Два метода выявления русских заимствований в якутских текстах

Н. Кортегосо Виссио, В. П. Захаров

***Аннотация*— В этой статье рассматриваются два метода выделения русскоязычные заимствований в якутских текстах. Под русскоязычным заимствованием понимаются лексические элементы, корни которых не адаптированы к якутской фонетике и пишутся как в исходном языке. Исходя из того, что большинство заимствований в якутских текстах происходит из русского языка, предполагается, что они имеют определенную форму, по которой их можно отличить от якутских словоформ. Первый метод опирается на правилах. В нем реализован алгоритм, выявляющий сочетания букв, чуждые якутскому языку. Второй метод применяет статистический подход к моделированию сочетаний якутских и русских букв.**

Эффективность обоих методов извлечения заимствований сравнивается с результатами ручного выделения носителями русского языка в 6 якутских текстах. Данная работа является продолжением статьи [1].

***Ключевые слова*— автоматическая обработка текстов, выделение русских заимствований, якутский язык, n-грамм.**

Введение

Якутский язык начал заимствовать русские слова при первых контактах в XVII веке. Это тенденция получила новый импульс после Октябрьской революции и продолжается до настоящего времени. Русские заимствования в якутском языке были и остаются предметом филологических исследований. Среди классических исследований можно назвать работу П.А. Слепцова [2] и других исследователей, например, Л.Н. Харитонова [3]. В этих работах разбираются периоды, когда русские заимствования входили в состав письменного якутского языка. В основном, первые заимствования, пришедшие в язык, были приспособлены к якутской фонологии, но усиление двуязычия среди якутов привело к тому, что более поздние термины стали употребляться с их русским написанием. С 1939 года якутский язык пишется вариантом кириллицы, включающей в себя все символы русского алфавита, плюс 5 специальных букв { ү, ө, ҕ, ҥ, h }. Символы { е, я, ю, ё, в, г, ж, з, ф, ц, ш, щ, ъ } используются только в лексических элементах, заимствованных из русского языка. Что касается типа заимствованных слов, то на первом месте стоят существительные, за ними следуют прилагательные.

Более современным можно назвать эмпирическое исследование, проведенное Н.М. Васильевой [4], где она изучает адаптацию гласных и применение гармонии гласных в адаптациях русских слов двуязычными русско-якутскими носителями. Васильева отмечает, что написание русских заимствований, которые издавна широко употребляются в разговорной речи, было адаптировано к якутской фонологии [4, с. 166-167]. Например: *остуол* «стол», *куорат* «город», *сокуон* «закон», *норуот* «народ», в то время как лексические элементы, которые представляют общественно-политические, научно-технические понятия, сохраняют русскую форму. Например: «архитектура», «неолит», «материализм».

Также выделяются заимствования, которые часто встречаются в текстах как в русском, так и в якутском написании. К этой группе относится заимствования, до сих пор не разрешенные в разговорной речи, как например, *биисинэс* «бизнес», *сийиэс* «съезд», *эрэнгиэн* «рентген», имена собственные и некоторые географические названия, например, *Дьоппуон* «Япония», *Эмиэрикэ* «Америка», *Уйбаныап* «Иванов», *Маарыйа* «Мария» и т. д.

Васильева также отмечает следующее: «в географических наименованиях, в мужских фамилиях в форме полного прилагательного и в названиях городов на *-ск* в конце пишется *-ай* или *-эй*, например, *Новай Гвинея*, *Пекарскай*, *Горькай*, *Минскэй*, а также в заимствованных прилагательных, фиксируемых в русской форме, окончания передаются через *-ай*, *-эй*: *этиловай испиир*, *физическэй география*» [4, с. 166-167].

Написание русских заимствований является актуальной областью лексических исследований якутского языка. Лексикографы, занимающиеся вопросом эволюции использования заимствованных слов в прессе и в Интернете, рассматривают большое количество письменного материала. В этой статье предлагаются и оцениваются два подхода для автоматического выделения заимствований в якутских текстах.

Результаты этих подходов сравниваются с ручной разметкой. Под заимствованием понимаются лексические элементы, корни которых не адаптированы к якутской фонетике и пишутся так, как в исходном языке.

Остальная часть статьи включает обзор статистического подхода, использованного в

Статья получена 1 октября 2022.

Кортегосо Виссио Николас, магистр в компьютерной и прикладной лингвистике, Санкт-Петербургский государственный университет, ORCID 0000-0003-1683-7270 (st082534@student.spbu.ru).

Захаров Виктор Павлович, канд. филол. наук, Санкт-Петербургский государственный университет, доцент кафедры математической лингвистики, ORCID 0000-00030522-7469 (v.zakharov@spbu.ru).

