ИНФОРМАЦИОННАЯ СИСТЕМА «ЭЛЕКТРОННЫЙ КОРПУС ЯКУТСКОГО ЯЗЫКА»

INFORMATION SYSTEM "ELECTRONIC CORPUS OF YAKUT LANGUAGE"

N. Leontiev
N. Neustroev

Summary. This paper describes the structure and capabilities of the information system for the collection, analysis and processing of electronic texts in the Yakut language. The architecture of the software is described, the problems connected with preparation and input of texts are considered. The schemes of the information system are given.

Keywords: machine language corpus, text processing, Yakut language.

Леонтьев Ньургун Анатольевич
К.т.н., доцент, Северо-Восточный федеральный университет им. М. К. Аммосова, Якутск
leonza@mail.ru
Неустроев Никита Сергеевич
Северо-Восточный федеральный университет им. М. К. Аммосова, Якутск
neustroev-nikita@bk.ru

Аннотация. В данной работе описывается структура и возможности информационной системы по сбору, анализу и обработке электронных текстов на якутском языке. Описывается архитектура программного обеспечения, рассматриваются проблемы связанные с подготовкой и вводом текстов. Приводится схемы работы информационной системы.

Ключевые слова: машинный языковой корпус, обработка текста, якутский язык.

Для развития методов обработки естественного языка в мире часто применяют машинный электронный корпус. В мире существуют множество машинных корпусов на различных языках, в том числе и на русском, в России тоже имеются корпуса на других языках кроме русского [1], такие как татарский, башкирский, чувашский, коми, калмыкский, тувинский [2] и других. Существуют программные средства для создания корпусов и базы данных распространенных языков, но проблема автоматизированной обработки миноритарных языков остается острой, так как необходимы специалисты по данным языкам и соответственно, большая работа по сбору и анализу данных.

Якутский язык — язык одного из восточных тюрских народов Российской Федерации (самоназванием саха). Якуты — коренное население республика Саха (Якутия) Общая численность якутов по данным всероссийской переписи населения 2010 г. составляло 478,1 тысяч человек. Письменность основана на кириллице, для применения имеются 5 дополнительных национальных символов, на якутском языке издаются учебники, пишутся художественные книги, выпускаются радио и теле-материалы, создаются электронные ресурсы. Имеется богатый материал для сбора машинного языкового корпуса.

Авторами был разработан и собран машинный газетный корпус якутского языка [3], который содержит более 20 тыс.статей, более 12 млн. словоупотреблений. На его основе были созданы программы для определения автора текста [4], для определения языка текста [5], составления дикторского текста для задачи синтеза и распознавания речи. При обработке корпуса был созданы электронные словари, в частности словарь имен собственных [6], частотные словари. В результате имеются базы данных словоформ из 250 тыс. единиц, базы аффиксов и лемм, базы N-грамм из словосочетаний и словари.

Информационная система основана на сервере Apache и при использовании базы данных MySQL. Доступ осуществляется через тонкий клиент — браузер. Веб-страницы созданы с применением HTML и CSS. Применяемая кодировка Unicode UTF-8, которая поддерживает в дополнительной раскладке специальные символы якутского языка. Информационная система написана на языке PHP, для корректной работы с кодировкой UTF-8 была использована библиотечка Multibyte String Functions.

Размер текстового поля для ввода текста в базе данных определен как LONGTEXT, т.е. текст объемом до 4 Гб. Для представления обычно берут отрывок в тексте до десятки тысяч слов, так что такой размер является все-таки немного избыточным, но некоторые тексты могут быть размером больше 64 Кб.

Для более качественной работы с машинным языковым корпусом разрабатывается информационная система «Электронный корпус якутского языка».

Для морфологической разметки была разработана многопользовательская система ручной разметки [7].

[T1]= NКыым-85 [T2]Сонун, [T3]сэргэх [T4]буолан [T5]иһэр [T6]Туох [T7]ханнык [T8]иннинэ [T9]саамай [T10]сөбүлээн [T11]ааҕар [T12]«Кыым» [T13]хаһыаппыт [T14]85 [T15]сааһын [T16]туолар [T17]үбүлүөйүнэн [T18]истиҥник [T19]эҕэрдэлиибин. [T20]Хаһыат [T21]үлэһиттэригэр [T22]туйгун [T23]доруобуйаны, [T24]дьолу-соргуну, [T25]айымньылаах [T26]үлэни [T27]баҕарабын!

