УДК 519.769+ 81.322

Леонтьев Ньургун Анатольевич
ФГАОУ ВПО «Северо-восточный федеральный университет им. М.К. Аммосова»
Россия, Якутск[1]
Доцент каф. радиотехники и информационных технологий
Кандидат технических наук
E-Mail: leonza@mail.ru

Автоматическая коррекция букв в текстовом сообщении на якутском языке

Аннотация. Внедрение современных информационных технологий в жизнь дало новый толчок для общения, появились Интернет-форумы, чаты, социальные сети, дневники и другие средства сетевого общения. Отсутствия четких стандартов национальных алфавитов и кодировок привело к использовании в общении букв кириллического русского алфавита. Пользователи сети часто используют транслитерацию сообщений на якутском языке с помощью букв русского языка. Документов описывающие правила транслитерации с якутского языка на русский не существует, так как в письменности применялся либо русский язык либо якутский язык. В данной работе рассматриваются варианты транслитерации букв национального алфавита якутского языка в Интернет-форумах пользователями сети. Приводятся примеры слов написанных в транслитерации, проведен анализ совпадения словарных слов при транслитерации, а также используемой транслитерации букв якутского алфавита с помощью букв русского алфавита. Собраны варианты используемой транслитерации букв якутского языка в текстовых сообщениях. Автором впервые созданы правила замены и автоматической коррекции букв, написанных в транслитерационном написании. Создан скрипт на языке программирования PHP и проверена работа скрипта по автоматической коррекции букв в текстовом сообщении на якутском языке с помощью разработанных правил коррекции.

Ключевые слова: якутский язык; автоматическая коррекция букв; текстовые сообщения; транслитерация сообщений; Интернет-форумы; скрипт PHP.

Идентификационный номер статьи в журнале 96TVN314

[1] 677013, Республика Саха (Якутия), г.Якутск, ул.Белинского 58, СВФУ ФТИ РТиИТ, КФЕН ауд.614

Введение

Внедрение современных информационных технологий в жизнь дало новый толчок для общения, появились Интернет-форумы, чаты, социальные сети, дневники и другие средства сетевого общения. Сетевое общение дает много разных возможностей для публикаций, обладая при этом также функцией анонимности и мобильности. Использование программных средств, предназначенных для английского языка в среде использующих другой язык для общения проходить не без проблем. Одной из проблем было ограничение традиционных кодировок количеством отображаемых символов, а также отсутствием соответствующих шрифтов. Введение стандарта Unicode решило частично проблему национальных кодировок, желающие народы и народности могут подать заявку на включение своих кодировок в данный стандарт. В стандарте Unicode имеется символы якутского национального алфавита, в кодах 0400-04FF, принадлежащих кириллице, что описывается документом стандарта Unicode 6.3 на сайте http://www.unicode.org.

Вопросы транслитерации рассматриваются в разных документах, например существует ГОСТ 7.79-2000 (ИСО-95) межгосударственный стандарт «Правила транслитерации кирилловского письма латинским алфавитом», в нем приведены правила транслитерации на кириллическом алфавите, на русском, белорусском, украинском, болгарском и македонском языках, также в приложении А стандарта приведен список языков охваченный кирилловской письменностью, в их числе упомянут якутский язык. Но этот стандарт описывает правила транслитерации с кириллического на латинский алфавит. Документов описывающие правила транслитерации с якутского языка на русский не существует, так в письменности применялся либо русский язык либо якутский язык.

В разных странах и регионах были рассмотрены проблемы автоматической коррекции и транслитерации. В [1,2] приводятся работы по автоматической коррекции таджикского языка, в работе [3] показан подход к автоматической коррекции ошибок сочетаемости слов в текстах на естественном языке, в работах [4,5] обсуждаются транслитерация имен собственных. Транслитерация русских слов на другие иностранные языки приведены в работе [6,7]. Исследования по автоматической коррекции естественных языков ведутся различными коллективами, но с разными языками. Естественные языки имеют большие различия, из-за этого методы и подходы к автоматической коррекции сильно отличаются.

