лексем с парадигматическими формами. Каждый словарный вход содержит:

1. Начальную форму слова, которая представлена: 1) формой из словаря соответствующего диалекта (в традиционной орфографической записи), 2) фонологическую или фонетическую форму слова, 3) для словарей, материал которых собран в полевых условиях, аудиофайл с произнесением начальной формы три раза, 4) спектрограмму, созданную в Праате, содержащую данные о физических характеристиках каждого звука в слове (интенсивности, длительности, частоте, тоне), которую можно скачать, и проверить правильность проведенной обработки в Праате.

2. К каждой начальной форме, по возможности, прикреплены: парадигматические формы слова в изолированном произношении и в контекстах. Каждая из форм парадигмы представлена тем же способом, что и начальная.

3. Каждая начальная форма, по возможности, соединена этимологическими связями с родственными словами из других диалектных словарей. Их можно сделать в тех случаях, когда авторы словарей родственных диалектов дали создателю словаря согласие на соединение этимологическими связями. При нажатии на кнопку “Когнаты” можно увидеть список родственных слов из других словарей.

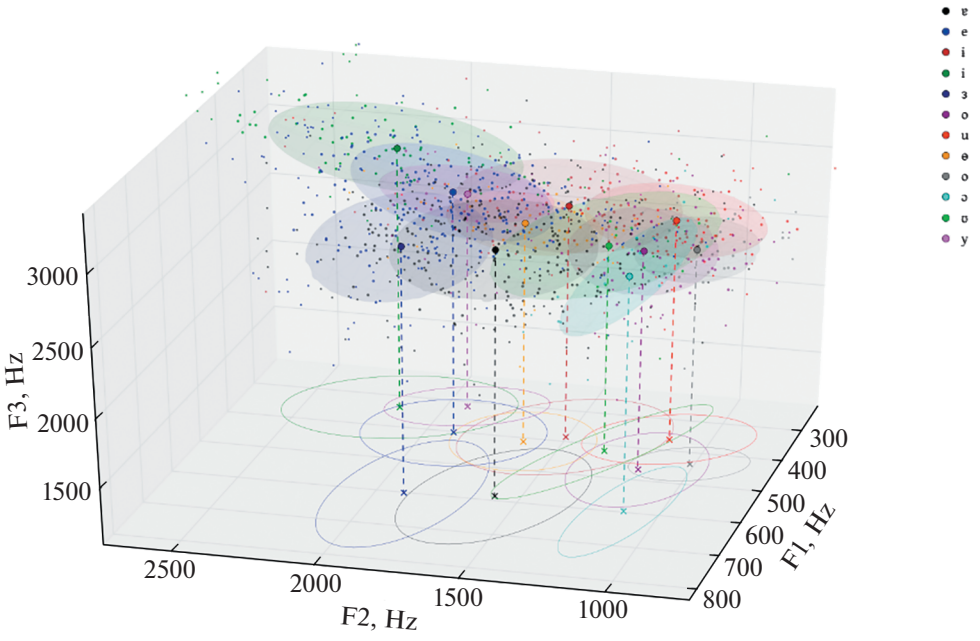

Рис. 15. Пример обработки с помощью программы “Результаты анализа спектрограмм” правильно выполненной транскрипции ваховского диалекта хантыйского языка http://lingvodoc.ispras.ru/dictionary/1230/570/perspective/1230/571/view.