предыдущей статье, который повторно тестируется здесь. Затем следует короткое описание закона гармонии гласных на якутском языке и то, как её можно применять в виде простого алгоритма. Далее раскрывается метод обработки корпуса, который применялся при обучении статистического метода. И наконец, сравниваются результаты выделения заимствований, полученные с помощью метода, основанного на правилах, и статистического метода.

Предыдущие работы

Стохастический метод автоматического выделения русских заимствований был уже предложен в статье «Выделение русских заимствований в якутских текстах» [1].

Там поясняется, что классификацию русских заимствования в якутских текстах можно рассматривать в рамках задачи идентификации языков. Один из наиболее распространённых методов для идентификации языка опирается на N-граммы. Основная идея состоит в том, что каждый язык в его письменной форме имеет уникальные сочетания букв, по которым его можно идентифицировать. Данное сочетание букв можно смоделировать при помощи N-грамм, то есть последовательность n элементов, где вероятность встретить элемент n зависит от n-1 предыдущих элементов. Если модель учитывает только один предыдущий элемент, то она называется биграммной (анализируются биграммы), если два - триграммной (3-граммы) и т. д. Эти N-граммы имеют разные вероятности в разных языках. Например, вероятность биграммы «аы», то есть вероятность встретить букву «а» после «ы» в русском гораздо меньше, чем в якутском, где «аы» является обычным дифтонгом.

С целью идентификации языков создаются N-граммные модели для каждого языка, который необходимо определить. Эти модели содержат вероятность последовательностей n букв. Когда N-граммные модели созданы, можно их использовать для определения языка.

Чтобы отличить русские заимствовании от якутских словоформ, требуется создать две модели N-грамм, одна для русских заимствований и одна для якутских словоформ. Если для создания моделей использовались 3-граммы, словоформа или фрагмент текста, язык которых следует определить, разделяются также на 3-граммы. Например, словоформы «город» разделяются на следующие элементы, где ^ и _ являются идентификаторами начала, а # показывает окончание словоформы:

*_*г, *го, гор, оро, род, од#*

Затем извлекаются вероятности для всех 3-грамм по модели заимствований и по модели якутских словоформ, от каждой из них берется логарифм и эти значения суммируются. Например:

по модели якутских словоформ:

$$-8.08+(-9.88)+(-10.29)+(-6.92)+(-10.02)+(-13.55)=-58.73$$

по модели русских заимствований:

$$-5.41+(-7.14)+(-7.51)+(-6.95)+(-7.51)+(-8.86)=-43.38$$

Чтобы решить, является ли словоформа «город» заимствованием, сравниваются полученные результаты и выбирается большее значение. В этом случае, -43.38 больше, чем -58.73, поэтому словоформа «город» выделяется как заимствование.

В процессе классификации словоформы *куорот* (город) получится следующая сумма логарифмов 3-грамм { ^_к, _ку, куо, уор, оро, рот, от# } для каждой модели:

по модели якутских словоформ:

$$-4.44+(-6.32)+(-7.32)+(-6.75)+(-6.92)+(-10.50)+(-9.10)=-51.35$$

по модели русских заимствований:

$$-4.61+(-6.93)+(-9.68)+(-8.99)+(-6.95)+(-7.78)+(-9.21)=-54.15$$

Теперь -54,15 меньше, чем -51,35, а значит словоформа *куорот* не определяется как заимствование.

В процессе классификации можно применить априорное значение, то есть, ожидаемую вероятность встретить или заимствование, или якутские словоформы. Например, если ожидается (наблюдается), что девять из десяти словоформ в якутских текстах является заимствованием, то при классификации словоформы «город» можно взять логарифм от вероятности 0.1 для якутских словоформ (-2.30) для якутских словоформ, от вероятности 0.9 для заимствований (-0.10) и тогда мы получим:.

по модели якутских словоформ:

$$-2.30+(-8.08)+(-9.88)+(-10.29)+(-6.92)+(-10.02)+(-13.55)=-61.03$$

по модели русских заимствований:

$$-0.10+(-5.41)+(-7.14)+(-7.51)+(-6.95)+(-7.51)+(-8.86)=-43.48$$

При применении априорных значений словоформа «город» классифицируется как заимствование (-43.48 > -61.03), но расстояние между двумя показателями увеличилось (раньше было -43.38 > -58.73). Использование априорных значений придает модели некоторую гибкость для создания классификаций. Манипуляции с априорными значениями влияют на точность и полноту модели 3-грамм. Более высокая априорная вероятность заимствований повысит точность классификации за счет ее полноты и наоборот.

