БАШКИРСКИЙ ЯЗЫК И ЛИТЕРАТУРА В УСЛОВИЯХ ГЛОБАЛИЗАЦИИ И ПОЛИЭТНИЧЕСКОЙ СРЕДЫ: ОПЫТ И ПЕРСПЕКТИВЫ

ISBN 978-5-91608-181-7 С. 72-76

https://doi.org/10.31833/conf_ihll2019.014

УДК 81.512.157

В. В. Бочкарев

ИГИиПМНС СО РАН Петровского, 1, Якутск, 677027, Россия
E-mail: Uus-Aldan@mail.ru

ЭЛЕКТРОННЫЙ КОРПУС ЯЗЫКА САХА

В статье описывается процесс создания электронного корпуса якутского языка, проблемы и пути решения. Автором для ускорения работы по набору материала базы данных предлагается неаннотированный вариант корпуса, основанный на работе универсального лингвистического анализатора.

Ключевые слова: корпусная лингвистика, якутский язык, обработка текста, лингвистический анализатор.

Проделана огромная работа учеными-якутоведами по выявлению и интерпретации основных базовыхкатегорий языка саха. Дальнейшие исследования по углублению знаний о языке требуют обработки колоссального количества информации, исчисляемыми уже не тысячами, а миллионами и даже миллиардами примеров. Такой объем перебрать вручную невозможно или же потребует многолетнего чернового труда, что в условиях ускоряющегося ритма жизни недопустимо.

Поэтому в последнее время все более актуальными становятся различные способы автоматизации научного труда ученого-лингвиста. В этом отношении выделяется корпусная лингвистика, занимающаяся созданием общих унифицированных принципов представления сверхбольших массивов языковых данных – корпусов, непосредственным их составлением, а также выполнением конкретных экспериментальных лингвистических исследований на базе этих данных [Сиразитдинов, 2013, 66].

Поскольку дальнейшее развитие якутской филологии напрямую зависит от автоматизации и отсутствие которой может привести к значительному отставанию от вызовов современного мира, перед разработчиками электронного корпуса якутского языка стоит задача не только создания работающей модели, но и скорейшего доведения ее до уровня национальных языковых корпусов других народов, начавших работу ранее. На данный момент все мажоритарные языки обзавелись своими национальными корпусами. Ведутся корпусные разработки и по языкам народов России. Отдельно отметим научные разработки и корпусные проекты по языкам тюркской группы [Сиразитдинов, 2013, 66]. Например: национальный корпус русского языка в своей базе имеет 364 881 378 словоупотреблений [2-10]; «Письменный татарский корпус» — более 356 миллионов слов, около 4,5 миллионов словоформ [2-10]; «Бурятский корпус» — более 2 миллионов 200 тыс. словоупотреблений [2-10]; татарский национальный корпус «Туган тел» — более 26 миллионов словоупотреблений.

Использование возможностей вычислительно-аналитической мощи современных компьютеров позволят получить инновационные результаты как в области теоретической лингвистики (получение новых знаний об устройстве языка), так и в области прикладной лингвистики (модернизация методов лингвистических исследований, внедрение новых технологий для автоматической обработки текстов) [Торотоев и др, 2018].

По этой причине в Институте гуманитарных исследований и проблем коренных малочисленных народов Севера СО РАН ведется активная работа по созданию электронного корпуса якутского языка. Перед Институтом стоит задача не только разработки корпуса, но и скорейшего его доведения до репрезентативного уровня, годного для научных исследований. Автором статьи предлагается нестандартный вариант стартового неаннотированного корпуса, роль разметок в котором выполняет универсальный лингвистический анализатор.

Якутский язык относится к агглютинативному типу. В этих языках словоформа образуется путем присоединения к основе в строгом порядке однозначных стандартных аффиксов; границы морфем отчетливы, фонетические изменения на стыках морфем подчиняются строгим правилам [Дыбо, 2014, 20], что делает якутский язык очень удобным для компьютерного анализа, но в то же время попытки построить парадигму конкретного слова демонстрируют ее чрезвычайную сложность и многомерность, что обусловлено большим числом словоизменительных аффиксов [Дыбо, 2014, 20]. Поэтому при разработке алгоритма универсального лингвистического анализатора был выбран двукомпонентный метод, который основывается на логической переборке частей слова и сопоставлении с матрицей, составленной на основе работ выдающихся ученых-языковедов в соответствии с законами грамматики якутского языка. Такая двойная система анализа исключает ошибки и позволяет программе анализировать не только существующие слова, но и дает возможность генерации новых слов, т.е. имеет зачатки «искусственного интеллекта», способного понять значение неизвестных, не записанных в словарь слов и присваивать им соответствующие переменные, что в будущем станет основой программ, имитирующих мыслительные процессы.

