Желтов Валериан Павлович
канд. техн. наук, профессор
ФГБОУ ВО «Чувашский государственный
университет им. И.Н. Ульянова»
Желтов Павел Валерианович
канд. техн. наук, ведущий инженер
АУ Чувашской Республики «Научно-исследовательский
институт экологии и природопользования»
Министерства природных ресурсов и экологии
Чувашской Республики
г. Чебоксары, Чувашская Республика

АНАЛИЗ НАЦИОНАЛЬНЫХ КОРПУСОВ И НОВЫЕ ВОЗМОЖНОСТИ В ИЗУЧЕНИИ РОДНЫХ ЯЗЫКОВ

Аннотация*: в статье рассмотрены национальные корпуса с поисковыми системами. Одно из главных направлений корпусного языкознания – формирование национальных языковых корпусов в их естественном виде. Проведен краткий анализ национальных корпусов и показаны новые возможности в изучении родных языков.*

Ключевые слова*: национальный корпус, русский язык, татарский корпус, чувашский язык.*

Современные компьютеры в состоянии обрабатывать огромный лексический текстовый материал, благодаря чему стало возможным объяснить выдвигаемые теоретические модели лингвистических явлений и выявить новые, ранее не изученные лингвистические закономерности. Большинство классических гипотез языкознания проверяются новыми способами, и поэтому конечная цель их достигается в разы доказательнее и легче [1; 2]. Благодаря этому отмечается быстрый качественный переход в лексикологии и лексикографии: значительно облегчается работа по составлению словарей и тезаурусов, во время которой необходимо учитывать, как практические, так и теоретические составляющие.

Content is licensed under the Creative Commons Attribution 4.0 license (CC-BY 4.0)

Лингвистические исследования, касающиеся изучения письменного языка и устной речи, все больше и больше ускоряются. Резко расширяется круг возможностей по наблюдению и последующему изучению речи.

Любой язык, в том числе его лексика, постоянно развивается. Появляются новые значения слов. В некоторых случаях они даже заменяют старые значения. Имеет место и появление новых слов, число которых увеличивается с каждым годом, включая заимствованные слова из других языков. Новые слова, в основном, представляют собой профессиональные и общеиспользуемые термины из области науки и техники, политики и экономики, а также социальных взаимоотношений и повседневного быта. В результате этого, количество терминов практически в каждом языке насчитывает миллионы, включая разнообразные сочетания слов. Из-за нарастающей информатизации общества (так называемого информационного взрыва) и из-за того, что каналы межъязыковых коммуникаций расширились, увеличение лексических значений стало играть особую роль.

Современные методы исследования по корпусной лингвистике предполагают появление нового инструментария, который не был известен самым первым лингвистам, и который позволяет проверить гипотетические теории, как способом обратной связи, так и современными формализованными и количественными методами.

Национальный корпус должен характеризовать чувашский язык республики и диаспоры в целом, так как должен содержать практически исчерпывающее собрание текстов на чувашском литературном языке, в том числе и в различных переводных вариациях. Корпус должен включать в себя не только классические художественные произведения, но и статьи периодических изданий, специальные технические тексты.

Национальные корпуса предоставляют возможность следить за поведением единиц языка в естественных, реально существующих контекстах. Речь идет о словах, словоформах, грамматических, синтаксических языковых конструкциях.

Содержимое доступно по лицензии Creative Commons Attribution 4.0 license (CC-BY 4.0)

Одно из главных направлений корпусного языкознания – формирование национальных языковых корпусов в их естественном виде. Такие корпуса созданы для большинства языков. Образцом заслуженно можно считать BNC – это корпус британского языка. Существует такой корпус (Ruscorpora.ru) и для русского языка. Благодаря нему ученые, учителя, учащиеся, иностранцы имеют возможность получить профессиональные ответы на свои вопросы, которые связаны с русским языком. По объему русский языковой корпус насчитывает более шестисот миллионов слов [3]. Для национального корпуса русского языка характерно следующее.

Тексты, которые представляют русский язык, можно разделить на две большие группы:

– тексты современности (XX–XXI век);

– тексты середины XVIII–XX столетия.

Поиск по данным группам (массивам) по умолчанию проводится одновременно. Если необходимо задать другие параметры, к примеру, хронологический диапазон, то это делается вручную.