Модели N-грамм обучаются на основе корпуса. За подробностями о том, как вычисляются вероятности N-грамм, читатель отсылается к статье Canvar и Trenkle [5]. В предыдущей статье были рассмотрены проблемы, по которым невозможно создавать профили N-грамм для выделения заимствований напрямую из русских и якутских корпусов. В целом эти сложности можно свести к двум причинам:

1. Заимствования в якутских текстах утрачивают грамматические особенности, которыми они обладают в русском языке, и подчиняются правилам аффиксации якутского языка. Таким образом, заимствования, сохраняющие русскую орфографию в корне, могут встречаться с прикрепленным аффиксами в якутских текстах. Например: Москваҕа «в Москве», космонавтар, «космонавты». Следовательно, модель, обученная на русском тексте, теряла бы способность выделить заимствовании с прикрепленными суффиксами.
2. Якутские тексты содержат некоторое количество русских заимствований. Наличие в

них лексических единиц, сохраняющих русскую орфографию, могло внести шум в модель якутских N-грамм и ухудшить способность выделения заимствований. В работе, опубликованной в 1947 г., Харитонов подсчитал, что доля русских заимствований в газетах (включая иностранные слова) составляла до 41 процентов. Согласно Харитонову, в стихах этот показатель уменьшился до 6 процентов [3, с. 24].

В качестве попытки обойти эти две проблемы была специально создана обучающая выборка на основе корпуса Википедии на якутском языке (108715 уникальных словоформ) [6]. В ней словоформы были размечены либо как «якутские», либо как «заимствования». Процесс разметки обучающей выборки реализован на основе фильтрации специальных символов, которые встречаются только в заимствованиях { е, я, ю, ё, в, г, ж, з, ф, ц, ш, щ, ъ }. Еще 5 дополнительных фильтров применялись для разметки русских заимствований в корпусе. Данные фильтры основаны на различиях, которые Харитонов описывает между якутскими и русскими словоформами [3, с. 61-65]:

1. в начале якутского слова не встречается более одного согласного;
2. в конце якутского слова не встречается более одного согласного (исключение составляют сочетания «рт» и «лт»);
3. в середине якутского слова не встречается более двух согласных рядом;
4. в начале якутского слова никогда не встречаются согласные { ҕ, й, ҥ, р, ль };
5. в конце якутского слова не встречаются { б, г, ҕ, д, дь, ль, нь, h, ч }.

Если словоформа не соответствует пунктам 1-5, она помечается как заимствование. Все словоформы из корпуса, которые не были «уловлены» фильтрами, размечены как якутские.

После применения фильтров 31695 словоформ (29%) были помечены как заимствования. На основе этого размеченного корпуса были обучены модели 3-грамм. Превалирование 3-грамм над другими типами N-грамм обосновано тем, что 3-граммы могут моделировать все вышеперечисленные фильтры. Например, фильтр 3 относится к последовательности из трех согласных. Ожидается, что обученные модели на этой выборке будут способны статистически отразить разницу между якутскими словоформами и русскими заимствованиями. Следует отметить, что 3-граммы в меньшей степени приспособлены моделировать явление гармонии гласных, чему посвящен следующий раздел.

Гармония гласных

Наиболее отличительной чертой якутской фонологии по сравнению с русской является гармония гласных. В то время как в русских словоформах допускается любое сочетание гласных, закон гармонии гласных в якутском языке требует, чтобы гласные подчинялись определенным правилам. В якутских словоформах наблюдаются сильные закономерности сочетания и последовательности гласных. В результате негармонические формы являются хорошими кандидатами в заимствования. Далее описывается закон гармонии гласных, описанный Харитоновым [3, 59-61].

Гласные звуки якутского делятся по своему произношению на следующие основные группы:
задние гласные { а, ы, о, у } и передние гласные { э, и, ө, ү }, с одной стороны, и широкие гласные { а, э, о, ө } и узкие гласные { ы, и, у, ү }, с другой.

Произношение гласных заднего и переднего ряда различается в зависимости от положения языка во рту. В то время как при произнесении гласных заднего ряда язык оттягивается назад, произношение гласных переднего ряда требует, чтобы язык выдвигался вперед. При этом образуются следующие пары (задние | передние): { а|э, ы|и, о|ө, у|ү }.

Произношение широких и узких гласных различается в зависимости от ширины открывания рта. У первых рот открывается шире, у вторых рот открывается гораздо слабее. При этом образуются следующие пары (широкие | узкие): { а|ы, э|и, о|у, ө|ү }.

Третье различие связано с ролью губ в их произношении: огубленные и неогубленные гласные. Неогубленные гласные – это { а, ы, э, и }, а огубленные – { о, у, ө, ү }.

