Инновации и технологии

УДК 811.511.131’35

М. П. Безенова, Г. Л. Григорьев

РОЛЬ ПРОГРАММЫ ПРОВЕРКИ УДМУРТСКОЙ ОРФОГРАФИИ В ПОПОЛНЕНИИ НАЦИОНАЛЬНОГО КОРПУСА УДМУРТСКОГО ЯЗЫКА

Корпусная лингвистика – на данный момент один из самых популярных разделов языкознания. Большинство крупных языков мира сегодня уже имеет свои электронные корпуса объемом в десятки и сотни миллионов словоупотреблений. В последнее время созданию корпусов текстов на языках народов России также уделяется особое внимание, поскольку, с одной стороны, корпусные исследования позволяют взглянуть на устройство языка с совершенно иного ракурса; с другой стороны, корпус – это своеобразная форма сохранения языковых данных. В статье описывается Национальный корпус удмуртского языка (Удмурт йӧскалык кылшыкыс), который разрабатывается с конца 2019 года сотрудниками отдела филологических исследований Удмуртского института истории, языка и литературы УдмФИЦ УрО РАН. Подробно говорится о возможностях создаваемой информационно-справочной системы на данный момент, а также о перспективах использования корпуса текстов при проведении исследований, подготовке словарей и создании различных программ по удмуртскому языку.
В статье идет речь также о программе проверки удмуртской орфографии на основе Hunspell, разработанной Григорием Григорьевым, которая играет немаловажную роль в пополнении Национального корпуса удмуртского языка. Перед загрузкой на сайт новых текстов все они подвергаются обязательной проверке на наличие орфографических ошибок, которые могли остаться при их вычитке. Данное расширение для текстовых редакторов, благодаря словарной базе, связанной с файлом аффиксов, в котором заложены по возможности все морфологические варианты лексем основного словаря, выявляет орфографические ошибки, позволяя загружать на сайт Национального корпуса удмуртского языка максимально выверенные тексты.

Ключевые слова: корпусная лингвистика, корпус текстов, удмуртский язык, национальный корпус, возможности и перспективы корпуса, проверка орфографии, Hunspell.

DOI: 10.35634/2224-9443-2020-14-3-549-556

Корпусная лингвистика, появившись относительно недавно (лишь во второй половине XX века), в настоящее время стала одним из ведущих направлений в современном языкознании, включая в себя и создание корпусов, и корпусные исследования.

Языковой корпус представляет собой информационно-справочную систему, основанную на собрании текстов на некотором языке в электронной форме. Важнейшая составляющая любого корпуса – это аннотация [McEnery, Hardie 2012, 13], т. е. «лингвистический разбор всех языковых единиц на выбранном языковом уровне» [Копотев 2014].

Большинство крупных языков мира сегодня уже имеет свои корпуса, в частности: Британский национальный корпус (British National Corpus, BNC)[1], Американский национальный корпус (American National Corpus, ANC)[2], Венгерский национальный корпус (Magyar Nemzeti Szövegtár, MNSZ)[3], Чешский национальный корпус (Český národní korpus, ČNK)[4], Цифровой словарь немецкого языка (Digitales Wörterbuch der deutschen Sprache, DWDS)[5] и др. В последние годы корпусным исследованиям уделяется особое внимание и в России. В итоге сейчас создан Национальный корпус русского языка[6], насчитывающий более 300 млн словоупотреблений, развитый действительно на высочайшем мировом уровне. Корпуса текстов в настоящий момент существуют и на отдельных языках народов Рос-

[1] https://www.english-corpora.org/bnc/
[2] https://www.anc.org
[3] http://corpus.nytud.hu/mnsz/index_eng.html
[4] https://ucnk.ff.cuni.cz/cs/
[5] https://www.dwds.de
[6] http://www.ruscorpora.ru/new/

сии: Машинный фонд башкирского языка[7], Письменный корпус татарского языка[8], Двуязычный корпус чувашского языка[9], Корпус вепсского языка[10], Корпус коми языка[11] и некоторые другие.