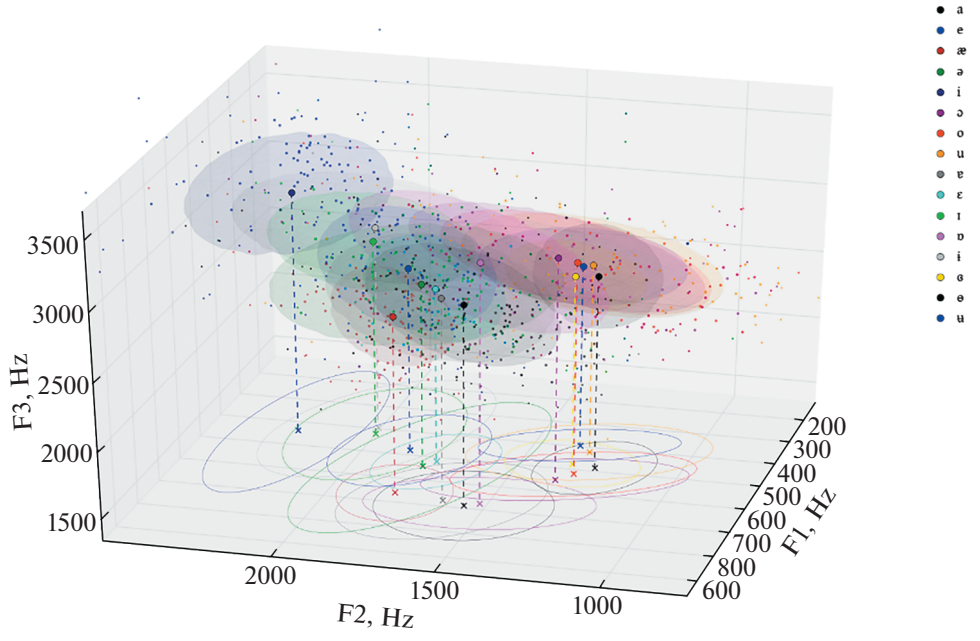

Рис. 16. Пример обработки с помощью программы “Результаты анализа спектрограмм” некорректно выполненной транскрипции юкондинского диалекта мансийского языка http://lingvodoc.ispras.ru/dictionary/1230/570/perspective/1230/571/view.