В якутском языке существуют еще 4 дифтонга. Дифтонги делятся на задние { ыа, уо } и передние { иэ, үө }, а также на нелабиализированные { ыа, иэ } и лабиализированные { уо, үө }.

Согласно закону гармонии гласных якутского языка в одной словоформе могут сочетаться друг с другом только определенные гласные. Этот закон может быть выражен в следующих двух основных наблюдениях:

1. если в первом слоге словоформы есть гласная заднего ряда, то и в следующих слогах должны быть гласные заднего ряда. И, в свою очередь, если в первом слоге слова стоит гласная переднего ряда, то за ним могут следовать только гласные переднего ряда. Таким образом, словоформы на якутском языке состоят только из гласных заднего ряда или только из гласных переднего ряда;
2. за широкой гласной может следовать только широкая гласная или соответствующая ей узкая гласная. За узкой гласной может следовать только узкая гласная или соответствующая ей широкая гласная.

Это наблюдение можно проиллюстрировать в таблице 1:

Таблица 1. Гармония гласных (Харитонов [3: 60])

если в первом слоге есть:	то в следующем слоге может быть:	если в первом слоге есть:	то в следующем слоге может быть:
а	а, ы, ыа	ы	ы, а, ыа
э	э, и, иэ	и	и, э, иэ
о	о, у, уо	у	у, а, уо
ө	ө, ү, үө	ү	ү, э, үө

Поскольку закон гармонии гласных в якутском языке очень регулярен, его легко смоделировать с помощью простого алгоритма.

Подготовка обучающей выборки

Как упоминалось в разделе «Предыдущая работа», обучающая выборка состоит из 108715 словоформ, из которых 31695 помечены как заимствовании. В таблице 2 показано количество заимствований, обнаруженных каждым фильтром, и их доля в общем числе заимствований, если они применяются по отдельности.

Таблица 2. Количество заимствований, обнаруженных фильтров

фильтр	пояснение	кол-во заимст.	доля	примеры
содержит специальные символы	слово содержит символы, встречающиеся только в заимствованиях	28166	0.89	Россия, Иван, Мария, колхоз
начинается с двух согласных	в начале якутского слова не встречается более одного согласного	5683	0.18	спорт, Дмитрий, группа, драма
содержит кластеры более 2 гласных	в середине якутского слова не встречается более двух согласных рядом	5069	0.16	мусульманскай, донской, Татарстан, абстрактнай, курска
оканчивается недопустимой буквой	в конце якутского слова не встречаются { б, г, ҕ, д, дь, ль, нь, h, ч }	1646	0.05	Араб, Самарканд, Бог, Хань
оканчивается двумя согласными	в конце якутского слова не встречается более одного согласного (исключение сочетания «рт» и «лт»)	968	0.03	Маркс, банк, митинг, класс
начинается с недопустимой буквы	в начале якутского слова никогда не встречаются согласные { ҕ, й, ҥ, р, ль }	105	0.003	йорк, йунус, йогурт

При отдельном применении фильтра гармонии гласных он способен выделить 26560 словоформ как заимствования. Если его применить с оставшимися фильтрами, то количество обнаруженных заимствований вырастет с 31695 до 38326, то есть число заимствований увеличится на 6631, или на 21%. Словоформы, такие как «кандидат», «радио», «курорт», «поэма» являются примерами заимствований и выделены фильтром гармонии гласных, которые были пропущены остальными фильтрами.

Рисунок 1 показывает накопленную сумму заимствований, когда фильтры применяются каскадно.

На первые два фильтра приходится 95 процентов всех заимствований.

На основе этой новой размеченной выборки модель 3-грамм обучается для обнаружения заимствований (далее статистический метод). Также те 7 фильтров, которые были применены для создания обучающей выборки, были повторно использованы в качестве алгоритма выделения заимствований. Если хотя бы одна словоформа попадает в один из фильтров, она помечается как заимствование. Далее этот метод называется методом, основанным на правилах.

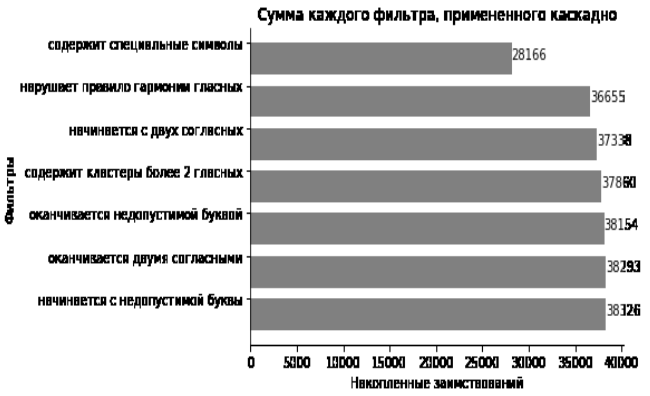

Рисунок 1. Сумма каждого фильтра, примененного каскадно

В следующем разделе и статистический метод, и метод, основанный на правилах, применяется на текстах на якутском языке.

