Сервис по размещению коллекции словарей татарского языка

И.Л. Александрова[1], М.Ф. Насрутдинов[1]

[1] Казанский (Приволжский) федеральный университет

Аннотация. В работе представлен опыт по созданию программного обеспечения для организации и размещения в открытом доступе словарей татарского языка. Проект стартовал в 2018 году под эгидой Института языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан. Целью проекта является создание и развитие многофункционального электронного свода словарей татарского языка, призванного расширить возможности информационного сопровождения применения, изучения и исследования татарского языка. Источниками фонда являются языковые справочники (представленные до этого только в бумажном виде), созданные в предшествующие годы в Институте языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан, предоставленные республиканскими издательствами, а также авторами словарей. В фонде размещаются словари различного типа: толковые, двуязычные и многоязычные, общие и частные, аспектные. На данный момент коллекция содержит 49 словарей. В статье обсуждаются проблемы, которые возникают при организации поисковой системы и работе с несколькими алфавитами (кириллица, латиница, арабский алфавит), способы разметки в словарях для поддержки сквозного поиска по всем источникам.

Ключевые слова: коллекция словарей татарского языка, оцифровка специализированных словарей.

Service for hosting a collection of Tatar language dictionaries

I.L. Alexandrova[1], M.F. Nasrutdinov[1]

[1] Kazan (Volga Region) Federal University

Abstract. The experience of creating software for organizing and sharing Tatar language dictionaries is described. The project started in 2018 under the leadership of the Ibragimov Institute of Language, Literature and Art of the Academy of Sciences of the Republic of Tatarstan. The aim of the project is to create and develop a multifunctional electronic collection of Tatar language

dictionaries. The sources of the fund are language guides (previously presented only in paper form), created in previous years at the Institute of Language, Literature and Art. The fund contains dictionaries of various types: explanatory, bilingual and multilingual, general and particular, aspect. At the moment the collection contains 49 dictionaries. The article discusses the problems of organizing a search system and working with several alphabets (Cyrillic, Latin, Arabic), methods of markup in dictionaries to support end-to-end search across all sources.

Keywords: collection of dictionaries of the Tatar language, digitization of specialized dictionaries.

1. Введение

Проект по созданию объединенного Электронного фонда словарей (ЭФС) татарского языка стартовал в 2018 году на базе Института языка, литературы и искусства им. Г. Ибрагимова Академии наук Республики Татарстан (ИЯЛИ). Основу коллекции составляют словари, созданные в ИЯЛИ в различные годы и представленные до этого в основном в печатном виде. Ссылка на коллекцию в Навигаторе информационных ресурсов по общественным наукам (НИРОН) http://niron.inion.ru/linguistics/130. Словари размещены на платформе сайта Академии наук Республики Татарстан по адресу: http://suzlek.antat.ru

В настоящее время размещено свыше 70 источников, некоторые из них представляют многотомные издания. Самым старым по году издания является словарь 1965 года – Арабско-татарско-русский словарь заимствований, самым новым – Толковый словарь татарского языка в 6 томах (2015-2021 г.). Фонд словарей размещен в открытом доступе.

Кратко о содержании проекта ЭФС с точки зрения филологов можно прочитать в статье [1] одного из первых руководителей проекта профессора Казанского университета, член-корреспондента Академии наук РТ К.Р. Галиуллина (1951-2021). ЭФС в 2019 году содержал материалы татарского языка, с различными входными языками: татарским, русским и английским [1]. Сейчас к ним добавились турецкий и арабский язык.

На рисунках 1-3 представлен вид начальной страницы поиска и пример запроса по базе словарей.