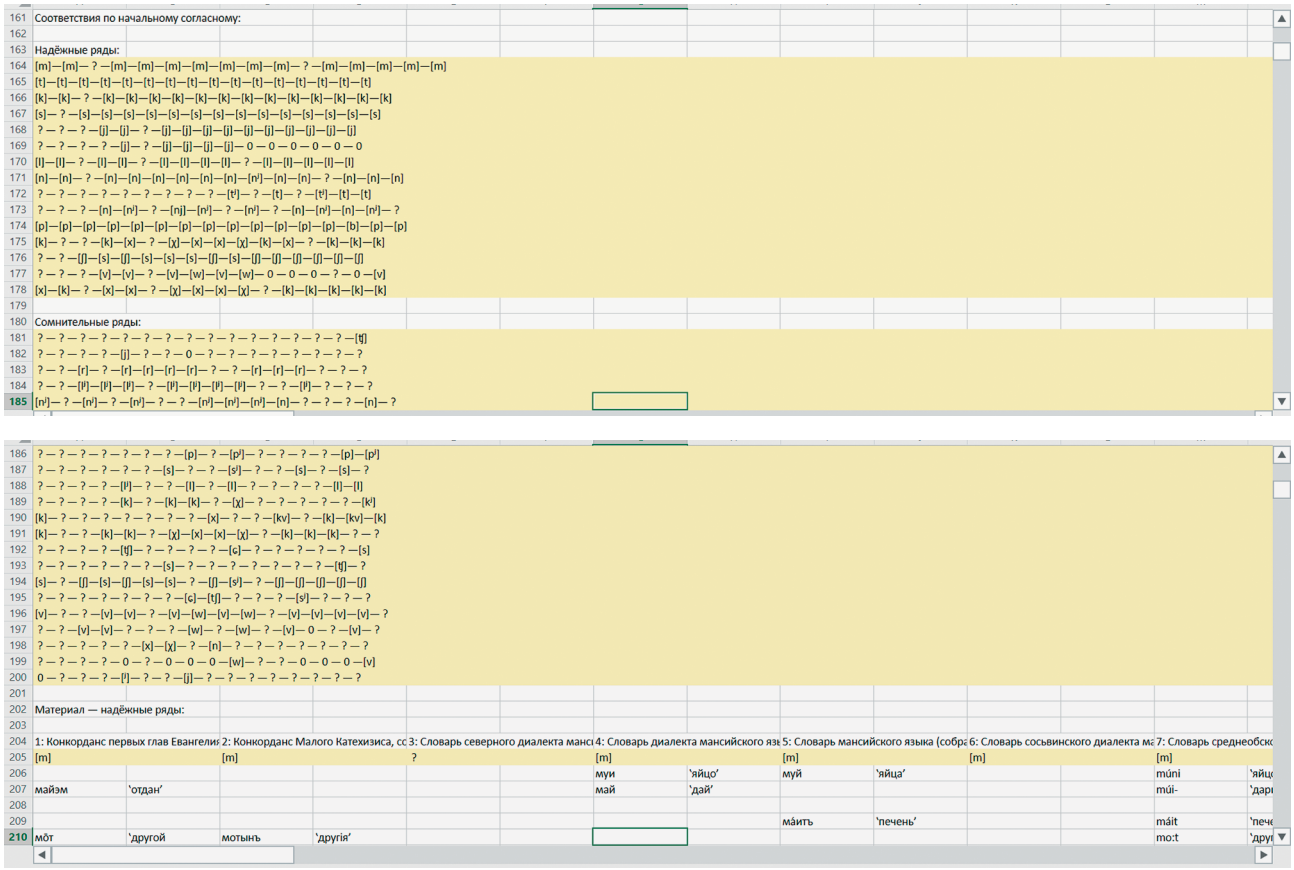

Рис. 17. Пример фрагмента автоматически полученных соответствий в мансийских диалектах: согласные в позиции начала слова.

В каждом словаре при выборе опции **“Инструменты”** можно увидеть список программ для анализа языкового материала и, в частности, выявления ошибок в его фонетической и этимологической обработке.

Программа LingvoDoc “Результаты анализа спектрограмм”, позволяющая оценить правильность транскрипции и описать системы гласных в диалектах с точки зрения экспериментальной фонетики.

На LingvoDoc есть возможность комплексного анализа фонетических характеристик гласных всех лексем из выбранного словаря, полученных в результате обработки их спектрограмм в фонетической программе Праат. Для этого надо выбрать в открытом словаре того или иного диалекта опцию Инструменты > Результаты анализа спектрограмм. Для использования этой программы в словаре должны быть загружены звуковые файлы в формате .wav и их спектрограммы, где в Праате поставлены границы отдельных звуков в формате textgrid.

Программа “Результаты анализа спектрограмм” для каждого гласного звука в выбранном словаре собирает его физические характеристики во всех произнесениях, обрабатывая спектрограммы, размеченные предварительно в фонетической программе Праат на отдельные звуки. На выходе получается таблица со списком всех физических характеристик каждого звука в множестве лексем выбранного диалекта, см. рис. 14.

При создании этой таблицы используются алгоритмы анализа интенсивности звука и формантных характеристик Праат, см. подробнее [8], и разделы Sound: To Intensity... и Sound: To Formant (burg)... в справочнике Праата, доступном на Praat’s website.

В результате работа программы “Результаты анализа спектрограмм” эта таблица Эксель с указаниями физических параметров каждого звука также преобразуется в 3D-график, построенный на основании показателей трех формант для каждого гласного звука. На графике произнесения каждого гласного оформлены как элипсы разных цветов, см. рис. 15.

Как показано на рис. 16, если транскрипция некорректная, облака формант разных гласных будут значительно пересекаться.