Эксперименты

Эксперимент проведен на 6 статьях из якутоязычной газеты «Саха Сирэ» (см. таблица 3). В качестве критерия отбора текстов был установлен объем от 700 до 800 токенов. Первые 3 статьи извлечены из № 13, (апрель 2021 года), а остальные относятся к №1 (январь 2021 года). Текст №1 - это тот же текст, который использовался в эксперименте в предыдущей статье. Все тексты доступны по ссылке [7].

В каждом тексте применяются три метода выделения заимствований. Первой метод – это ручная разметка. Для этого мы попросили 6 русскоязычных (без знания якутского языка) прочитать текст и отметить все слова, корень которых они могут распознать. Таким образом, выделенная словоформа может включить в себя суффиксы, например, как в «Канадаҕа». Второй и третий метод – это статистический метод и метод, основанный на правилах, описанные в предыдущем разделе. Результаты показаны в таблицах.

Таблица 3. Список статей

№	Название статьи	Русский перевод
1	Арктика баайа норуот туһатыгар тахсыа дуо?	Пойдет ли богатство Арктики на пользу народу?
2	Ураты эйгэлээх “уолан” гимназия	Гимназия “Сыктывкар" с особым статусом
3	Дьиэҕэ олорон үлэ	Личная жизнь и работа
4	Саҥа дьыллааҕы сүпсүлгэн	Новогодняя тревога
5	Биһиги ытыктыыр салайааччыбыт – Иван Сивцев	Наш уважаемый руководитель – Иван Сивцев
6	Тыйаатыр иһигэр тыйаатыры көрдүбүт	Внутри театра мы увидели театр

Таблицы 4, 7, 10, 13, 16 и 19 содержат общие

показатели каждого текста. Отмеченные словоформы представлены в виде таблиц 5, 8, 11, 14, 17 и 20, где они распределены по типам словоформ (столбец «Категория»). Если участник эксперимента выделил словоформу, не являющуюся заимствованием, она помещается в категорию «Ошибочная». Если словоформа встречается в тексте больше одного раза, то количество повторений указывается в скобках. Столбец «Количество» показывает, сколько выделено словоформ по каждой категории, а столбец «процент» её долю в общем количестве.

Ручная разметка являются основание (baseline) для сравнения двух других автоматических методов. Процедура автоматического выделения заимствований применяется аналогично для всех 6 статей. Процесс токенизации выделяет словоформы по пробелам, пропуская знаки препинания. Каждая словоформа обрабатывается отдельным методом, основанным на правилах и статистических методах. В дальнейшем результаты автоматического выделения заимствований сравниваются с ручной разметкой.

В таблицах 6, 9, 12, 15, 18 и 21 показана разница между заимствованиями, выделенными человеком и заимствованиями, обнаруженными обоими автоматическими методами.

С одной стороны, ожидается, что метод, основанный на правилах, должен показать более высокую точность. То есть, если характеристики якутской словоформы, описанные Харитоновым, соблюдаются, тогда не должно появиться ложноположительных результатов. Но это также подразумевает, что метод, основанный на правилах, будет пропускать заимствовании, когда словоформы не противоречат требованиям, определенным в фильтрах. Это приведет к увеличению ложноотрицательных результатов. С другой стороны, несмотря на то, что статистический метод, вероятно, покажет более низкую точность, чем метод, основанный на правилах, ожидается, что он будет иметь лучший показатель полноты, т. е. будет иметь меньше ложноотрицательных результатов.

В качестве априорного значения статистического метода используется процент заимствований, выделенных вручную. Например, если человек выделил 75 заимствований в тексте 1, содержащем 713 словоформ, тогда процент заимствований равен 11% (75/713). Таким образом, при выделении заимствований на данном тексте, априорная вероятность для обнаружения заимствований будет 0.1, а для якутской словоформы 0.9.

Текст 1. Пойдет ли богатство Арктики на пользу народу?