В конце 2019 г. сотрудники отдела филологических исследований Удмуртского института истории, языка и литературы УдмФИЦ УрО РАН положили начало подготовке Национального корпуса удмуртского языка. Ниже подробно рассмотрим структуру данной информационно-справочной системы, а также познакомимся с программой проверки удмуртской орфографии и возможными сферами ее применения.

Национальный корпус удмуртского языка

Национальный корпус удмуртского языка (Удмурт йӧскалык кылшыкыс)[12] – это информационно-справочная система, основанная на собрании удмуртских текстов в электронной форме[13]. Стоит отметить, что на сегодняшний день лингвистические корпуса удмуртского языка в интернете представлены только разработками Тимофея Архангельского (НИУ ВШЭ)[14]. Однако основной его корпус объемом 9,57 млн токенов[15] на 91,3 % состоит из прессы [Архангельский 2019, 530], представляющей собой значительно упрощенный вариант удмуртского языка в сравнении с классическим удмуртским. На наш взгляд, корпусные исследования целесообразнее проводить, в первую очередь, на основе анализа литературных текстов, поэтому на данный момент в рамках информационно-справочной системы «Национальный корпус удмуртского языка» идет формирование корпуса современного литературного языка; в будущем же предполагается разработка диалектологического и фольклорного корпусов, корпуса параллельных текстов и др.

На сегодня система Национального корпуса удмуртского языка открывает возможности:

– поиска определенных словоформ и словосочетаний по всем загруженным на сайт текстам;

– подбора словоформ по грамматическим показателям (например: «найти все существительные в родительном падеже», «найти все глаголы в прошедшем времени» и др.);

– поиска по авторам;

– поиска по годам публикаций;

– поиска по подкорпусам (проза и поэзия);

– просмотра метаинформации найденного словоупотребления, включая сведения об авторе, названии произведения, источнике и др.

– просмотра контекстов, с возможностью видеть предыдущее и последующее предложения;

– просмотра морфологического разбора слова и его перевода на русский язык[16];

– скачивания результатов поиска в виде электронной таблицы Excel, в которой дан набор предложений с указанием источника.

Сегодня корпуса широко используются в различных научных направлениях и сферах деятельности, в том числе в лексикографии, компьютерной лингвистике, культурологии, диалектологии, лингводидактике и др. [Жевнерович 2018, 26–28]. Национальный корпус удмуртского языка создается лингвистами для решения ряда задач.

---

[7] http://mfbl2.ru

[8] https://www.corpus.tatar

[9] http://ru.corpus.chv.su

[10] http://vepsian.krc.karelia.ru

[11] http://komicorpora.ru

[12] http://udmcorpus.udman.ru/

[13] В систему Национального корпуса удмуртского языка также интегрированы электронные версии удмуртско-русского [УРС] и русско-удмуртского [РУС I, РУС II] словарей, поэтому сайт предназначен не только для профессиональных лингвистов, но и для обычных школьников, а также всех, кто интересуется удмуртским языком.

[14] http://udmurt.web-corpora.net

[15] «*Токен* / текстоформа (англ. *token*) – минимальная единица морфологического анализа, аналог словоформы в лингвистике, чаще всего понимаемая как единица „от пробела до пробела“. Например, в предложении: „Я буду приезжать в ваш город“ – 6 текстоформ, но 5 словоформ („буду приезжать“ – одна словоформа, но две текстоформы)» [Копотев 2014].

[16] В системе Национального корпуса удмуртского языка используется морфологический анализатор Тимофея Архангельского, который свободно распространяется и доступен по адресу https://github.com/timarkh/uniparser-grammar-udm.