В Республике Саха (Якутия) сетевое общение имеет давние корни, имеются несколько популярных ресурсов для общения. Одним из популярных ресурсов является система форумов на сервере www.ykt.ru. Поддержка стандарта Unicode позволяет использовать якутские национальные буквы для сетевого общения. Так же имеется кнопки на форме ввода сообщения с помощью которых можно вводить якутские буквы, с использованием манипулятор «мышь». В мобильных приложениях существуют надстройки клавиатуры, которые позволяют вводит национальные буквы.

Автором был осуществлен обзор Интернет-форумов на якутском языке, были рассмотрены более полутора тысяч сообщений, в среднем получилось, что только 6,4% сообщений содержат буквы национального алфавита, остальные сообщения набраны в транслитерации или же не имеют в составе якутских букв. Средняя длина сообщений около десяти слов, что позволяет отнести большую часть сообщений к коротким типам сообщениям.

Отдельная задача, это определение языка текста, на котором записано сообщение. Имеются несколько видов решения, например через N-граммы [8], словарное определение [9]. В большой степени человек, определяя язык текста, ориентируется на словарь, а потом на правила словообразования в языке. При работе с текстом на русском языке, обычно данная проблема не возникает, так как она имеет отличия от текста на основе латинского алфавита,

так как кириллические символы имеют другой код страниц и код букв. Хотя раньше до внедрения стандарта Unicode существовало несколько кодовых страниц, такие как cp1251, OEM866, KOI8-R, cp437, ASCII, ГОСТ 19768-87. С внедрением Unicode вопросы совмещение в тексте разных кодировок отпали. Все виды кодировок и обработка текста на русском языке поддерживаются компаниями-разработчиками, языки народов России имеют меньшую поддержку на уровне компаний-разработчиков программного обеспечения, так коммерческое использование обработки текста на национальных языках не имеет большой финансовой перспективы из-за малого контингента.

Проблема ошибок в сообщениях на якутском языке на Интернет-форумах заключается в том, что многие пользователи Интернета имеют привычку набирать якутские национальные буквы, используя транслитерацию русских букв. Имеется несколько причин, во-первых отсутствие якутской раскладки клавиатуры на некоторых операционных системах, во-вторых отсутствие якутской раскладки клавиатуры, а также привычный, более скоростной набор. Слово получается в транслитерации, но при чтении другими пользователями понятен в какой-то мере, но бывают случаи, когда слово становится трудночитаемым или появляются совпадение с другими словами.

Среди остальных народов России и Азии идут споры о замене алфавита и системе транслитерации при использовании кириллического и латинского алфавита. Например, в казахском языке 42 знака и как уместить их в другом алфавите, в результате чего происходит игнорирование национальных знаков[10].

Таким образом, была поставлена цель автоматической коррекции национальных букв написанных транслитерацией в текстовых сообщениях на Интернет-форумах использующих якутский язык.

Анализ текстов

Для анализа совпадение якутских и русских слов были сравнено более 3 тысяч русских и якутских написанных в транслитерационной форме. Для этого было якутское слово было преобразовано в транслитерационную форму и сравнено с словами русского словаря.

Было найдено только одно совпадение при замене букв. Это якутское слово **үс** (три) – превращенное в слово **ус**. Использование этого слова в сложных формах совпадений уже не дает – **үстэх, үстэн, үһүс** (*с тремя, с трех, третий – якут.язык*)**.**

Имеется пять национальных букв якутского алфавита, их заменяют следующими транслитерационными символами. В таблице 1, приводятся якутские буквы и их транслитерационные аналоги.

Таблица 1

Якутские буквы и их транслитерационные аналоги
(составлено автором)

Буква	Аналог
Ү	у, Y (англ.), «Ұ», «ю»
ө	Е, 8
һ	Ь, h (англ.)
ҕ	Цифру «5»
ҥ	«н»

В случае замены получается из слова «бөтүүк» слово «бетуук», «б8туук», «бетYYk». Для некоторых слов получаются до десяти различных форм написания. В таблице 2, приводятся слова на якутском языке и их аналоги в транслитерационной записи.