Таблица 4. Общие показатели текста

Токены	755
Словоформы	713
Уникальные словоформы	493
Количество человеческих разметок	75
Процент вручную выделенных заимствований на тексте	11%

Таблица 5. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	26	35%	абориген, арктика, арктикаҕа (2), геолог, геолога (2), депутатскайга, империятын, инвестиция, инвестордар (2), кураторынан, металлургияҕа, миллиард, недра, протекционизм, ртуть, рудник (2), рудниктар, субъекка, субъект, субъектарга, сырье, фабрика,
прилагательные	16	22%	атомнай, геологическай (2), корпоративнай, материальнай (2), регистрациялаах, социальнай (2), стратегическай (2), транснациональнай, федеральнай (2), экологическай (2)
ФИО и имена собственные	29	39%	Валентин, Гаврил, Главсевморпуть, Госком, Госкомгеологияҕа, Дальстрой (2), Дмитриевич (3), Ефимов, Индигирзолото, Кириллин (3), Куларзолото, Магадан, Марианна, Николай (5), Тыртыкова, Цареградскай, Цветмет, Юрий, Янзолото
акронимы	3	4%	НКВД, СГУ, ССРС
ошибочные	1	1%	баран

Таблица 6. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	баран, Госком, Кириллин, Магадан	-
3-грамм	баран, Кириллин	гааска, Канадаҕа, онтон, Томпо

При ручной классификации участник эксперимента выделил якутскую модальную частицу *баран* из-за ее омонимии с русским существительным. Как и ожидалось, метод, основанный на правилах, не имеет ложноположительных результатов. Метод не обнаружил ни одного заимствования, которое не было выделено вручную. Статистический метод ошибся при классификации *гааска* (газ), *онтон* (временной союз *затем, потом, после*) и словоформы *Томпо*, указывающей на реку в Якутии. Статистический метод выделил словоформу *Канадаҕа* (в Канаде), которая была пропущена при ручной разметке.

Текст 2. Гимназия «Сыктывкар» с особым статусом

Таблица 7. Общие показатели текста

Токены	774
Словоформы	762
Уникальные словоформы	547
Количество человеческих разметок	97
Процент вручную выделенных заимствований на тексте	12%

Таблица 8. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	40	44%	академиятын (2), гимназия (12), гимназиябыт (3), гимназиялара,

			гимназияларыгар, гимназияны, гимназиятыгар, гимназияҕа, лабораториялара, лауреатынан (2), лидер, логикалара, математика (2), математикаҕа , моделирование, наукаларын, номинациятыгар, олимпиадаларга, робототехника, робототехникаҕа, технологиялары, управлениетын, философиятын, этнопедагогикаҕа
Прилага-тельные	15	17%	ассоциированнай, инновационнай, интеллектуальнай (2), лазернай, методическай, национальнай, патриотическай, педагогическай (2), практическай, спортивнай, федеральнай, цифровой (2)
ФИО и имена собственные	34	38%	Алексеев, Аммосов, Антипин, Афанасий, Борислав, Валентин (2), Василий, Васильевич, Ворлдскилл, Ворлдскиллс (2), Гена, Гран-При, Захарова, Иванович, Кирилловы, Колесов, Максим, Мигалкин, Михаил, Николаев, Петрова, Петровна, Розалия, Руслан (2), Татаринов (2), Федор (2), Федортатаринов, Филиппов (2)
акронимы	1	1%	РФ
ошибочные	0	0%	-

Таблица 9. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	Руслан	грант, гранын, креатосфераны, МүрүГИС, политиэхиньиичэскэй, станокка, телеустуудьуйатын, ХИФУ
3-грамм	-	грант, гранын, инники, креатосфераны, мин, политиэхиньиичэскэй, станокка, телеустуудьуйатын, ХИФУ

Метод, основанный на правилах, пропустил только имя *Руслан*. И метод, основанный на правилах, и статистический метод выделили практически такие же заимствовании, хотя статистический метод неправильно классифицировал *инники* (передний) и *мин* (я). Здесь интересно, как якутские словоформы сочетаются с заимствованными префиксами «поли» и «теле»: *политиэхиньиичэскэй* (политехнический) и *телеустуудьуйатын* (телестудия).

Текст 3. Личная жизнь и работа

Таблица 10. Общие показатели текста

Токены	773
Словоформы	759
Уникальные словоформы	523
Количество человеческих разметок	39
Процент вручную выделенных заимствований на тексте	5%

Таблица 11. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	16	41%	доку, ипотека, коронавирус, коронавирустан, офис (2), офискар (2), пресс, психолог, технология, фриланс (4), фрилансер, юристар
прилагательные	9	23%	больничнай (2), гражданскай, дистанционнай, канцелярскай, правовой, психологическай, социальнай, спортивнай
ФИО и имена собственные	5	13%	Байгожаева, Марианна, Тыртыкова, Яна (2)
акронимы	5	13%	БАД, НДФЛ (2), км (2)
ошибочные	4	10%	баран, да (3)

Таблица 12. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	баран, да	аэропуордун, бассейннға, ГПХ, кодексатын, принтергиттэн, сервергэ, фитнескэ
3-грамм	баран, да	аэропуордун, бассейннға, ГПХ, кодексатын, принтергиттэн, сервергэ, фитнескэ

Оба метода достигли аналогичных результатов. Якутская частица да не была выделена как заимствование, как и произошло в случае ручных разметок. Здесь снова появилась модальная частица баран. Следует отметить, что человеческая разметка пропустила словоформу бассейннға (в бассейн), которая должна быть знакома для носителей русского языка.