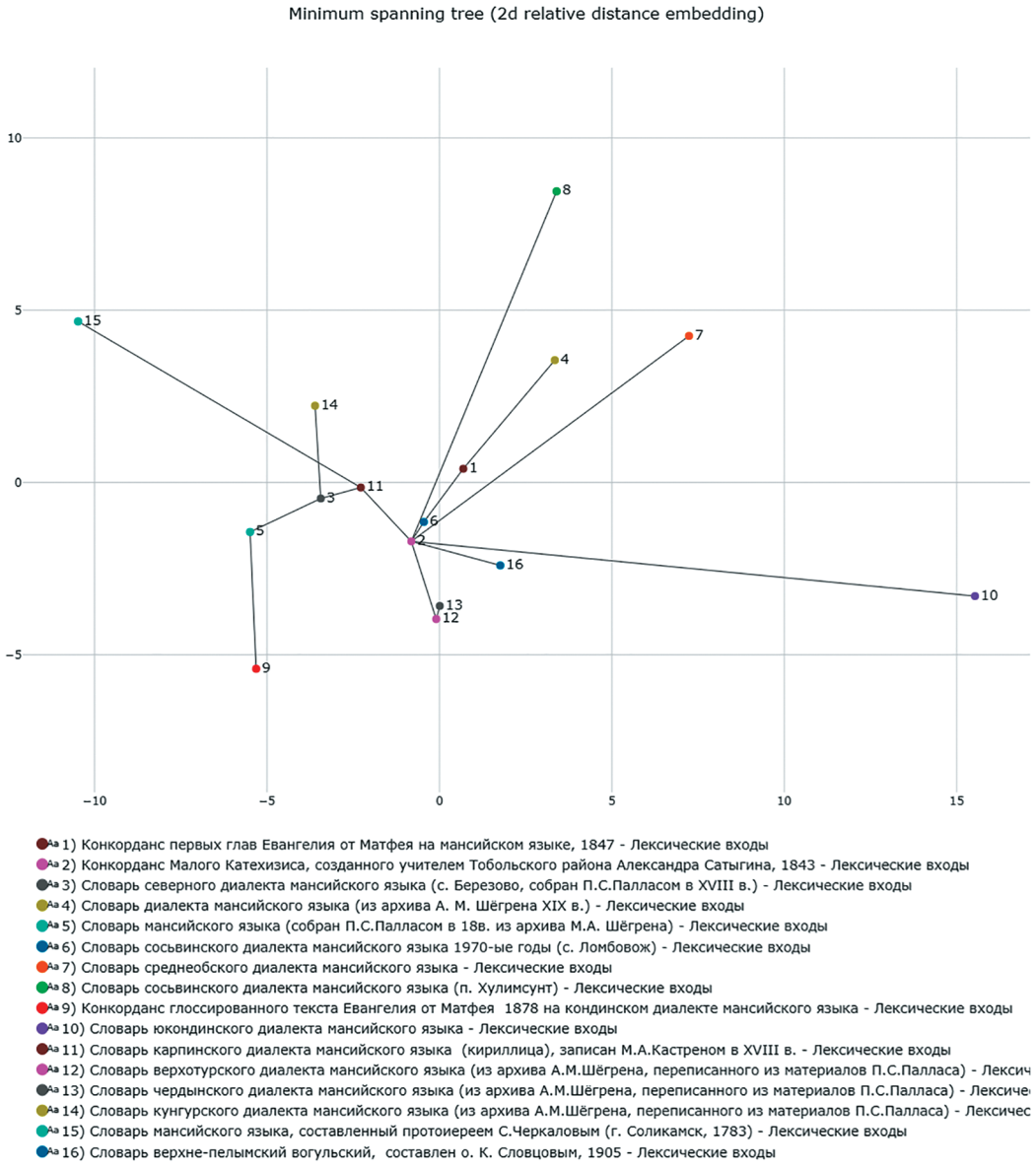

Рис. 18. Пример 2D-графика близости мансийских диалектов, полученного в программе “Анализ конатов”.

Так, обработка аудиословарей в программе “Результаты анализа спектрограмм” позволяет математически точно оценить правильность фонетической транскрипции, принятой для каждого из диалектов, и существенно уточнить ее. Далее, уже опираясь на уточненную фонетическую транскрипцию, можно перейти к следующему этапу: анализу степени близости родственных диалектов.

Этимологические программы LingvoDoc “Поиск когнатов в разных диалектах / языках”, “Анализ когнатов в разных диалектах/языках”, позволяющие уточнить системы регулярных соответствий звуков и оценить степень близости разных языков и диалектов[2].

В настоящее время на LingvoDoc представлены не только словари современных диалектов, собранные в полевых условиях, но и архивные материалы XIX в., созданные в результате деятельности Переводческой комиссии Православного Миссионерского Общества, учрежденной при Братстве Св. Гурия. Анализ и сопоставление архивных данных с современными диалектами

[2] Эти опции доступны только тем пользователям, которые либо сами создали один из словарей, либо получили права на работу с ним.