В первую очередь, корпус предоставляет богатый материал для **проведения новых исследований** и верификации уже полученных ранее результатов. Подобное собрание текстов в электронной форме дает детальное представление об устройстве конкретного языка и о его функционировании; помогает изучить специфику употребления отдельных слов и выражений; подготовить полную статистическую информацию о частотности лингвистических явлений и др. В частности, на основе корпуса удмуртских текстов у нас появляется возможность описать глагольную систему удмуртского языка, которая, к сожалению, на данный момент изучена совершенно неудовлетворительно по сравнению с другими частями речи удмуртского языка, а также с учетом роли глагола в организации речи в целом. При этом, по сравнению с русским языком, функциональная нагрузка глагола в удмуртском языке выше, что объясняется наличием системы причастных и деепричастных форм, а также изобилием отглагольных существительных, т. е. в тех случаях, когда в русском языке многие глаголы образуются от существительных, в удмуртском, наоборот – существительные от глаголов. Однако, несмотря на важность этой темы, рассмотрению категории времени в современном удмуртском языке посвящены пока лишь отдельные статьи В. И. Алатырева [1959], Б. Ш. Загуляевой [1986], Р. Ш. Насибуллина [1984] и ряда других, но работ монографического характера по данной проблеме на сегодня не существует. Также корпус позволяет изучить многие ранее не описанные темы в удмуртском языкознании, в том числе послеложное управление, которое по сей день представлено лишь единичными статьями Д. А. Ефремова [2009, 2013].

Во-вторых, корпус текстов способствует **подготовке различных видов словарей**. Так, на основе Национального корпуса русского языка на данный момент созданы следующие электронные словари[17]: «Грамматический словарь новых слов русского языка» (Е. А. Гришина, О. Н. Ляшевская), «Новый частотный словарь русской лексики» (О. Н. Ляшевская, С. А. Шаров), «Словарь русской идиоматики. Сочетания слов со значением высокой степени» (Г. И. Кустова), «Словарь глагольной сочетаемости непредметных имен русского языка» (О. Л. Бирюк, В. Ю. Гусев, Е. Ю. Калинина). Что касается Национального корпуса удмуртского языка, то разработка подкорпуса памятников письменности и диалектных текстов, несомненно, поможет создать исторический словарь удмуртского языка, которого на сегодня у нас нет; а также полный современный диалектологический словарь, поскольку имеющиеся на данный момент диалектологические словари зарубежных ученых Ю. Вихманна [Wichmann 1987] и Б. Мункачи [Munkácsi 1896] содержат материалы, собранные еще в конце XIX века.

В третьих, корпус, представляющий собой готовые размеченные тексты в электронном виде, послужит основой **для разработки переводчиков, систем распознавания и синтеза речи, а также других программ по удмуртскому языку**.

Программа проверки удмуртской орфографии

Процесс пополнения Национального корпуса удмуртского языка включает в себя три основных этапа: 1) сканирование и распознавание книг на удмуртском языке в программе ABBYY FineReader; 2) вычитку распознанных текстов (сверка с оригиналом); 3) загрузку на сайт подготовленных текстов. Перед загрузкой новых текстов обязательна процедура их обработки специальной программой для проверки удмуртской орфографии, поскольку как бы внимательно мы ни старались вычитывать распознанные *PDF*-файлы удмуртских книг, в любом случае, к сожалению, остаются ошибки, в том числе орфографические, устранить которые помогает специальное расширение для текстовых редакторов на основе Hunspell. Что собой представляет данная программа? Как она устроена?

Hunspell – свободная программа для проверки орфографии, предназначенная для языков со сложной системой словообразования и обширной морфологией [Hunspell]. Для проверки орфографии в Hunspell требуются два файла: 1) файл словаря (.dic), содержащий как можно больше слов определенного языка, и 2) файл аффиксов (.aff), определяющий значения специальных меток (флагов) в словаре.

На основе данной программы Григорий Григорьев разработал специальное расширение для проверки удмуртской орфографии[18]. Основой удмуртского Hunspell-словаря послужил удмуртско-русский словарь 2008 года [УРС], дополненный неологизмами русско-удмуртского словаря 2019 года [РУС I, РУС II].

[17] Словари находятся в открытом доступе на сайте http://dict.ruslang.ru/.
[18] Программа находится в свободном доступе на сайте https://github.com/vorgoron/udmspell.

Кратко рассмотрим устройство данной программы[19].