Все тексты из основного корпуса проходят два вида разметки: морфологическую и мета-разметку.

Первая производится посредством специализированных аналитических программных ресурсов. В главной части языкового корпуса есть возможность снять вручную омонимию и подкорректировать выдачу ресурса. Эта часть корпуса может стать плацдармом для морфологических исследований, тестирования разных поисковых программ, изучения современной русской морфологии – всего, что требует повышенной поисковой точности. Те примеры, которые взяты из этой части (подкорпуса) ресурса помечены тэгом «[омонимия снята]». При помощи русского грамматического словаря тексты с пометкой «[омонимия снята]» снабжены автоматической акцентуацией.

Национальный корпус русского языка (НКРЯ) находится в постоянном развитии. На данный момент подкорпусы НКРЯ отличаются тем, что:

Content is licensed under the Creative Commons Attribution 4.0 license (CC-BY 4.0)

– каждый корпус глубоко аннотирован, каждое предложение имеет дерево зависимостей, то есть выстроенную морфологическую и синтаксическую структуру;

– можно найти практически любые переводы с английского на русский и наоборот;

– присутствует материал диалектов с их уникальной спецификой;

– поэтические произведения можно искать не только по лексическим или грамматическим признакам, но и по типу рифмовки и т. д.;

– имеет обучающий корпус для школьников;

– содержит устную речь с магнитофонных пленок и фильмов 1930–2000 годов.

В структуру НКРЯ входит разметка по мета-тэгам, морфологии, синтаксису, семантике.

Сегодня НКРЯ состоит из следующих корпусов:

– главный (проза, драматургия 18–21 столетия);

– с глубоким деревом зависимостей (синтаксический);

– современных СМИ (с 1990 по 2000 год);

– с переводом русских слов и словосочетаний на другие языки мира (к примеру, французский, испанский, итальянский, польский, украинский, белорусский и т. д.);

– диалектов Российской Федерации;

– поэтических произведений;

– обучающего корпуса для школьников и студентов;

– устной речи (киноленты и магнитофонные записи);

– истории русского ударения;

– мультимедиа (видео- и аудиопленки 1930–2000 гг.).

В качестве еще одного хорошего примера из корпусов по тюркским языкам можно привести татарский корпус (http://web-corpora.net/TatarCorpus). В корпусе собраны различные жанры. У каждого документа свое мета-описание и морфо-

Содержимое доступно по лицензии Creative Commons Attribution 4.0 license (CC-BY 4.0)

логическая разметка, которая реализуется автоматически посредством программного инструмента PC-KIMMO (двухуровневый морфологический анализатор языка) [4].

Под этот корпус адаптирован поисковик Восточноармянского языкового корпуса, с помощью которого можно искать необходимую информацию по лексике, словоформам и определенным грамматическим особенностям.

Стоит обратить внимание на то, что регулярность морфологии в татарском языке некоторым образом нарушена из-за большого количества заимствований и несовершенства современной орфографии татарского языка, что затрудняет процесс автоматической обработки.

Татарский языковой корпус размещен на платформе EANC (East Armenian National Corpora), которая была разработана для Восточноармянского языкового корпуса. Платформа включает в себя поисковик, веб-интерфейс, индексатор, являющий собой модуль, который преобразовывает входящие данные (тексты) в базы данных.

Эта платформа, изначально предназначенная для армянского языка, по сути, универсальна и может быть использована для любого другого языка.

Поиск, выбор и сохранение материала на татарском языке.

Основа результатов – это сбор текстов, доведение их до единого формата, создание таблиц исходных данных с мета-данными: автор текста, название, дата издания, объем текста, его тип и т. д.

Морфологическая разметка национального корпуса татарского языка, первоочередной целью имеет представить все естественные грамматические формы, которые не всегда отражаются в исследованиях татарской грамматики или имеют неоднозначные трактовки.

Поиск по корпусу осуществляется поисковой системой Яndex.Server 3.1 Professional.