Используется система тэгов, базирующуая на Лейпцигских правилах глоссирования, что соответствует унифицированной морфологической разметке тюркских языков.

На данный момент программа представляет собой универсальный «научный калькулятор» для лингвиста и является первой разработкой подобного рода для якутского языка. До сих пор все исследования лингвистов производились вручную, без автоматизации, тогда как в других отраслях, например, у физиков и инженеров, используются программы для расчетов и моделирования.

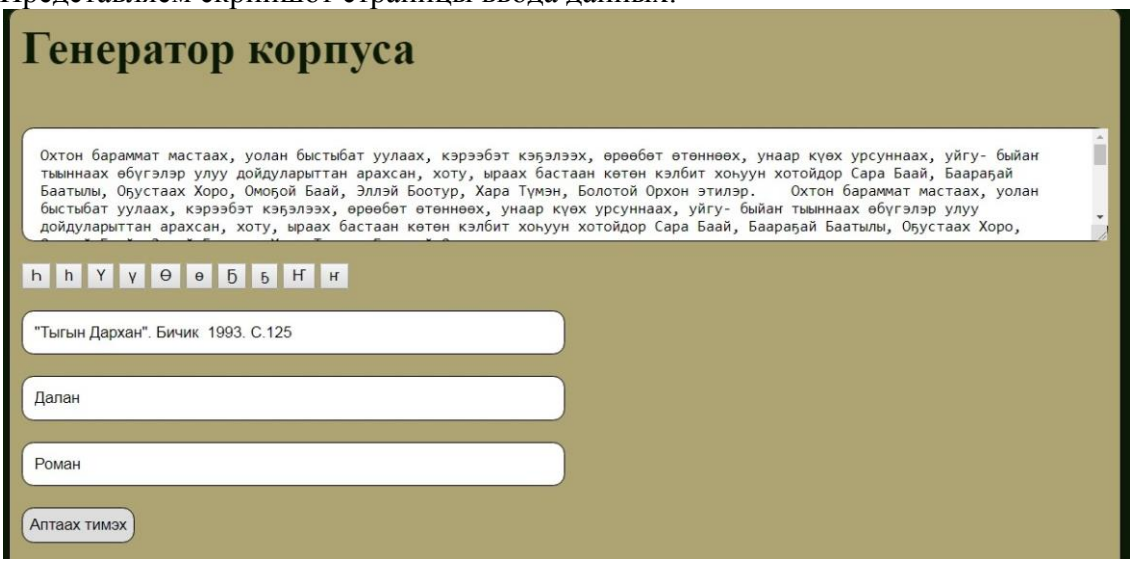
Основными функциями программы являются: Разбивка по слогам; фонетический разбор; морфологический разбор; визуализация морфологического разбора в виде схемы; нахождение древнего этимона; поиск лексического значения слова из 15-томного “Большого толкового словаря якутского языка”.

Использование анализатора вместо многочисленных меток в базе данных корпуса позволяет на начальном этапе обойтись максимально упрощенной схемой без существенного урона для качества, что, в свою очередь, позволяет многократно ускорить процесс заполнения информацией.

Схема вывода информации:

<table>
<tr><td colspan="2">Ввод слова</td></tr>
<tr><td>Корпус</td><td>Анализатор</td></tr>
<tr><td>Поиск в базе соответствий с введенным словом.</td><td>Анализ введенного слова:
1. Разбивка по слогам;
2. Фонетический разбор;
3. Морфологический разбор;
4. Визуализация морфологического разбора в виде схемы;
5. Нахождение древнего этимона;
6. Поиск лексического значения слова из 15-томного “Большого толкового словаря якутского языка”.</td></tr>
<tr><td colspan="2">Вывод обобщенного результата на экран.</td></tr>
</table>