Таблица 2

Слова на якутском языке и их аналоги в транслитерационной записи
(составлено автором)

№	*Якутское слово*	*Транслитерационный*
1	алҕас	ал5ас
2	анньыалаһар	анньыалаьар
3	быраҕар	быра5ар
4	бөтүүк	бетуук
5	түргэтэттэ	тургэтэттэ
6	хатааһылатар	хатааьылатар
7	үчүгэйкээн	учугэйкээн
8	үҥкүүлүүр	ункуулуур
9	үөрбүт	уербут
10	мөһөччүк	Меьееччук

Проверка текста и определение написанного языка осуществляет с помощью подпрограммы, использующий N-граммы с точностью 70-100%, что вполне достаточно для случая определения языка в текстовом сообщении, длиной от десяти слов. Использование слов из якутского и русского языка вперемешку усложняет определение языка текста из-за смешанной структуры. Особенности якутско-русского билингвизма порождают языковые конструкции с использование русских слов с якутскими окончаниями или якутских слов с русским склонением, что усложняет задачу при определении языка текстового сообщения.

Правила замены букв

По результатам анализа автором впервые были созданы правила замены букв транслитерационной записи, данные правила были разработаны автором для набора сообщений с сайта www.ykt.ru.

Буква «ь» и английская буква «h» заменяется на якутскую букву «h» исходя из-за следующих правил: в якутском языке обычно букв h произносится в контексте гласных букв, например в словах: «кыргыһыы», «мөһөөччүк», «куһаҕан». В якутском языке имеются диграфы «нь» и «дь», где используется буква «ь», такие случаи надо исключать из замены.

Буква «у» заменяется на букву «ү», по следующим правилам: буква «ү», пишется, если предыдущем слоге «ү» или «э» или «үө», т.е. используем правило гармонии гласных. Остальные случаи замены букв происходят по словарю. Исключается слово «түү», так как есть слово «туу».

Буква «е» и цифра «8» заменяются на букву «ө», по следующим правилам: отсутствие в якутском языке буквы «е» позволяет проводить замены во всех словах, кроме русских. Принадлежность слова к русскому словарю, проверяется через базу данных русского словаря. А цифра «8» заменяется в том случае, если его окружают буквы.

Буква «ю», «Ұ» заменяется на букву «ү», так как отсутствие в якутском языке этих букв позволяет проводить замены во всех словах, кроме русских.

Цифра «5» заменяется на букву «ҕ», по следующим правилам: происходит проверка на контекст букв. Замена не производится, если цифра «5» написана первой буквой, последней буквой или в контексте чисел.

Буква «н» заменяется на букву «ҥ», по следующим правилам: в случае пары «нк» заменяется на «ҥк». Остальные случаи проверяются по словарю.

Английские буквы «Y» и «h» заменяются на буквы «ү» и «һ» в случае, если их в слове окружают буквы не английского алфавита.

Заключение

Для замены букв написанных транслитерационной записи для букв «ө», «һ», «ҕ» в большинстве случаев достаточно простых правил, приведенных выше.

Для замены букв «ү» и «ҥ» необходимо применять словарный поиск, так как они не обладают четким критерием по правилам грамматики. Использование правила гармонии гласных покрывает только часть слов [11], которые являются многосложными, то есть состоят из трех и более слогов.

Таким образом, автоматическая коррекция букв в словах, написанных на якутском языке с помощью транслитерационной записи возможна. Используя простые правила замены и словарный поиск можно произвести автоматическую коррекцию в текстовых сообщениях. Сложность словарного поиска заключается в его объеме и большом времени на поиск в базе данных. Для поиска всех возможных вариаций написания слов и букв необходимо набрать материалы на Интернет-ресурсах на якутском языке, где происходит ежедневное общение пользователей и выставление новостей. При словарном поиске необходимо учитывать возможные вариации написания слова, что увеличивает критерии поиска, что отрицательно сказывается на быстродействии.