Content is licensed under the Creative Commons Attribution 4.0 license (CC-BY 4.0)

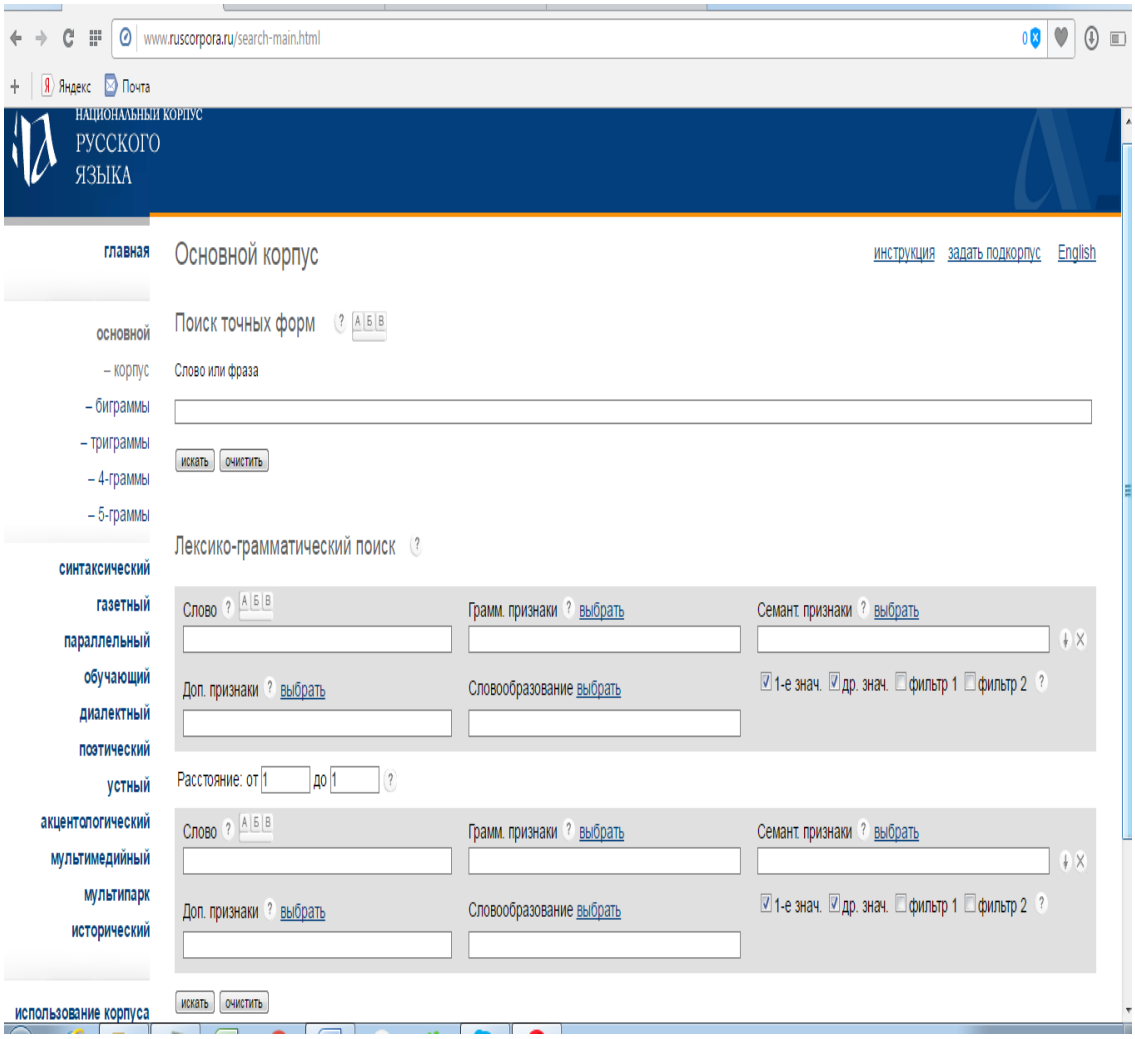

Рис. 1

Поиск лексических единиц может вестись как во всем корпусе текстов, так и в отдельных его разделах (подкорпусах). Например, можно ограничить число текстов, выбрав тип или жанр текста, место и время описываемых событий [5].

При переходе на страницу «Поиск» в национальном корпусе русского языка откроется «Основной корпус», слева будет список других корпусов (рис. 1)/

Поисковая система интернет – портала национального корпуса татарского языка http://web-corpora.net/TatarCorpus/search/?interface_language=ru позволяет реализовать поиск по:

– лемме (лексеме);

– словоформе;

– заданному набору морфологических параметров.