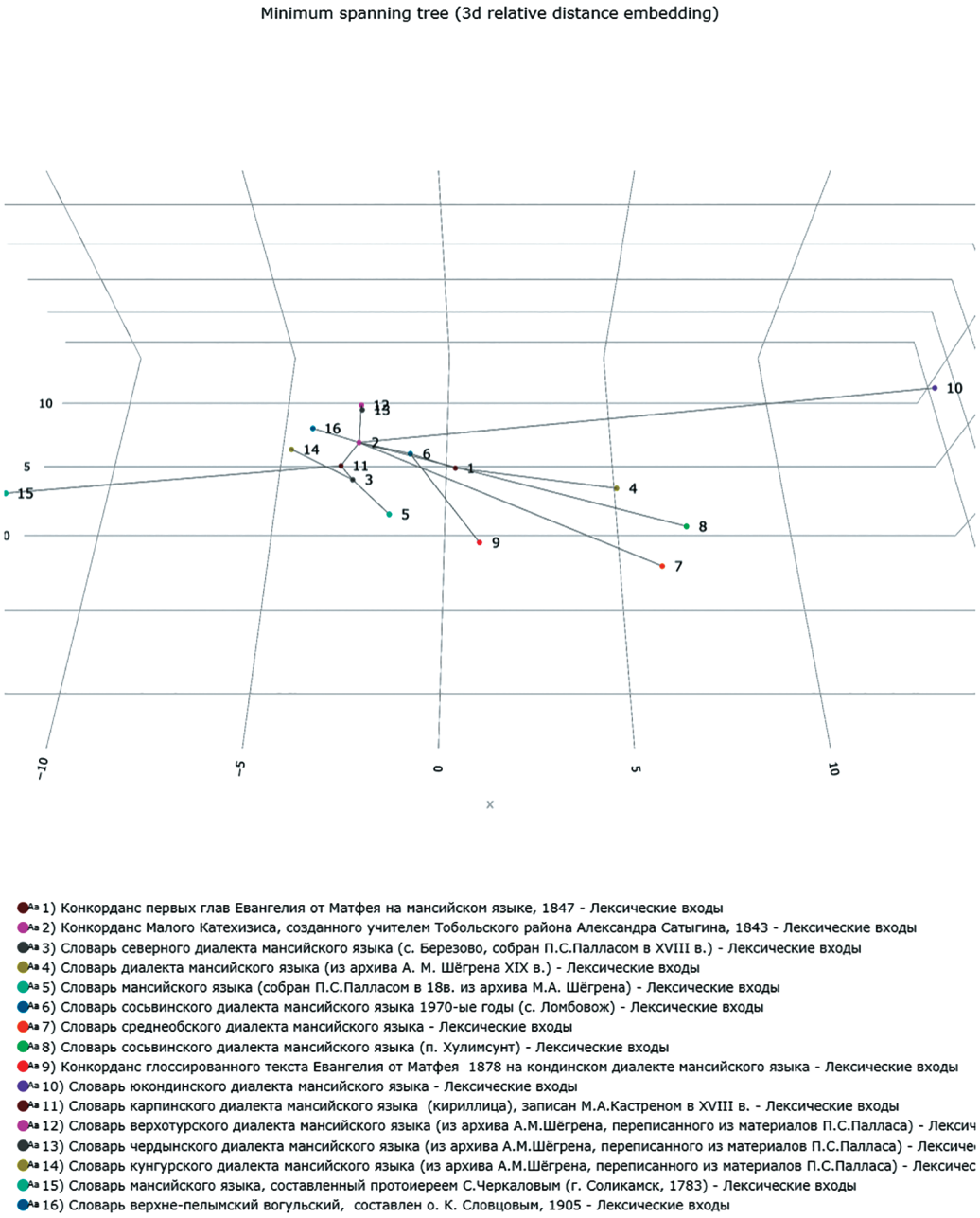

Рис. 19. Пример 3D-графика близости мансийских диалектов, полученного в программе “Анализ когнатов”.

показывают, что ошибки в рассматриваемых памятниках практически отсутствуют, они имеют регулярные соответствия с наиболее точными словарями обско-угорских языков. Это связано с тем, что, как мы знаем из истории их создания, первые книги неоднократно выверялись носителями соответствующих диалектов. Поскольку ранее эти тексты не были включены в научный оборот, то весьма важной задачей является их диалектная классификация.

1.“Фонетические соответствия”

Для автоматического проведения этого анализа на первом этапе необходимо выбрать несколько словарей родственных языков или диалектов для пословного соединения этимологическими связями с помощью разработанной нами программы “Поиск когнатов в разных диалектах/языках”. В результате работы этой программы появляется список потенциально родственных слов, состоящий в среднем из 5–8 тысяч предложений возможных когнатов в разных язы-

Рис. 20. Фрагмент автоматической реконструкции прамансийского языка на основании сравнения 16 диалектов.

ках/диалектах, которые далее необходимо вручную утвердить как родственные или отклонить. Результат этой ручной обработки можно увидеть в каждом словаре при нажатии кнопки Когнаты.