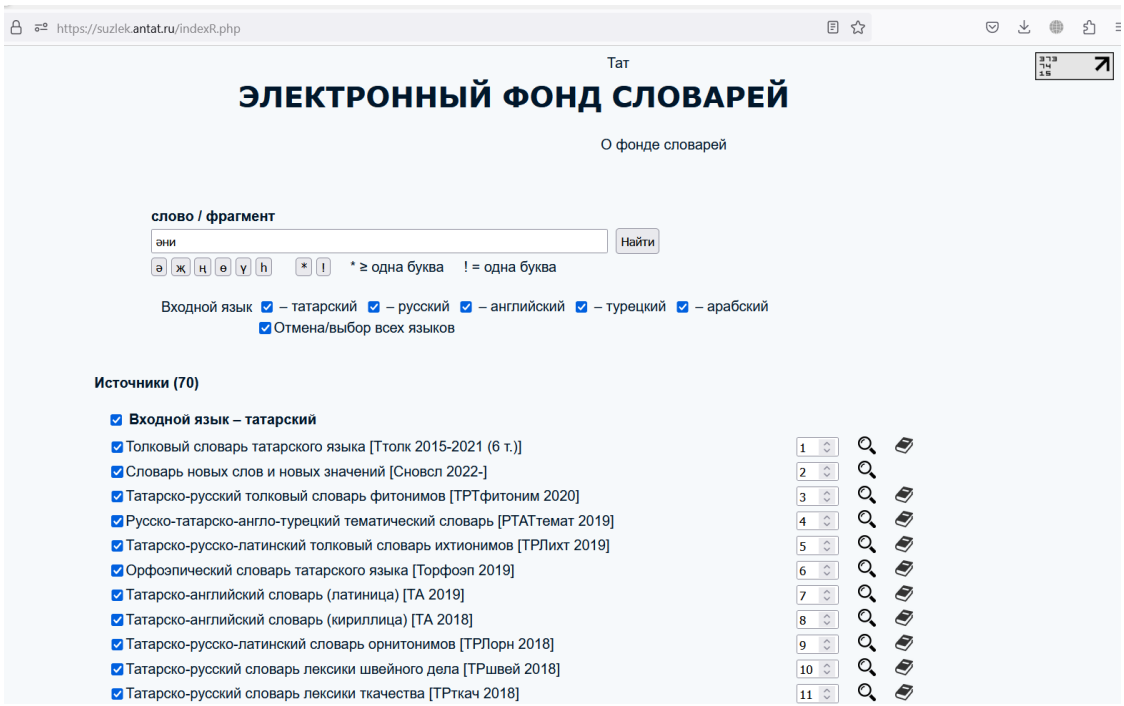

Рис. 1. Начальная страница поиска

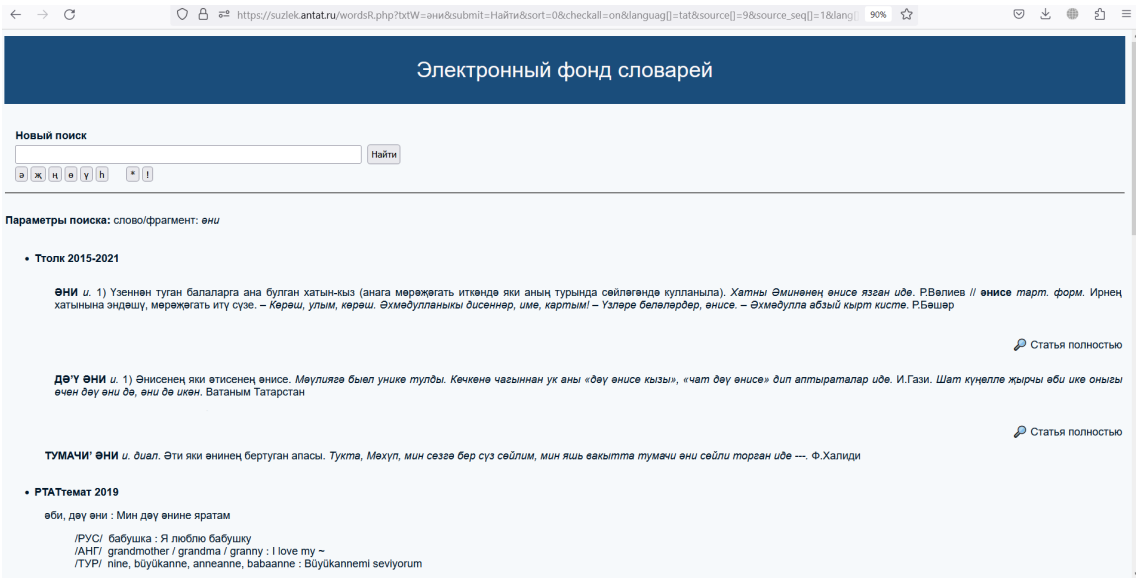

Рис. 2. Список словарных статей, соответствующих запросу

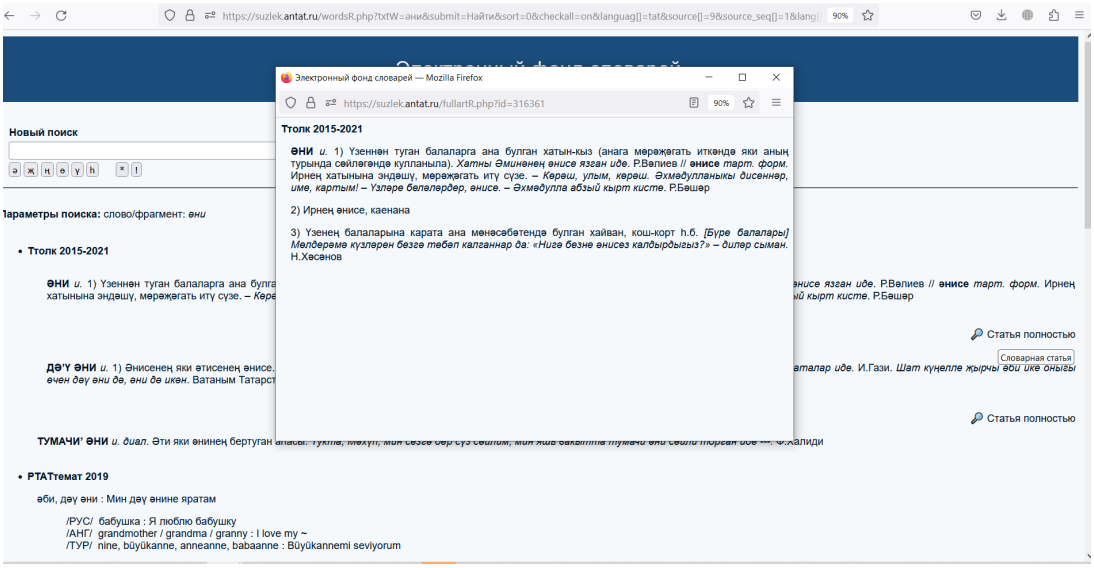

Рис. 3. Словарная статья из выбранного словаря

Создание электронных словарей имеет достаточно давнюю историю в Казанском университете. Уже в 1995 году вышел в свет сборник [2] по компьютерной лингвографии. "Лингвография – междисциплинарная область языкознания, теория и практика создания языковых справочников, словарей, подразделами лингвографии являются лексикография, фразеография, морфемография, паремиография и др. " [3]. Приведем еще цитату из [3]: "Компьютерные языковые справочники функционируют не только как аккумулятор информации, не только как источник материалов для последующих изысканий, но также как инструмент обработки данных, инструмент научного исследования. Это одна из отличительных особенностей современной лингвографии." Подходы к построению словарей и сервисов по их размещению были оформлены в виде патентов [4-5], см. также [6].

По статистике в день отмечается около 300 просмотров страниц проекта. Следует отметить, что словари коллекции в большей степени предназначены специалистам, для которых татарский язык является областью профессиональной деятельности. "Обычный" пользователь, скорее всего, воспользуется существующим переводчиком от Яндекса или Google. В то же время словарные статьи по конкретному слову будут в этом случае сильно урезаны или отсутствовать. Отметим также, что в Институте прикладной семиотики Академии наук Татарстана осуществляется исследование в области автоматического перевода, см., например, https://translate.tatar.

Тема создания электронных словарей и их отличие от простой электронной версии бумажного словаря обсуждалась многими авторами [7-9].

Наша статья посвящена только технической составляющей проекта. Обсуждаются проблемы, которые возникают при организации поисковой

системы, работой с несколькими алфавитами (кириллица, латиница, арабский алфавит), способы разметки в словарях для поддержки сквозного поиска по всем источникам.

2. Процесс создания базы данных словарей

Процесс наполнения базы данных словарей состоит из нескольких этапов.