Текст 4. Новогодняя тревога

Таблица 13. Общие показатели текста

Токены	775
Словоформы	775
Уникальные словоформы	552
Количество человеческих разметок	19
Процент вручную выделенных заимствований на тексте	2%

Таблица 14. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	16	68%	гитара, гуашь, клоун, мультик (2), пуфик, такси, утреннигыттан, утренник (5)
прилагательные	0	0%	-
ФИО и имена собственные	2	11%	Захарова, Ульяна
акронимы	0	0%	-

ошибочные	4	21%	бары, хата (3)

Таблица 15. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	бары, хата	Аммосова, пакекка, утреннига, феннанар, фольгаттан
3-грамм	бары, хата	Аммосова, пакекка, утреннига, феннанар, фольгаттан

Участник в эксперименте ошибся при выделении модального слова *хата* и *бары* (все) и пропустил заимствование *утреннига* (в утренник), *феннанар* (из фена) и *фольгаттан* (из фольги). Словоформа *пакекка* (в пакет) ч трудом распознается русскоязычными людьми как заимствование.

Текст 5. Наш уважаемый руководитель – Иван Сивцев

Таблица 16. Общие показатели текста

Токены	799
Словоформы	786
Уникальные словоформы	570
Количество человеческих разметок	53
Процент вручную выделенных заимствований на тексте	7%

Таблица 17. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	11	21%	бары, инфраструктуратын, лидерэ, отделениетын, партия, позициятын, совхоз, страховкалааһын, тыл, управляющайынан (2)
прилагательные	6	11%	материальнай, партийнай, принципиальнай, советскай, социальнай, якутскай
ФИО и имена собственные	31	58%	Иван (12), Госагропромна, Дмитриевна, Ефимович, Маргарита, Михаил, Николаев, Сивцев (2), Степанович (10), Степановиһы
аккронимы	2	4%	АССР, РФ
ошибочные	3	6%	Баран (2) тыл

Таблица 18. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	баран, бары, тыл	альтернативнай, дьиэуот, Женни, идеялардаах, комсомолецтаахтара, позициялаах, Сивцевтэр, Стрюкова, ТХПК, түмүгу
3-грамм	баран, бары, тыл	альтернативнай, Женни, идеялардаах, комсомолецтаахтара, обкомун, онтон, позициялаах, Сивцевтэр, Стрюкова, ТХПК

В этом случае человек неверно классифицировал частицы *баран* и *бары* и существительное *тыл* (язык) в качестве заимствований. Человеком были пропущены заимствования: *альтернативнай, идеялардаах, комсомолецтаахтара, позициялаах, Сивцевтэр, Стрюкова*, которые обнаружены обоими автоматическими методами. Метод, основанный на правилах, пропустил заимствование «обкомун», которое было выделено статистическим методом. Метод, основанный на правилах, обнаружил опечатку в тексте, которая противоречит закону гармонии гласных: где читается *түмүгу* должно быть *түмүгү* (результат вин. п.). С другой стороны, статистический метод ошибочно классифицировал временной союз *онтон* как заимствование.

Текст 6. Внутри театра мы увидели театр

Таблица 19. Общие показатели текста

Токены	725
Словоформы	721
Уникальные словоформы	510
Количество человеческих разметок	97
Процент вручную выделенных заимствований на тексте	13%

Таблица 20. Заимствовании, выделенные человеком

категория	**кол.**	**проц.**	**словоформы**
существительные	39	41%	автора, актера, актердар, актрисата, драма, жанрыгар, интервьютугар, искусство, итальянец, лауреата, литератураҕа, персонаж (3), персонажей, персонажтар (4), персонажтара, персонажтарга, персонажтарыгар, поисках, пьеса (3), пьесаларыгар, пьесата, пьесатын, режиссер (3), режиссертан, режиссеру, режиссурата, сцена, театрга, трагедия, трагикомедияны
прилагательные	4	4%	интеллектуальнай, риторическай, трагическай, эмоциялаах
ФИО и имена собственные	50	52%	Айталина (3), Айтматов, Алдан, Алексеев, Альберт, Аммосов (2), Анатолий (3), Андрей (4), Борисов (4), Готовцев, Докторова, Дорофеев, Егоров, Егорова (2), Изабелла, Лаврнова (3), Леонтий, Луиджи, Мария, Матрена, Михаил, Москваҕа, Мотхонов, Надежда, Николаев (2), Николай, Павлов, Павлович, Пиранделло (2), Роман, Седельникова, Федотова, Чингиз, Шардина, Ширин
аккронимы	0	0%	-
ошибочные	3	3%	баран (2), дойду