Содержимое доступно по лицензии Creative Commons Attribution 4.0 license (CC-BY 4.0)

Поисковая система татарского корпуса (рис. 2) поддерживает поиск минус-слов, поиск по части слова, поиск с использованием логических формул; таким образом, пользователь может задавать сложные запросы, требуемые спецификой своего научного исследования.

Основные запросы, которые может произвести пользователь в татарском корпусе с помощью разработанного интерфейса – это поиск по точной форме, поиск по лемме (начальной форме) и поиск по набору грамматических параметров.

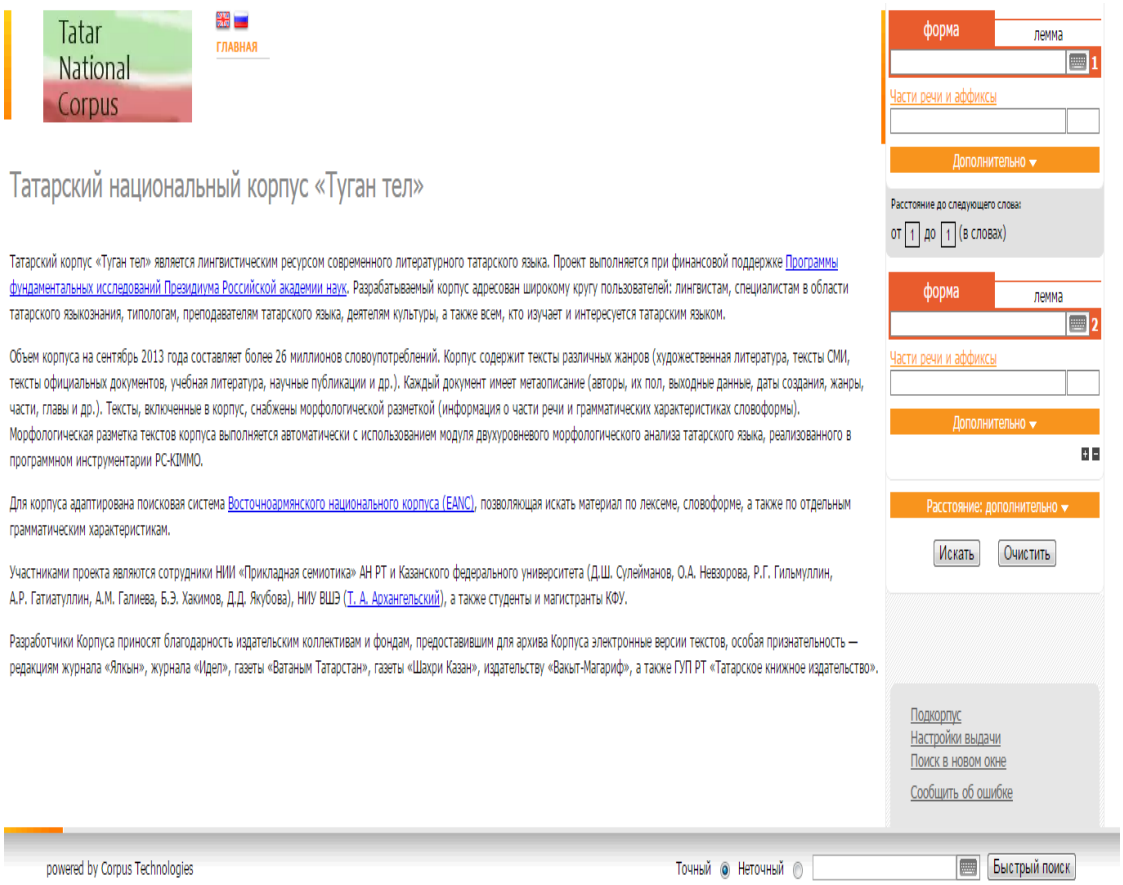

Рис. 2

Результатом обработки запроса являются все предложения, которые содержат словоформы, соответствующие заданным критериям. Все эти предложения отображаются на странице.