Сначала проводится сканирование и распознавание текста отдельного словаря. Затем специалисты ИЯЛИ выверяют текст и в отсканированные документы вносят вспомогательные символы разметки.

Далее размеченный текст обрабатывается специально разработанной нами программой-парсером. Программа-парсер разбивает текст на словарные статьи, в которых выделяются заголовочные единицы и другая служебная информация: омонимы, синонимы, происхождение, функционально-стилистические признаки, фразеологизмы и др. Вся полученная информация записывается в базу данных.

2.1. Предварительная обработка данных

Работа программы-парсера основана на вспомогательных символах разметки, которые были придуманы нами совместно с проф. Галиуллиным К.Р. Эти символы универсальны и подходят для любого словаря с любым входным языком, размещаемого в ЭФС.

Для многоязычных словарей словарные статьи на разных языках часто размещены в разных столбцах таблицы. В случае, когда текст на другом языке встречается внутри словарной статьи без разбиения на колонки, вводится пометка \язык\ перед текстом на другом языке.

Символы разметки в отсканированный текст вносят специалисты ИЯЛИ. Какие-то символы можно вносить с помощью инструмента автозамены, например, зону для обозначения омонимов, или символ начала отсылки.

Далее текст из документа в формате Word переводится в текстовый файл с сохранением разметки: вносятся теги разметки абзацев, выделения курсивом и полужирным шрифтом так, как это принято в html.

2.2. Процесс работы программы-парсера

На вход программе-парсеру подается текстовый документ в кодировке юникод, содержащий служебную разметку, а также указывается тип словаря.

Сейчас все словари разбиты на три типа. Первый тип – словарь с простой структурой, в котором поиск осуществляется по заголовочной единице или по зонам со служебной информацией, и в котором встречаются отсылки к другим статьям этого словаря. Второй тип –

словари, в которых кроме возможностей словаря первого типа добавлены вложенные словарные статьи. По вложенным словарным статьям также возможен поиск. Третий тип – словари, в которых словарные статьи не распознаны в виде текста, а хранятся в формате pdf. Для словарей третьего типа исходный файл отличается от файла для словарей первых двух типов. В нем указаны заголовочные единицы и номер страницы в pdf файле.

Программа-парсер производит обработку файла построчно. Для нее важно, чтобы в одной строке файла располагалась только одна словарная статья. В случае некорректной разметки строки в исходном файле, программа останавливает свою работу и сообщает пользователю о проблеме в строке. Прерывание работы программы было сделано специально, т.к. для конечных пользователей важно сохранить порядок отображения статей. После исправления проблемной строки программа продолжает загружать оставшиеся строки файла. После прохода по файлу формируется база данных.

Далее происходит процесс обработки отсылок. На этом этапе в базе данных ищутся строки – словарные статьи обрабатываемого словаря, которые содержат отсылки. Выделяется слово-отсылка, и слово-отсылка ищется в базе данных среди заголовочных единиц этого словаря. В случае успешного поиска пометки отсылки в словарной статье заменяются гиперссылкой на статью с найденной заголовочной единицей. В случае если заголовочная единица, на которую ссылается статья, найдена не была, информация записывается в файл ошибок. Далее специалист по этому файлу с ошибками может дополнительно внести правки в словарь. Обычно ошибки, связанные с отсутствием заголовочных единиц, возникают из-за неправильно распознанных при сканировании символов.

3. Особенности структуры базы данных

На рисунке 4 представлена обобщенная схема базы данных словарей.