Файл словаря (.dic) содержит список удмуртских слов (по одному слову в каждой строке). После каждой лексемы следует слэш («/») и один или несколько флагов[20], обозначающих, к какой группе аффиксов файла (.aff) относится данное слово. Всего в системе используется 9 флагов:

a – для обозначения глаголов I-го спряжения (*-ыны*);

b – для обозначения глаголов, оканчивающихся на *-йыны* (данные глаголы выделены в отдельный класс, чтобы не загромождать группу глаголов I-го спряжения и в связи со специфичностью образования слов на *-йыны*);

c – для обозначения глаголов II-го спряжения (*-аны*/*-яны*);

d – для обозначения неодушевленных существительных;

g – для обозначения одушевленных существительных;

h – для обозначения имен существительных, образующих притяжательные формы индивидуального обладателя единственного числа с помощью *ы*-овых вариантов морфологических показателей посессивности[21], например: *пель* ‘ухо’ – *пель**ы*** ‘мое ухо’, *пель**ыд*** ‘твое ухо’, *пель**ыз*** ‘его ухо’; *ныл* ‘дочь’ – *ныл**ы*** ‘моя дочь’, *ныл**ыд*** ‘твоя дочь’, *ныл**ыз*** ‘его дочь’ и др.

j – для обозначения неодушевленных существительных, оканчивающихся на *-ие*, *-ия*; например: *авиация*, *Россия*, *предприятие* и др.

q – для обозначения числительных;

p – для обозначения прилагательных.

Например:

```
бергатыны/a
бырйыны/b
вераны/c
гудыри/d
нылаш/g
ныр/h
Удмуртия/j
тямыс/q
ӵуж/p
бергес
```

Второй файл (.aff) содержит группы аффиксов, обозначенные флагами. Например:

```
SFX a ыны й [дзлнст]ыны
SFX a ыны йд [дзлнст]ыны
SFX a ыны йз [дзлнст]ыны
SFX a ыны ймы [дзлнст]ыны
SFX a ыны йды [дзлнст]ыны
SFX a ыны йзы [дзлнст]ыны
SFX a ыны и [^дзлнстьй]ыны
SFX a ыны ид [^дзлнстьй]ыны
SFX a ыны из [^дзлнстьй]ыны
SFX a ыны имы [^дзлнстьй]ыны
SFX a ыны иды [^дзлнстьй]ыны
SFX a ыны изы [^дзлнстьй]ыны
```

[19] Более подробную информацию о данной программе можно найти на сайте http://mozilla-russia.org/projects/dictionary/hunspell.html

[20] Как правило, флаг представляет собой один символ (обычно, алфавитный).

[21] Согласно [ГСУЯ 1962], *ы*-овые варианты притяжательных маркеров при выражении индивидуального обладателя в единственном числе употребляются в удмуртском языке: 1) с существительными, именующими части тела (*йыр* ‘голова’, *ныр* ‘нос’, *тыбыр* ‘спина’ и др.); 2) с существительными, обозначающими отдельные внутренние или внешние свойства человека и некоторых других живых существ (*мылкыд* ‘настроение, желание’, *сям* ‘характер’, *лул* ‘душа’ и др.); 3) с существительными, именующими части или стороны вещи, предмета (*тыш* ‘тыльная часть предмета’, *йыл* ‘конец, кончик’, *выл* ‘поверхность’ и др.); 4) с существительными, выражающими отношения времени (*дыр* ‘время’, *нунал* ‘день’, *арес* ‘возраст’); 5) с существительным, именующими родственные отношения (*вын* ‘младший брат’, *ныл* ‘дочь’) [ГСУЯ 1962, 82–83].

SFX – наименование группы суффиксов;

а – наименование флага (в данном случае для обозначения глаголов I спряжения);

ыны – обозначение части слова, которое будет убираться с конца лексем, приведенных в первом файле словаря;

й, йд, йз, ймы, йды, йзы, и, ид, из, имы, иды, изы – суффиксы, которые будут подставляться вместо *ыны*.

[дзлнст]ыны – условие, при котором будут подставляться суффиксы **й, йд, йз, ймы, йды, йзы**.