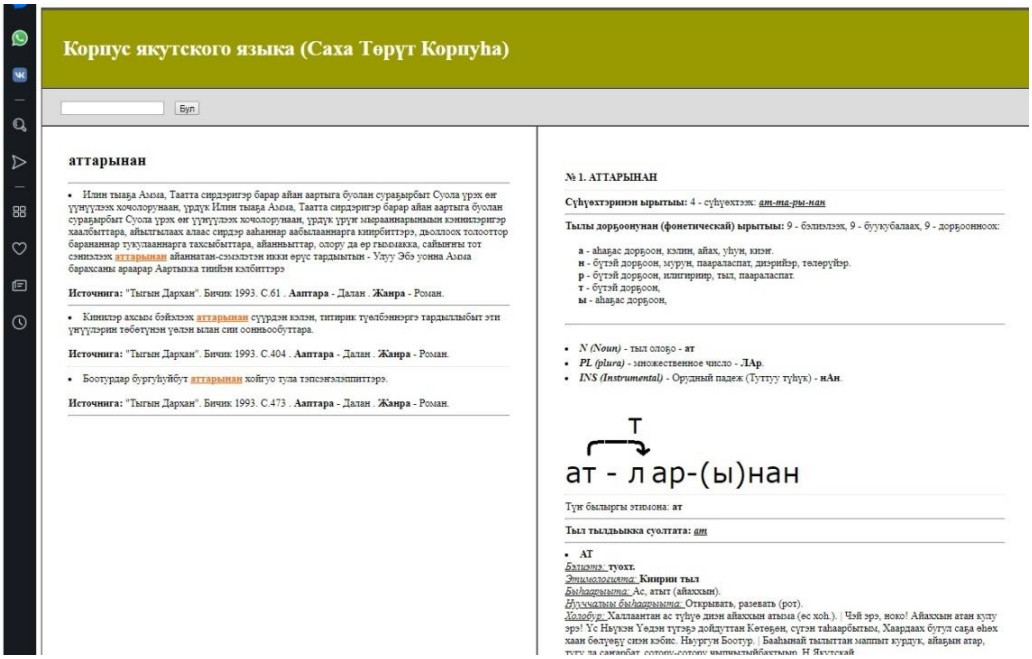
Представляем скриншот страницы ввода данных:



1.Поле для текста; 2. Источник; 3. Автор; 4. Жанр.

После нажатия кнопки база автоматически пополняется новыми данными.

При чтении информации работа скрипта анализатора обеспечивает вывод минимально необходимого количества обработанной информации для исследовательской работы лингвиста. В будущем, скрипт модуля чтения и вывода информации можно как угодно настраивать и модернизировать под определенные цели и задачи, не опасаясь повредить всю систему, поскольку это не затронет базу данных.



После достижения достаточного объема базы (по подсчетам Леонтьева Н.А. составляет примерно 7 млн. словоформ [Леонтьев, 2019]), финишной настройки универсального анализатора и окончательного определения нужных функций для фильтра при помощи специального скрипта, на основе цикла по количеству полученных в итоге словоформ, будет произведено автоматическое преобразование простейшей базы на стандартную с пометами, что повысит надежность, точность и скорость работы системы.

Есть возможность сделать систему таким образом, чтобы анализатор работал не при чтении информации, а при записи в базу и тем самым сразу же получить аннотированный

текст. Но такой вариант плох тем, что любые вносимые изменения (дополнение и удаление функций, разделов, оптимизация алгоритма работы и другие всевозможные измения), которых, ввиду того, что корпус создается с нуля, будет много, потребуют редактирования всей структуры базы с миллионами записей. Это с большой вероятностью может вызвать ошибки, поиск и исправления которых очень трудоемко. Поэтому будет лучше, если анализатор работает на выводе. Изначально такая схема организации работы выбрана для экономии людских, финансовых и временных ресурсов.

В данное время система работает в закрытом режиме и находится на стадии заполнения базы, тестирования и отладки скриптов.

Помимо вышеизложенного в настоящее время параллельно ведется работа по созданию электронного корпуса якутского языка к.т.н., доцентом СВФУ Леонтьевым Н. А. Им был разработан машинный газетный корпус якутского языка, который содержит более 20 тыс. статей, более 12 млн. словоупотреблений и 250 тыс. словоформ [Леонтьев, 2018, 94]. В данное время эта информационная система тоже работает в закрытом режиме, идет тестирование [Леонтьев, 2018, 97].