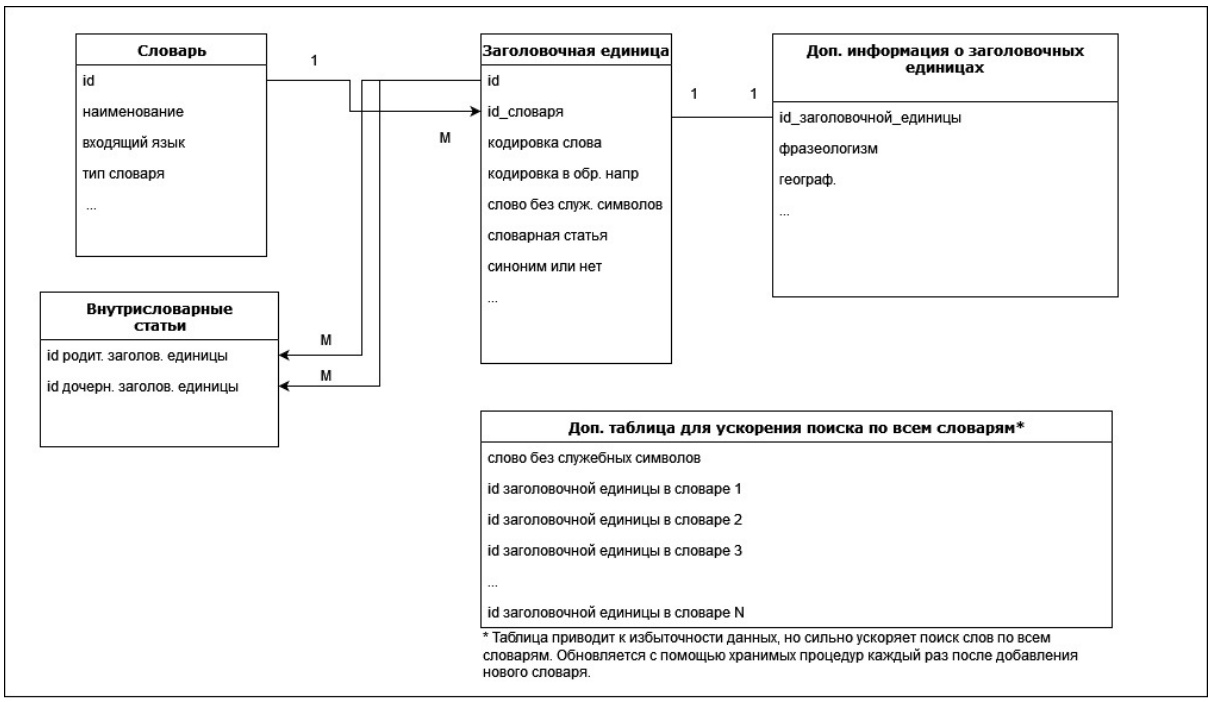

Рис. 4. Обобщенная схема базы данных словарей

Одной из основных проблем при обработке запросов стало долгое время поиска по базам всех словарей. Для решения была создана дополнительная таблица для ускорения поиска.

Заключение

Проект по наполнению баз словарей продолжается. Дальнейшее направление работы может включать как "простое" добавление словарей, так и добавление новых сервисов, связанных с интеллектуальной обработкой текстов. Например, одной из проблем поиска является разные варианты написания слов в зависимости от года издания словаря.

Литература

1. Галиуллин К.Р., Каримуллина Г.Н., Каримуллина Р.Н. Словари татарского языка: электронный металингвографический фонд // И.А. Бодуэн де Куртенэ и мировая лингвистика: междунар. конф.: VII Бодуэновские чтения (Казан. федер. ун-т, 28–31 окт. 2019 г.): тр. и матер.: в 2 т. / под общ. ред. К.Р.Галиуллина, Е.А.Горобец, Э.А.Исламовой. – Казань: Изд-во Казан. ун-та, 2019.– Т.2.– 280 с
2. Компьютерная лингвография / Науч. ред. Н.К. Замов, К.Р. Галиуллин. – Казань: Изд-во Казан. ун-та, 1995. – 119 с.
3. Галиуллин К.Р. Интернет-лингвография: русские текстоописывающие словари [Текст] // Проблемы истории, филологии, культуры. - Вып. 2(24).- Магнитогорск; Новосибирск: Аналит, 2009.- С.635-639.