[^дзлнстьй]ыны – условие, при котором будут подставляться суффиксы **и, ид, из, имы, иды, изы**.

В условиях используются регулярные выражения (подробнее см. [Forta 2004]). В частности, в квадратных скобках перечисляются буквы, одна из которых обязательно должна быть на том месте, где стоят эти скобки. Так, например, условию **[дзлнст]ыны** удовлетворяют слова с окончаниями *-дыны*, *-зыны*, *-лыны*, *-ныны*, *-сыны*, *-тыны*. То есть если в файле с расширением .dic есть лексемы, помеченные флагом «а» и заканчивающиеся на *-дыны*, *-зыны*, *-лыны*, *-ныны*, *-сыны*, *-тыны*, например, *бергатыны*/*а*, то будут образовываться новые слова с определенным суффиксом: *бергатӥ*, *бергатӥд*, *бергатӥз* и т. д. При этом если в условии после открывающей квадратной скобки добавлен символ «^», это означает, что те буквы, которые заключены в такие скобки, не должны находиться в данной позиции (происходит обратное, т. е. отрицание). Например, условию **[^дзлнстьй]ыны** подходят такие слова, как *кыткыны*/*а*, *пырыны*/*а*, *шуыны*/*а* и др.

Недостатком Hunspell является возможность определять одновременно в одном слове не более двух суффиксов. В удмуртском языке, относящемся к языкам агглютинативного типа, словоформы зачастую содержат в себе два и более формальных показателей. Например, в слове *кылзӥсьёсмылы* ‘нашим слушателям’ представлено 4 аффикса *ӥсь-ёс-мы-лы*, но в словаре аффиксов нельзя определить все эти аффиксы по отдельности, надо обозначить все либо как один длинный суффикс (*-ӥсьёсмылы*), либо как два аффикса (*-ӥсь-ёсмылы*), иначе программа не поймет. При этом для определения второго суффикса в конец впереди идущего добавляется *слэш* с нужным флагом, например:

SFX а ыны йсь/d [дзлнст]ыны

Здесь видим, что от глагола может образоваться существительное, т. к. флаг «а» используется для глаголов, а флаг «d» используется для существительных.

Заключение

Таким образом, благодаря словарной базе, связанной с файлом аффиксов, в котором заложены по возможности все морфологические варианты лексем основного словаря, программа выявляет орфографические ошибки, оставшиеся при вычитке текстов для корпуса. Кроме того, с такой же целью данное расширение проверки правописания можно использовать и в браузерах *Google Chrome*, *Opera* и *Firefox*.

Программа проверки удмуртской орфографии используется также в электронном удмуртско-русском словаре для *GoldenDict*, благодаря чему при вводе любой словоформы в поисковую строку мы в результатах получаем начальную форму удмуртского слова и его перевод на русский язык. Такой формат словаря, в первую очередь, удобен при переводе текстов с удмуртского языка на русский лицам, недостаточно хорошо владеющим удмуртским языком.

В целом, по нашему мнению, повсеместное использование данного расширения позволило бы стандартизировать удмуртское правописание и этим существенно облегчить работу корректоров, поскольку на данный момент, несмотря на наличие орфографического словаря по удмуртскому языку, в современных печатных изданиях (журналах, газетах, книгах и др.) встречаются различные варианты написания одного и того же слова.

Таким образом, корпусная лингвистика, сочетающая в себе как большой объем эмпирических данных, так и современные методы статистических расчетов и обработки информации, представляет собой относительно новый подход в языкознании и тем самым вызывает большой интерес у современных лингвистов и специалистов в области компьютерных технологий. По нашему мнению, и специалисты по удмуртскому языку в ближайшее время в своих исследованиях станут опираться на корпусные данные. Однако опыт ведущих мировых лингвистов показывает, что привлечение к анализу подобных материалов существенно меняет наши представления об устройстве языка. В связи с этим,

вероятно, в ближайшее время могут появиться труды, содержащие в себе совершенно новую концепцию, отличающуюся от традиционных точек зрения, что, несомненно, повлияет на развитие, в том числе, и удмуртского языкознания.