Затем нужно открыть словарь, который будет оценен с точки зрения его места в диалектной классификации, и выбрать в меню Инструменты программу “Анализ когнатов в диалектах/языках”. В следующем меню будет возможно галочками отметить те словари, которые планируется автоматически сравнить с выбранным. В результате работы этой программы мы получаем ряды соответствий фонем в выбранном словаре звукам в других диалектах или языках в разных позициях: в начале слова, после первой согласной, после первой гласной и т.д., см. рис. 17.

2. “Диалектная классификация”

Далее для каждой пары различных фонем в одном ряду соответствий подсчитывается, сколько переходов необходимо по Международному фонетическому инвентарю для изменения фонемы X в фонему Y, каждому переходу присваивается равный вес. Таким образом, в выбранной паре диалектов или языков обсчитываются все ряды соответствий между ними и степень различия фонем в этих рядах. В результате в файле Экселя строится матрица близости всех выбранных диалектов или языков, которая отображается также на он-лайн 2D и 3D-графиках, см. рис. 18, 19.

3. Реконструкция рядов соответствий фонем и слов праязыка

Как показано ниже на рис. 20, в результате обработки набора словарей родственных языков или диалектов с помощью программы “Анализ когнатов” мы получаем систему соответствий и реконструкцию праязыковых фонем и слов.

В результате этой работы на материале полного анализа словарей мы получаем информацию о фонетической близости генетически родственных диалектов и языков между собой. В LingvoDoc для этого используются алгоритмы, построенные на основе принятого в лингвистике “принципа регулярных соответствий”, а не на поверхностном сходстве словоформ, как в настоящее время принято в большинстве проектов, посвященных определению близости родства: LingPy, EVOLAEMP. Известно, что поверхностное сходство, скорее, указывает на наличия заимствований, чем на глубокое родство, поскольку доказательством родства языков являются лишь многочисленные нетривиальные регулярные соответствия между согласными и гласными звуками. Единственный на-

дежный автоматический способ оценки степени близости родства языков – это поиск правил: рядов “регулярных соответствий”, которые преобразуют слова между двумя языками, или из прото-языка в дочерний язык, и последующий анализ этих рядов. Ранее именно таким образом работали ученые-этимологи, обрабатывали материал вручную. Теперь это можно сделать в сотни раз быстрее с помощью компьютерных программ LingvoDoc. Очень важно, что эти результаты полностью верифицируемы, поскольку в результате работы программ получается файл Экселя объемом несколько сотен страниц, в котором для каждого ряда соответствий приведены все примеры, его подтверждающие. Ранее при ручной обработке такой материал никогда не был доступен, вероятно, из-за трудоемкости его получения. Значительно более высокая степень точности этой этимологической работы в LingvoDoc связана еще с тем, что на вход этимологического анализа поступают не транскрипции, сделанные авторами словарям на слух, а непосредственно звук, и уточнение транскрипции в результате его обработки с помощью экспериментально-фонетических программ, является одним из этапов анализа.

Таким образом, становится ясно, что без возможностей LingvoDoc качественное и полностью верифицируемое описание и этимологизация обско-угорских словарей невозможны. Еще раз подчеркнем, что этот ресурс позволит и будущим исследователям получить доступ непосредственно к звуковым записям и полностью самостоятельно воспроизвести процессы 1) правильности транскрипции диалектов, 2) определения степени близости между диалектами.

Результаты, которые мы получили, привлекая значительное количество новой информации, собранной как в полевых условиях, так и в архивах по обско-угорским языкам, позволили в очень сжатые сроки новые результаты, важные как для лингвистики, так и изучения истории России, поскольку они позволяют уточнить степень генетической близости носителей разных диалектов, существовавших в течение последних 250 лет, и время распада хантыйского и мансийского праязыков на разные диалекты.

6. ЗАКЛЮЧЕНИЕ

В статье представлена виртуальная лаборатория ЛингвоДок, которая позволяет проводить документацию языков с помощью создания словарей любой структуры, этимологических связей между ними и программ их анализа. В ЛингвоДоке не уровне интерфейса также возможно использование результатов обработки звука в программах Элан (глоссирование звучащих текстов) и Праат (экспериментально-фонетическая обработка). Бэкенд предоставляет интерфейсы взаимодействия при помощи GraphQL HTTP API. Доступные вызовы могут быть интроспектированы стандартными утилитами работы с GraphQL и позволяют использовать данные, хранящиеся в системе для целей вычислительного анализа снаружи самой системы. Проект является открытым и лицензируется по лицензии Apache 2. В настоящее время в системе зарегистрировано более 300 активных пользователей.