4. Свидетельство № 2015620420 от 02.03.2015 о государственной регистрации базы данных "Электронный конкорданс к поэтическим произведениям Габдуллы Тукая" (Авторы - К.Р.Галиуллин, Р.Н.Каримуллина, Н.А.Обносова, И.Л.Александрова)
5. Свидетельство № 2013620723 от 21.06.2013 о государственной регистрации базы данных "Русский язык в словарных описаниях середины XIX века" (Авторы - К.Р.Галиуллин, Д.А.Мартьянов, М.Р.Загидуллин, И.Л.Александрова)
6. Галиуллин К.Р. Русский язык в словарных описаниях середины XIX века: справочный интернет-комплекс [Текст] / К.Р.Галиуллин, Д.А. Мартьянов, И.Л.Александрова, М.Р.Загидуллин // Международный журнал экспериментального образования. – М., 2013.– № 5.– С. 92-94.
7. Климова О. В. Создание электронных словарей в рамках разработки концепции электронных изданий [Электронный ресурс] / О. В. Климова, Е. А. Березовская // Новые образовательные технологии в вузе: материалы XI международной научно-методической конференции. — Екатеринбург, 2014. — Режим доступа: http://hdl.handle.net/10995/24817
8. Зайковская И. А. Особенности реализации текстовых категорий в печатном и электронном словаре // Известия Российского гос. пед. ун-та, 2011 № 130, СПб., С. 150–155
9. Загорулько, М. Ю. Программный инструментарий разработки лингвистических ресурсов / М. Ю. Загорулько, Е. А. Сидорова // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2013): материалы III Междунар. научн.-техн. конф. (Минск, 21-23 февраля 2013г.) / редкол.: В. В. Голенков (отв. ред.) [и др.]. – Минск: БГУИР, 2013. – С. 159 – 164. https://libeldoc.bsuir.by/handle/123456789/4278

References

1. Galiullin K.R., Karimullina G.N., Karimullina R.N. Dictionaries of the Tatar language: electronic metalinguographic fund // I.A. Baudouin de Courtenay and World Linguistics: Intern. Conf.: VII Baudouin Readings (Kazan Federal University, October 28–31, 2019): tr. and mater.: in 2 volumes / under the total. ed. K.R.Galiullina, E.A.Gorobets, E.A.Islamova. - Kazan: Kazan Federal University, 2019. – V.2.– 280 p.
2. Computational linguography / Nauch. ed. N.K. Zamov, K.R. Galiullin. - Kazan: Kazan Publishing House. un-ta, 1995. - 119 p.
3. Galiullin K.R. The internet linguography: Russian text description // Problems of history, philology, culture. - Vol. 2(24). - Magnitogorsk; Novosibirsk: Analit, 2009.- P.635-639.

4. Certificate No. 2015620420 dated March 2, 2015 on state registration of the database "Electronic concordance to the poetic works of Gabdulla Tukay" (Authors - K.R. Galiullin, R.N. Karimullina, N.A. Obnosova, I.L. Alexandrova)
5. Certificate No. 2013620723 dated June 21, 2013 on state registration of the database "Russian language in dictionary descriptions of the middle of the 19th century" (Authors - K.R. Galiullin, D.A. Martyanov, M.R. Zagidullin, I.L. Alexandrova)
6. Galiullin K.R. Russian language in dictionary descriptions of the middle of the 19th century: Internet reference complex [Text] / K.R. Galiullin, D.A. Martyanov, I.L. Alexandrova, M.R. Zagidullin // International Journal of Experimental Education. - M., 2013. - No. 5. - P. 92-94.
7. Klimova O. V. Creation of electronic dictionaries as part of the development of the concept of electronic publications [Electronic resource] / O. V. Klimova, E. A. Berezovskaya // New educational technologies at the university: materials of the XI international scientific and methodological conference. - Yekaterinburg, 2014. - Access mode: http://hdl.handle.net/10995/24817
8. Zaikovskaya I. A. Features of the implementation of text categories in the printed and electronic dictionary. Izvestiya Rossiyskogo Gos. ped. University, 2011 No. 130, St. Petersburg, pp. 150–155
9. Zagorulko M.Yu. Tools for language resources software development/ Zagorulko M.Yu., Sidorova E.ASidorova // Open semantic technologies for designing intelligent systems = Open Semantic Technologies for Intelligent Systems (OSTIS-2013): materials of the III Intern. scientific-technical conf. (Minsk, February 21-23, 2013) / editorial board. : V. V. Golenkov (editor-in-chief) [and others]. - Minsk: BSUIR, 2013. - S. 159 - 164. https://libeldoc.bsuir.by/handle/123456789/4278