Размер словаря имеет большое значение при словарном поиске, в электронном виде имеются словари до 25 тыс. слов, что является достаточно большим словарем. Использование устаревших слов в транслитерационной записи не позволяет произвести автоматическую словарную коррекцию, также использование иностранных слов тоже усложняет автоматическую коррекцию текстового сообщения.

Автором была написана программа на языке PHP для автоматической коррекции букв для сообщений на якутском языке. Для работы программы необходим входной текст в стандарте Unicode, в кодировке UTF-8. В результате работы программы получается выходной текст в кодировке UTF-8. Для обработки текстовых сообщений в кодировке UTF-8 использовалась библиотека Multibyte String Function.

В ходе тестирования программа справилась со 90% текстовыми сообщениями, где существовали транслитерационные замены. Текстовые сообщения, в количество более 300 сообщений, были взяты с Интернет-форумов, где происходит общение на якутском языке. Необходимо производить обновление возможных правил замены, так как появляются попытки изменения правил транслитерации, что усложняет возможную коррекцию слов на якутском языке.

Для развития метода автоматической коррекции якутского языка необходимо развивать методы N-грамм, методы использования национального корпуса языка, методы словарного тезауруса. Развитие этих методов даст большой толчок для автоматической коррекции слов и предложений на якутском языке.

ЛИТЕРАТУРА

1. Усманов З.Д., Эвазов Х.А. Компьютерная коррекция таджикского текста, набранного без использования специфических букв // Доклады Академии наук Республики Таджикистан, том 54, №1, 2011, стр.23-26.
2. Усманов З.Д., Гращенко Л.А., Фомин А.Ю. Информационные основы автоматизированной таджикско-персидской транслитерации // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук, №1, 2008, стр.20-26.
3. Азимов А.Е., Большакова Е.И. Подход к автоматической коррекции ошибок сочетаемости слов в текстах на естественном языке // Новые информационные технологии в автоматизированных системах, №14, 2011, стр. 78-91
4. Сулейманова Р.А. Транслитерация башкирских личных имен на русский язык. // Сборник трудов конференции Городские башкиры: прошлое, настоящее, будущее Бирск, 2008, стр., стр.352-354
5. Караджа Бирсен, Актуальные проблемы перевода художественного текста с русского на турецкий язык (о проблеме транслитерации и транскрипции собственных имен) // Вестник Московского университета, серия 9: Филология, №2, 2007, стр.130-135
6. Кабакчи В.В., Юзефович Н.Г. Транслитерация русизмов в англоязычном описании русской культуры (к столетию поисков системы транслитерации русизмов) // Социальные и гуманитарные науки на Дальнем Востоке, №3, 2007, стр.115-124
7. Чэлич Ж.М., Проблема реализации транслитерации с русской кириллицы на хорватскую латиницу. // Язык. Словесность. Культура, №4, 2012, стр.84-97
8. Леонтьев Н.А. Распознавание языка текстовых сообщений с помощью биграмм на материалах якутского языка // Современное состояние естественных и технических наук. М: "Спутник+". 2014, XIV, стр.88-91
9. Леонтьев Н.А. Словарное определение якутского языка в текстовом сообщении // Научная перспектива.№2 (48)/ 2014. стр. 97-98.
10. Нужно ли менять казахский алфавит? http://online.zakon.kz/Document/?doc_id=30075826 (Дата обращения: 31.03.2014)
11. Леонтьев Н.А. Автоматическое исправление ошибок в якутском языке с помощью гармонии гласных // Сборник материалов XVII Международной научно-практической конференции, Новосибирск, 2014. стр. 25-27

Рецензент: Попов Василий Иванович, к.ф.-м.н., с.н.с. Арктического инновационного центра Северо-Восточного федерального университета им. М. К. Аммосова.