ЛИТЕРАТУРА

Алатырев В. И. Вопросы удмуртского языкознания. Т. 1. Ижевск, 1959. 213 с.

Архангельский Т. А. Интернет-корпуса финно-угорских языков России // Ежегодник финно-угорских исследований. 2019. Т. 13. Вып. 3. С. 528–537.

ГСУЯ – Грамматика современного удмуртского языка: фонетика и морфология / Удм. НИИ ист., экон., яз. и лит.; отв. ред. П. Н. Перевощиков. Ижевск, 1962. 376 с.

Ефремов Д. А. Удмурт кылын каронкыллэн управлениез сярысь // Вестн. Удм. ун-та. Сер. История и филология. 2009. Вып. 1. С. 43–54.

Ефремов Д. А. О послеложном управлении в удмуртском языке // Вестн. Удм. ун-та. Сер. История и филология. 2013. Вып. 2. С. 8–15.

Жевнерович Е. Э. Корпус текстов в научном исследовании // Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития: Материалы II Международной научно-практической конференции. Минск, 2018. С. 25–32.

Загуляева Б. Ш. Прошедшее длительное и прошедшее многократное время глаголов в удмуртском языке // Вопросы фонетики и грамматики удмуртского языка: Сборник статей / Науч.-исслед. ин-т при Совете Министров Удмурт. АССР; отв. ред. В. М. Вахрушев, В. К. Кельмаков. Устинов, 1986. С. 62–70.

Копотев М. Введение в корпусную лингвистику. Прага, 2014. (Электронное учебное пособие).

Насибуллин Р. Ш. О некоторых аналитических формах глагола в удмуртском языке // Вопросы грамматики удмуртского языка: Сборник статей / НИИ при Сов. Мин. Удмурт. АССР; отв. ред. В. М. Вахрушев. Ижевск, 1984. С. 38–44.

РУС I – Русско-удмуртский словарь: В 2 т. Более 55 000 слов. Т. 1 (А–О) / Л. М. Ившин, С. А. Максимов, О. В. Титова, Л. Е. Кириллова, Л. Л. Карпова, Т. Р. Душенкова, А. В. Егоров, А. А. Шибанов; отв. ред. Л. М. Ившин; УдмФИЦ УрО РАН. Ижевск, 2019. 936 с.

РУС II – Русско-удмуртский словарь: В 2 т. Более 55 000 слов. Т. 2 (П–Я) / Л. М. Ившин, С. А. Максимов, О. В. Титова, Л. Е. Кириллова, Л. Л. Карпова, Т. Р. Душенкова, А. В. Егоров, А. А. Шибанов; отв. ред. Л. М. Ившин; УдмФИЦ УрО РАН. Ижевск, 2019. 1016 с.

УРС – Удмуртско-русский словарь: Около 50 000 слов / Сост. Т. Р. Душенкова, А. В. Егоров, Л. М. Ившин и др.; отв. ред. Л. Е. Кириллова. Ижевск, 2008. 925 с.

Forta B. Sams Teach Yourself Regular Expressions in 10 Minutes. Indianapolis, 2004. 192 p.

Hunspell [Эл. ресурс]. URL: https://en.wikipedia.org/wiki/Hunspell (Дата обращения: 10.04.2020).

McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge, 2012. 312 p.

Munkácsi B. A votják nyelv szótára. Budapest, 1896. VXI + 758 l.

Wichmann Y. Wotjakischer Wortschatz / Aufgezeichnet von Yrjö Wichmann. Bearbeitet von T. E. Uotila, Mikko Korhonen. Herausgegeben von Mikko Korhonen. Helsinki, 1987. 421 S.