Мы показали, как программы анализа можно применять к материалу словарей обско-угорских языков и получать новые значимые результаты по экспериментально-фонетическому описанию отдельных говоров, генетической близости диалектов, реконструкции праязыка.

В будущем мы планируем улучшить версию, размещаемую локально в других местах с возможностью синхронизации, расширять сообщество ученых, разрабатывающих вычислительные алгоритмы, улучшить графический интерфейс, добавить новые опции лингвистического анализа материала.

СПИСОК ЛИТЕРАТУРЫ

1. Старостин и Бронников. Starling project. The Tower of Babel. Retrieved 05 September 2018. http://starling.rinet.ru/intrab.php?lan=en .
2. *Cavar D., Cavar M., Moe L.* The Lexicon Enhancement via the GOLD Ontology (LEGO) project. Retrieved 05 September 2018. http://lego.linguistlist.org
3. *Beermann D., Mihaylov P.* TypeCraft collaborative databasing and resource sharing for linguists. 2014. Language resources and evaluation. V 48. № 2. P. 203–225.
4. Fin-Clarin. Kielipankki (2015) – The Language Bank of Finland. Retrieved 05 September 2018. www.kielipankki.fi
5. *Druskat S., Gast V., Krause T., et al.* corpus-tools.org: An Interoperable Generic Software Tool Set for Multilayer Linguistic Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016. http://www.lrec-conf.org/proceedings/lrec2016/summaries/918.html
6. *Xiaoming H.* The application of queuing theory in harbour programming [J]. Journal of Qingdao university engineering & Technology edition. 1996. № 3.
7. Coubase project. Demystifying conflict resolution in Couchbase mobile. Retrieved 05 September 2018. https://blog.couchbase.com/conflict-resolution-couchbase-mobile/
8. *Boersma P., Weenink D.* Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 12 April 2018. http://www.praat.org/.
9. Date 2008 – Date C. The relation database dictionary. Apress, 2008.
10. *Normanskaya Y.V.* New field and archive data on the Mansi dialects and their meaning for the Proto-Mansi reconstruction of the first syllable vowel system. Ural-Altaic Studies. 2015. № 4. P. 40–59.
11. *Munkácsi B., Kálmán B.* Wogulisches Wörterbuch / Gesammelt von Munkácsi B. Geordnet, bearb. und hrsg. von Kálmán B. Budapest. 1986.

THE SOFTWARE SYSTEM LINGVODOC AND THE POSSIBILITIES IT OFFERS FOR DOCUMENTATION AND ANALYSIS OF OB-UGRIC LANGUAGES

Y. V. Normanskaja[a,b], O. D. Borisenko[a], I. B. Beloborodov[a], and Academician of RAS A. I. Avetisyan[a]

[a] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russian Federation*
[b] *Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russian Federation*

The LingvoDoc system (http://lingvodoc.ispras.ru) provides a service for collaborative language documentation and computations on the collected data. This software system provides GraphQL HTTP API for all the system components and allows its users to build their own extensions for data analysis or even to integrate it with their own software. Thanks to a special database and application design pattern, it is possible to construct offline applications integrated with the LingvoDoc system: these applications would need to have an internet connection only once to synchronize basic data types and for authentification purposes. The system itself allows users to construct multilayer dictionaries, attach them to the geographical map, fill documents with metadata, share access to dictionaries with other users or with everyone. The LingvoDoc system provides fine-grained access control lists for sharing, which allows to separate users into groups of dictionary editors, proofreaders and read-only users. The system also provides some computational algorithms on the stored data: phonology computations, automatic and guided deduplication inside the dictionaries etc. The system allows users to choose the dictionary structure. The system supports the following data types: text, images, sounds (wav, mp3 and flac), markups (ELAN and Praat formats), directed and undirected links between stored entities. A user can choose the most suitable format for their dictionary. Also, the system provides ELAN corpora storage, viewer and processing. In LingvoDoc there are 13 programs made for authors of the dictionary (only 4 of them are available for all users of the system). These programs analyze language data from phonetical, morphological and etymological point of view. This analysis previously was performed manually by linguists. Our programs allows do it tens and sometimes hundreds times faster. This paper presents the documentation and an analysis of Ob-Ugric languages using the LingvoDoc system.

Keywords: software system, documentation and analysis of languages, data analysis.