Силами группы энтузиастов ведется работа над проектом на сервисе Github.com – социальная сеть программистов, где все желающие могут присоединиться для создания корпуса якутского языка на основе облачных технологий.

Таким образом, в целом работа по созданию электронного корпуса якутского языка идет форсированно, разными коллективами, используя всевозможные пути и варианты решения данной задачи.

Литература

Бурятский корпус // URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 20.09.2019).

Дыбо А. В., Шеймович А. В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. №2(36) С.20-26.

Леонтьев Н. А. Вопрос о размере машинного корпуса на примере якутского языка // Электронный научно-практический журнал “Современные научные исследования и инновации”.URL: http://web.snauka.ru/issues/2015/11/58769 (дата обращения: 20.09.2019).

Леонтьев Н. А. Информационная система “Электронный корпус якутского языка” // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2018. №12. С.94-97

Национальный корпус русского языка // URL: http://www.ruscorpora.ru/new/corpora-stat.html (дата обращения: 20.09.2019).

Письменный татарский корпус // URL: https://www.corpus.tatar (дата обращения: 20.09.2019).

Сиразитдинов З. А., Полянин А. И., Ибрагимова А. Д., Ишмухаметова А. Ш. Корпусы башкирского языка: принципы разработки // Проблемы востоковедения. 2013/4 (62). С.66.

Сичинава Д. В. К задаче создания корпусов русского языка // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2002. №11 С. 25-31.

Татарский национальный корпус «туган тел». http://web-corpora.net/TatarCorpus/search/index.php?interface_language=ru (дата обращения: 20.09.2019).

Торотоев Г. Г., Торотоева С. Г. Морфологическое аннотирование словоизменительных категорий имени существительного в языке саха // Сравнительно-сопоставительное изучение тюркских и монгольских языков : материалы Международной научно-практической конференции. г.Якутск, 18-19 октября 2018 г. / Якутск, 2018. С. 254.

V. V. Bochkarev

Institute for Humanities Research and Indigenous Studies of the North
Petrovsky, 1, Yakutsk, 677027, Russia
E-mail: Uus-Aldan@mail.ru

COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES IN LINGUISTICS

The article describes the progress of work on the creation of the electronic corpus of the Yakut language, problems and solutions. The author offers a variant based on a universal linguistic analyzer to speed up text input into the database.

Keywords: corpus linguistics, Yakut language, text processing, linguistic analyzer.

References

Buryat building // URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (date accessed: 20.09.2019).

Dybo A. V., A. V. Samovich Automatic morphological analysis for corpora of Turkic languages // Philology and culture. 2014. No. 2(36) Pp. 20-26.

Leontiev N. A. The question of the size of the machine body on the example of the Yakut language // Electronic scientific and practical journal "Modern scientific research and innovation".URL: http://web.snauka.ru/issues/2015/11/58769 (date accessed: 20.09.2019).

Leontiev N. A. Information system "Electronic corpus of Yakut language" // Modern science: actual problems of theory and practice. Series: Natural and technical Sciences. 2018. No. 12. Pp. 94-97

National corpus of the Russian language // URL: http://www.ruscorpora.ru/new/corpora-stat.html (accessed 20.09.2019).

Written Tatar corpus // URL: https://www.corpus.tatar (date accessed: 20.09.2019).

Sirazitdinov Z. A., Polyanin A. I., Ibragimova A. D., Ishmukhametova A. Sh. Corpus of the Bashkir language: principles of development // Problems of Oriental studies. 2013/4 (62). Pp. 66.

Sichinava D. V. To the problem of creating Russian language corpus // Scientific and technical information. Series 2: Information processes and systems. 2002. No. 11 Pp. 25-31.

Tatar national corpus "Tugan tel". http://web-corpora.net/TatarCorpus/search/index.php?interface_language=ru (date accessed: 20.09.2019).

Torotoev G. G., Torotoeva S. G. Morphological annotation of noun inflection categories in the Sakha language/ /Comparative study of Turkic and Mongolian languages: materials of the International scientific and practical conference. Yakutsk, 18-19 October 2018 / Yakutsk, 2018. Pp. 254.