Поступила в редакцию 04.04.2020

Безенова Мария Петровна,
кандидат филологических наук, научный сотрудник
Удмуртский институт истории, языка и литературы УдмФИЦ УрО РАН
426004, Россия, г. Ижевск, ул. Ломоносова, 4
Институт системного программирования им. В. П. Иванникова РАН
109004, г. Москва, ул. А. Солженицына, 25
E-mail: mary_kaj@mail.ru

Григорьев Григорий Леонидович,
младший научный сотрудник
Лаборатория машинного обучения и обработки «больших» данных производственных киберсистем
УдмФИЦ УрО РАН
426068, Россия, г. Ижевск, ул. Барышникова, 53
E-mail: exey13@gmail.com

M. P. Bezenova, G. L. Grigoriev

THE ROLE OF THE UDMURT SPELL CHECKER IN REPLENISHMENT OF THE UDMURT NATIONAL CORPUS

DOI: 10.35634/2224-9443-2020-14-3-549-556

Corpus linguistics is currently one of the most popular sections of linguistics. Most of the major languages of the world today already have their own digital corpora of tens and hundreds of millions of word usage. Recently, special attention has also been paid to the creation of text corpus in the languages of the peoples of Russia, since, on the one hand, corpus research allows you to look at the structure of the language from a completely different perspective, on the other hand, the corpus is a kind of form of storing language data. The article describes the Udmurt National Corpus, which has been developed since the end of 2019 by the staff of the philological research department of the Udmurt Institute of History, Language and Literature of the Udmurt Federal Research Center of the Ural Branch of the Russian Academy of Sciences. It speaks in detail about the capabilities of the information and reference system being created at the moment, as well as about the prospects for using the corpus of texts when conducting research, preparing dictionaries, and creating various programs in the Udmurt language.

The article also deals with the Hunspell-based Udmurt spell checker developed by Grigory Grigoriev, which plays an important role in replenishing the Udmurt National Corps. Before uploading new texts to the site, all of them are subjected to a mandatory check for spelling errors that could remain during their proofreading. This extension for text editors, thanks to the vocabulary database associated with the affix file, which contains all possible morphological variants of the lexemes of the main dictionary, identifies spelling errors in the text, allowing you to upload the most verified texts to the website of the Udmurt National Corpus.

Keywords: corpus linguistics, corpus of texts, Udmurt language, national corpus, opportunities and prospects of the corpus, spell check, Hunspell.

Citation: Yearbook of Finno-Ugric Studies, 2020, vol. 14, issue 3, pp. 549–556. In Russian.

REFERENCES

Alatyrev V. I. *Voprosy udmurtskogo yazykoznaniya* [Issues of Udmurt Linguistics]. Izhevsk, 1959. Vol. 1. 213 p. In Russian.

Arkhangel'skij T. A. Internet-korpusa finno-ugorskikh yazykov Rossii [Internet-Corpora for Finno-Ugric Languages of Russia]. Ezhegodnik finno-ugorskikh issledovanij [Yearbook of Finno-Ugric Studies], 2019, vol. 13, no. 3, pp. 528–537. In Russian.

GSUYa – *Grammatika sovremennogo udmurtskogo yazyka: fonetika i morfologiya* [Grammar of the Modern Udmurt Language: Phonetics and Morphology]. Ed. P. N. Perevoshchikov. Izhevsk, 1962. 376 p. In Russian.

Efremov D. A. Udmurt kylyn karonkyllen upravleniez syarys' [About Verb Management in the Udmurt Language]. Vestnik Udmurtskogo universiteta. Seriya Istoriya i filologiya [Bulletin of Udmurt University. Series History and Philology], 2009, no. 1, pp. 43–54. In Udmurt.

Efremov D. A. O poslelozhnom upravlenii v udmurtskom yazyke [About Postposed Management in the Udmurt Language]. Vestnik Udmurtskogo universiteta. Seriya Istoriya i filologiya [Bulletin of Udmurt University. Series History and Philology], 2013, no. 2, pp. 8–15. In Russian.

Zhevnerovich E. E. Korpus tekstov v nauchnom issledovanii [Text Corpus in Scientific Research]. *Lingvistika, lingvodidaktika, lingvokul'turologiya: aktual'nye voprosy i perspektivy razvitiya: Materialy II Mezhdunarodnoj nauchno-prakticheskoj konferentsii* [Linguistics, Linguodidactics, Linguoculturology: Actual Issues and Development Prospects: Materials of the II International Scientific and Practical Conference]. Minsk, 2018, pp. 25–32. In Russian.