Результат ручной разметки: [T1]= N [T2]=ADJ [T3] =ADJ [T4]=CONV [T5]=V [T6]=PN [T7] =PN [T8]=POST [T9]=PART [T10] =CONV [T11]=PRT [T12]=N [T13]=N// [T14]=NUM [T15]=N// [T16]=PRT [T17]=N6 [T18]=ADV [T19]=V [T20]=N [T21]=N//3 [T22]=ADJ [T23]=N4 [T24]=N4-N4 [T25]=Nлаах [T26]=T4 [T27]=T [T28]=PN

Рис. 1. Ручная разметка текста

Морфологический анализ якутского языка

Якутские буквы h ҥ Y ү Ө ө Б б Ҥ ҥ ынахтарбытынааҕар Анализ

Анализ слова: **ынахтарбытынааҕар**
1 affiks="нааҕар" tip="noun" opis="Сравнительный падеж (В сравнении с кем? В сравнении с чем?)" razmetka="сравнительный падеж!!!"
1 affiks="ы" tip="noun" opis="Винительный падеж (Кого? Что?)" razmetka="acc"
1 affiks="быт" tip="noun" opis="Аффикс принадлежности (наш)" razmetka="принад.!"
2 affiks="быт" tip="verb" opis="Настоящее будущее время 1 лицо утвердительная положительная множественное число" razmetka="наст.буд.вр. 1 лицо множ."
3 affiks="быт" tip="verb" opis="Прошедшее результативное время 3 лицо " razmetka="Прош. результ. врем. 3 лицо"
4 affiks="быт" tip="verb" opis="причастие первичное" razmetka="причастие"
1 affiks="тар" tip="noun" opis="множественное число" razmetka="pl"
2 affiks="тар" tip="глаг" opis="условное наклонение" razmetka="условное наклонение"
1 slovo="ынах" lemma="ынах" rus="корова" tip="сущ"

Рис 2. Морфологический анализ

Данная система позволяет вести ручную разметку текста отдельно для каждого пользователя, что позволяет в дальнейшем выбрать лучший вариант. В связи с малой скоростью обработки вручную, была поставлена задача по разработке автоматизированной системы разметки.

Ускорение работы по разметки с помощью автоматизированной системы имеется, но все находится в стадии разработки, все усложняется из-за большого количества грамматических правил, согласования окончаний якутского языка.

На рисунке 2 приведен неоптимизированный результат морфологического анализа якутского слова «*ынахтарбытанааҕар*» — «*по сравнении с нашими коровами*».

Идет разработка программы для автоматизированной морфологической разметки. Имеются сложности в разрешении неоднозначностей и омонимов, так как