Nurgun Leontiev
M. K. Ammosov North-Eastern Federal University
Russia, Yakutsk
E-Mail: leonza@mail.ru

Automatic correction of letters in a text message in the Yakut language

Abstract. Development of modern information technologies gave a new impetus for communication, appeared online forums, chat rooms, social networks, blogs and other social networking tools. Lack of clear standards of national alphabets and character sets has led to use in communicating Cyrillic Russian alphabet letters. Net users often use transliteration posts in the Yakut language using letters of the Russian language. Documents describing the rules of transliteration from Russian to Yakut language does not exist, so as in writing used the Russian or the Yakut language. In this paper discusses the options for transliteration national letters Yakut language Internet forums. In paper are examples of words written in transliteration, the analysis matches dictionary words in transliteration and transliteration used Yakut alphabet letters using the letters of the Russian alphabet. In paper shown varieties used transliteration letters Yakut language in text messages. Author first created the replacement rules and automatic correction of letters written in transliteration writing. Created the script in PHP programming language and tested the script for automatic correction of letters in a text message in the Yakut language developed using adjustment rules.

Keywords: sakha (Yakut) language; automatic correction of letters; text messages; transliteration of messages; Internet forums; script PHP; computational linguistics.

Identification number of article 96TVN314

REFERENCES

1. Usmanov Z.D., Jevazov H.A. Komp'juternaja korrekcija tadzhikskogo teksta, nabrannogo bez ispol'zovanija specificheskih bukv // Doklady Akademii nauk Respubliki Tadzhikistan, tom 54, №1, 2011, str.23-26.
2. Usmanov Z.D., Grashhenko L.A., Fomin A.Ju. Informacionnye osnovy avtomatizirovannoj tadzhiksko-persidskoj transliteracii // Izvestija Akademii nauk Respubliki Tadzhikistan. Otdelenie fiziko-matematicheskih, himicheskih, geologicheskih i tehnicheskih nauk, №1, 2008, str.20-26.
3. Azimov A.E., Bol'shakova E.I. Podhod k avtomaticheskoj korrekcii oshibok sochetaemosti slov v tekstah na estestvennom jazyke // Novye informacionnye tehnologii v avtomatizirovannyh sistemah, №14, 2011, str. 78-91
4. Sulejmanova R.A. Transliteracija bashkirskih lichnyh imen na russkij jazyk. // Sbornik trudov konferencii Gorodskie bashkiry: proshloe, nastojashhee, budushhee Birsk, 2008, str., str.352-354
5. Karadzha Birsen, Aktual'nye problemy perevoda hudozhestvennogo teksta s russkogo na tureckij jazyk (o probleme transliteracii i transkripcii sobstvennyh imen) // Vestnik Moskovskogo universiteta, serija 9: Filologija, №2, 2007, str.130-135
6. Kabakchi V.V., Juzefovich N.G. Transliteracija rusizmov v anglojazychnom opisanii russkoj kul'tury (k stoletiju poiskov sistemy transliteracii rusizmov) // Social'nye i gumanitarnye nauki na Dal'nem Vostoke, №3, 2007, str.115-124
7. Chjelich Zh.M., Problema realizacii transliteracii s russkoj kirillicy na horvatskuju latinicu. // Jazyk. Slovesnost'. Kul'tura, №4, 2012, str.84-97
8. Leont'ev N.A. Raspoznavanie jazyka tekstovyh soobshhenij s pomoshh'ju bigramm na materialah jakutskogo jazyka // Sovremennoe sostojanie estestvennyh i tehnicheskih nauk. M: "Sputnik+". 2014, XIV, str.88-91
9. Leont'ev N.A. Slovarnoe opredelenie jakutskogo jazyka v tekstovom soobshhenii // Nauchnaja perspektiva.№2 (48)/ 2014. str. 97-98.
10. Nuzhno li menjat' kazahskij alfavit? http://online.zakon.kz/Document/?doc_id=30075826 (Data obrashhenija: 31.03.2014)
11. Leont'ev N.A. Avtomaticheskoe ispravlenie oshibok v jakutskom jazyke s pomoshh'ju garmonii glasnyh // Sbornik materialov XVII Mezhdunarodnoj nauchno-prakticheskoj konferencii, Novosibirsk, 2014. str. 25-27