Результатом обработки запроса являются все предложения, которые содержат словоформы, соответствующие заданным критериям. Все эти предложения отображаются на странице выдачи (количество предложений на одной странице

Content is licensed under the Creative Commons Attribution 4.0 license (CC-BY 4.0)

задаётся настройками, но не превышает 50). Для поиска по начальной форме или набору грамматических признаков необходимо, чтобы искомые слова имели грамматическую разметку. Поиск по точной форме работает и на тех словоформах, которым не было присвоено ни одного грамматического разбора. Грамматические пометы в татарском корпусе, в отличие от существующих корпусов на платформе EANC, кодируют не все грамматические категории, а только те из них, которые явно выражены аффиксами (при этом немаркированные элементы парадигм не получают специальных помет) [6].

Таким образом, при поиске по грамматике в татарском корпусе фактически осуществляется поиск по морфемам. Одна из связанных с этими проблемами касается тех случаев, когда один и тот же аффикс присоединяется к словоформе.

Поисковая система интернет – портала национального корпуса чувашского языка *http://ru.corpus.chv.su/* позволяет выделять корни слов – данный функционал реализован с помощью словаря Hunspell. Структура корпуса: тексты – предложения – слова. Тексты разбиты по их типам (публицистика, научные статьи, проза, поэзия, законы и т. д.), а также по тематикам (культура, вооруженные силы, сельское хозяйство, техника и т. д.). Также у текстов указаны авторы и их источники.

Выводы

Развитие современных информационных технологий приводит к тому, что появляются новые возможности в изучении языка в противовес традиционным – поиску нужной информации в словарях, художественных сочинениях, текстах классиков и других сохранившихся письменных источниках. Корпусная лингвистика в качестве основного практического метода исследования рассматривает использование Национального корпуса, то есть практически безразмерного сборника живого лексического материала, который был ранее создан, оцифрован, размечен, внесен в компьютер и непрерывно пополняется. Благодаря этому можно исследовать как речь, так и сам язык под новым ракурсом, применяя корпус как словарь и тезаурус.

Содержимое доступно по лицензии Creative Commons Attribution 4.0 license (CC-BY 4.0)

Практическая значимость чувашского национального корпуса заключается в том, что он позволит получать новые, более точные данные о распределениях языковых единиц чувашского языка. Данные корпуса будут способствовать разработке новых словарей чувашского языка, получению необходимых данных, а также развитию исследований в области русско-чувашского машинного перевода.

Национальный корпус чувашского языка должен стать не только самым первым (так как ранее исследователи не ставили перед собой настолько глобальные и всеобъемлющие задачи), но и самым крупным сборником национальных текстов [7; 8].

Список литературы

1. Копотев М.В. Современная корпусная русистика. Инструментарий русистики: корпусные подходы – Хельсинки: Yliopistopaino, 2008. – 36 с.

2. Корпусы текстов по русскому языку. Национальный корпус русского языка. Текстовая структура, поисковые возможности [Электронный ресурс]. – Режим доступа: http://mylektsii.ru/1–11436

3. Национальный корпус татарского языка «Туган Тел» [Электронный ресурс]. – Режим доступа: http://www.web-corpora.net/TatarCorpus/search/?interface_language=ru

4. Apertium Documentation [Электронный ресурс]. – Режим доступа: http://wiki.apertium.org/wiki/Documentation

5. Захаров В.П. Корпусная лингвистика: учебно-метод. пособие – СПб., 2005.

6. Селихов К.М. Татарский корпус текстов // Создание и развитие корпусных ресурсов по языкам народов России. Казань, 2012. – С. 4–14.

7. Желтов В.П. Средства разработки интернет-портала национального корпуса чувашского языка / В.П. Желтов, П.В. Желтов // Программные системы и вычислительные методы. – 2019. – №1. – С. 42–50 [Электронный ресурс]. – Режим доступа: https://nbpublish.com/library_read_article.php?id=28131

Content is licensed under the Creative Commons Attribution 4.0 license (CC-BY 4.0)

8. Желтов П.В. Национальный корпус чувашского языка: концепция и архитектура. – Чебоксары: Изд-во Чувашского ун-та, 2017. – 159 с.

Содержимое доступно по лицензии Creative Commons Attribution 4.0 license (CC-BY 4.0)