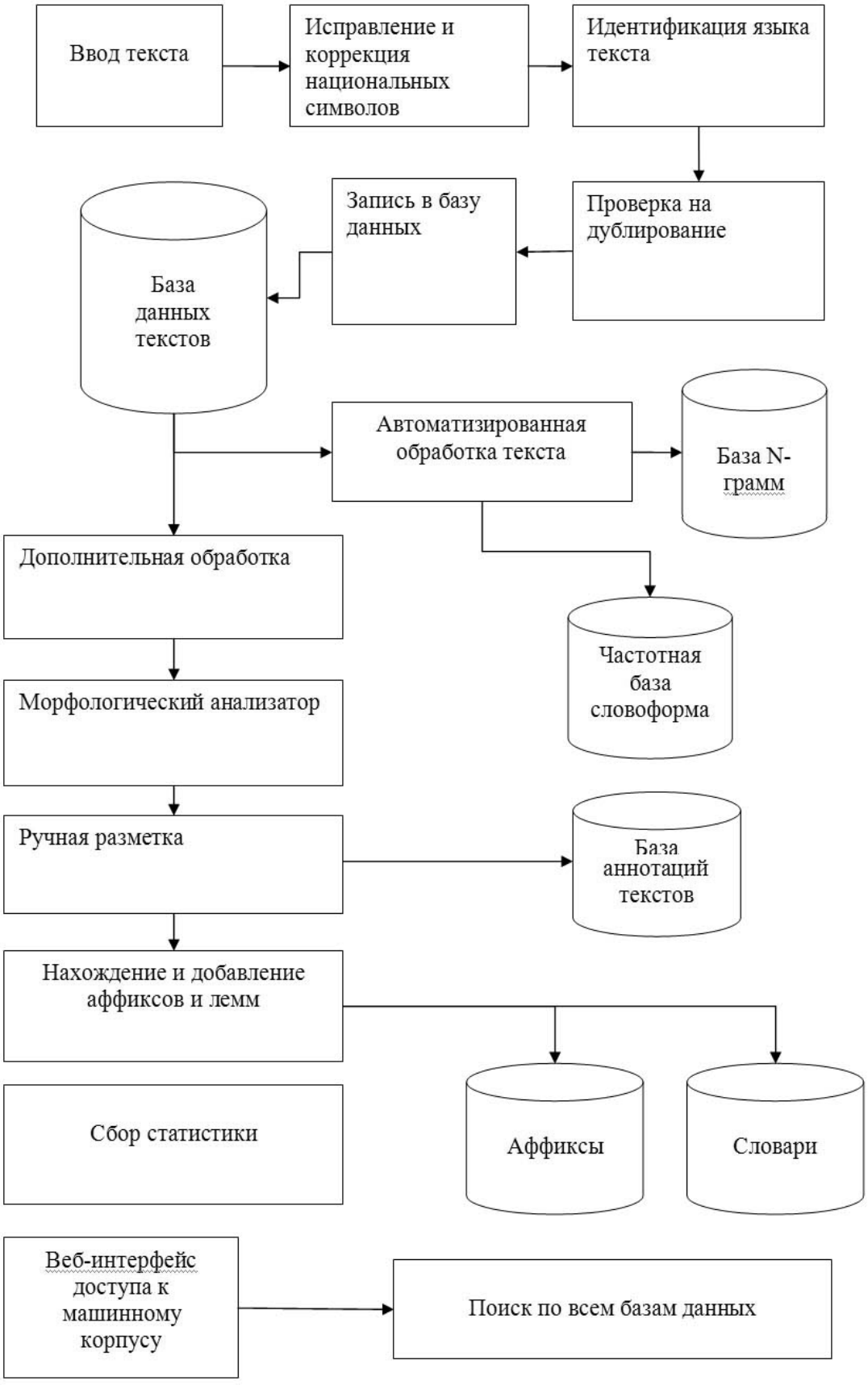

Рис. 3. Схема работы информационной системы

в якутском языке имеются более 1000 часто используемых омонимов, еще из-за агглютинативного типа языка образуются новые словоформы, совпадающие с основными и образующимися формами слов.

Имеется проблемы с правописанием заимствованных слов, иногда существуют десятки вариантов одного заимствованного слова, так как это бывают неустановившийся словоформы.

На рис. 3 приведена схема работы с указанием существующих баз данных и алгоритмом работы информационной системы.

Идентификация пользователя осуществляется по сочетанию уникального логина и пароль. Пароль сохраняется в базе пользователей в виде md5-хэш функции. Для восстановления пароля генерируется новый пароль и отправляется на почту. В настройках пользователь может поменять старый пароль на новый, с отправкой на почту сообщения об изменении пароля. Работа ведется без применения безопасного протокола SSL.

Новых пользователей с правами записи в базу данных может заводить только администратор. Существует три уровня доступа: гость — с правами поиска, редактор — с правами ввода текста, редактирования баз данных, администратор — с правами ввода пользователей и правами предыдущих уровней доступа.

Коррекция национальных символов необходимо, так как тексты раньше писались без использования кодировки Unicode. Существуют множество вариантов самодельных шрифтов и символов, которые не соответствуют стандарту Unicode.

Идентификация языка применяется для определения языка текста, так как он предотвращает ошибочный ввод текста на другом языке.

В ходе автоматизированной обработке идет проверка на дублирование текста, затем текст разбивается на буквы, дифтонги, диграфы, слова, словосочетания из N-грамм, до 6-грамм и создаются записи в соответствующих базах данных.

В ходе дополнительной обработки создаются аннотированные тексты, морфологическая разметка производится в ручном и автоматическом режиме.

Обычный пользователь может просмотреть частотные словари и использовать поисковую систему для поиска словоформы, также ему предлагаются словосочетания из биграмм, а также автоматический морфологический разбор словоформы.

В данное время информационная система работает в закрытом режиме, идет тестирование автоматической обработки текста.

ЛИТЕРАТУРА:

1. Arkhangelskiy T., Medvedeva M. Developing morphologically annotated corpora for minority languages of Russia // В сборнике: CEUR Workshop Proceedings Сер. "CLiF 2016 — Proceedings of Corpus Linguistics Fest 2016" 2016. С. 1–6.
2. Салчак А. Я. Электронный корпус текстов тувинского языка // Новые исследования Тувы. 2012. № 3 (15). С. 110–114.
3. Leontiev N. The newspaper corpus of the yakut language // В сборнике: Сборник трудов международной конференции Turklang-2015. Tatarstan Academy of Sciences L. N. Gumilyov Eurasian National University Ministry of Education and Science of the Republic of Kazakhstan Kazan Federal University Institute of Philology and Intercultural Communication. 2015. С. 233–235.
4. Леонтьев Н. А., Протопопова В. Ф. Программное определение автора текста на якутском языке статистическим методом // В сборнике: Высокие технологии и инновации: векторы, проблемы и приоритеты. Сборник научных трудов по материалам I Международной научно-практической конференции. 2017. С. 40–44.
5. Леонтьев Н. А., Слепцов И. А. Идентификация текстового документа с помощью триграмм на материалах якутского языка // Вестник Северо-Восточного федерального университета им. М. К. Аммосова. 2015. № 4 (48). С. 45–50.
6. Леонтьев Н. А., Протопопова В. Ф. Электронный словарь имен собственных якутского языка // Форум молодых ученых. 2017. № 1 (5). С. 326–328.
7. Леонтьев Н. А., Торотоев Г. Г. Многопользовательская морфологическая разметка корпуса якутского языка // В сборнике: Электронная письменность народов российской федерации: опыт, проблемы и перспективы. Сборник материалов Международной научной конференции. 2017. С. 101–103.

Журнал «Современная наука: актуальные проблемы теории и практики»