Zagulyaeva B. Sh. Proshedshee dlitel'noe i proshedshee mnogokratnoe vremya glagolov v udmurtskom yazyke [Past Long and Past Multiple Tenses of Verbs in the Udmurt Language]. Voprosy fonetiki i grammatiki udmurtskogo yazyka [Questions of Phonetics and Grammar of the Udmurt Language]. Digest of articles. Ed. V. M. Vakhrushev, V. K. Kelmakov. Ustinov, 1986, pp. 62–70. In Russian.

Kopotev M. *Vvedenie v korpusnuyu lingvistiku* [Introduction to Corpus Linguistics]. Prague, 2014. In Russian.

Nasibullin R. Sh. O nekotorykh analiticheskikh formakh glagola v udmurtskom yazyke [On Some Analytical Forms of the Verb in the Udmurt Language]. Voprosy grammatiki udmurtskogo yazyka [Grammar Issues of the Udmurt Language]. Digest of articles. Ed. V. M. Vakhrushev. Izhevsk, 1984, pp. 38–44. In Russian.

RUS I – *Russko-udmurtskij slovar': V 2 t. Bolee 55 000 slov* [Russian-Udmurt Dictionary: in 2 vol. Over 55 000 Words]. Comp. L. M. Ivshin, S. A. Maksimov, O. V. Titova, L. E. Kirillova, L. L. Karpova, T. R. Dushenkova, A. V. Egorov, A. A. Shibanov. Ed. L. M. Ivshin. Izhevsk, 2019. Vol. 1: A–O. 936 p. In Russian and Udmurt.

RUS II – *Russko-udmurtskij slovar': V 2 t. Bolee 55 000 slov* [Russian-Udmurt Dictionary: in 2 vol. Over 55 000 Words]. Comp. L. M. Ivshin, S. A. Maksimov, O. V. Titova, L. E. Kirillova, L. L. Karpova, T. R. Dushenkova, A. V. Egorov, A. A. Shibanov. Ed. L. M. Ivshin. Izhevsk, 2019. Vol. 2: P–Ya. 1016 p. In Russian and Udmurt.

URS – *Udmurtsko-russkij slovar': Okolo 50 000 slov* [Udmurt-Russian Dictionary: About 50 000 Words]. Comp. T. R. Dushenkova, A. Egorov, L. M. Ivshin, and others; Ed. L. E. Kirillova. Izhevsk, 2008. 925 p. In Udmurt. In Russian.

Forta B. *Sams Teach Yourself Regular Expressions in 10 Minutes*. Indianapolis, 2004. 192 p. In English.

Hunspell. URL: https://ru.wikipedia.org/wiki/Hunspell (accessed 10 April 2020). In English.

McEnery T., Hardie A. *Corpus Linguistics: Method, Theory and Practice.* Cambridge, 2012. 312 p. In English.

Munkácsi B. *A votják nyelv szótára* [Dictionary of the Votyak Language]. Budapest, 1896. VXI + 758 p. In Hungarian.

Wichmann Y. *Wotjakischer Wortschatz* [Votyak Dictionary]. Helsinki, 1987. 421 p. In German.

Received 04.04.2020

Bezenova Maria Petrovna,
Candidate of Philology, Researcher
Udmurt Institute of History, Language and Literature, UdmFRC UB RAS
Lomonosova st., 4, Izhevsk, 426004, Russian Federation
Ivannikov Institute for System Programming, RAS
A. Solzhenitsyna st., 25, Moscow, 109004, Russian Federation
E-mail: mary_kaj@mail.ru

Grigoryev Grigory Leonidovich,
Junior Researcher
Laboratory of Machine Learning and Processing of Big Data of Production Cybersystems, UdmFRC UB RAS
Baryshnikova st., 53, Izhevsk, 426068, Russian Federation
Email: exey13@gmail.com