Таблица 21. Разница с автоматическим методами

	выделены вручную, но не обнаружены автоматическим методом	обнаружены автоматическим методом, но не выделены вручную
правила	Алдан, баран, дойду	камернай, нобелевскай

3-грамм	Алдан, баран, дойду	камернай, маны, мин, нобелевскай,

Человеческая разметка пропустила заимствовании «камернай» и «нобелевскай» и ошибся при выделении *дойду* (родина, страна). Статистический метод неверно классифицировал «маны» (мест. указ. вин. п.) и *мин* (я).

Заключение

Эксперименты подтверждают наблюдение, что большинство заимствованных слов являются существительными, за которыми следуют прилагательные. заимствовании глаголов в обработанных текстах не найдены. В дальнейшем следует отфильтровывать акронимы перед экспериментами, так как они не представляют особого интереса для извлечения заимствований.

Оба автоматических метода хорошо справляются со своей задачей и способны уловить заимствования, которые упускает нетренированный глаз при ручной классификации. В то время как метод, основанный на правилах, способен сам по себе разделять заимствования, статистический метод обеспечивает дополнительную гибкость.

В основном, любой из двух подходов может быть полезен, когда необходимо обработать большое количество текстов. Как было замечено в ходе эксперимента, может быть полезным применять оба метода. С одной стороны, метод, основанный на правилах, обеспечивает большую точность и дает надежный способ обнаружения словоформ, не соответствующих якутской фонологии. С другой стороны, статистический метод оказался хорошей статистической транспозицией метода, основанного на правилах. За счет точности он может обеспечить дополнительную эвристическую силу, выделив некоторые заимствования, которые могут быть пропущены, если использовать только метод, основанный на правилах. Коды программ на Питоне, выделяющих заимствования по вышеуказанным методам, доступны в Github репозитории [8].

Библиография

[1] Кортегосо-Виссио Н., Захаров В.П. Выделение русских заимствований в якутских текстах» // Компьютерная лингвистика и вычислительные онтологии. Выпуск 5 (Труды XXIV Международной объединенной конференции «Интернет и современное общество, IMS-2022). СПб: Университет ИТМО, 2022 (в печати).

[2] Слепцов П.А. Русские лексические заимствования в якутском языке. Наука, 1975.

[3] Харитонов Л.Н. Современный якутский язык. Часть первая: фонетика и морфология. Научно-Исследовательский Институт языка, литературы и истории ЯАССР. Якутск: Госиздат ЯАССР, 1947.

[4] Васильева Н.М. К вопросу о правописании заимствованных слов современном якутском языке // Известия Российского государственного педагогического университета им. А.И. Герцена. 2011. № 131. С. 166-169.

[5] Canvar W.B., Trenkle J.M. N-Gram-Based Text Categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. 1994. P. 161–175.

[6] Goldhahn D., Eckart T., Quasthoff U. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012. P. 769-765. URL: http://www.lrecconf.org/proceedings/lrec2012/pdf/327_Paper.pdf (дата обращения: 16.06.2022).

[7] Сахамедия. Сетевое издание «sakhamedia.ru». URL: https://sakhamedia.ru/gazeta-saha-sire/ (дата обращения: 16.06.2022).

[8] A language identification classifier to extract Russian imports from Yakut texts. URL: https://github.com/nicolascortegoso/sakha_loanwords (дата обращения: 16.06.2022).

Two methods for identifying Russian words in Yakut texts

N. Cortegoso Vissio, V.P. Zakharov

***Abstract*— The article discusses two methods for extracting foreign words from Yakut texts. Foreign words refer to non-integrated lexical units, which have not been adapted to Yakut orthography and are therefore written as in the original language. Based on the fact that most foreign words in Yakut texts come from the Russian language, it is assumed that they have a particular form by which they can be distinguished from the Yakut word forms.**

The first method reviewed here is based on rules. It implements an algorithm that detects letter combinations that are foreign to the Yakut language. The second method applies a statistical approach to model and differentiate Yakut and Russian letter combinations.

The effectiveness of both methods in extracting Russian foreign words is compared with the results of manual highlighting performed by Russian speakers on 6 Yakut texts. This work is a continuation of the article "Identification of Russian borrowings in Yakut texts", published in "Computer Linguistics and Computational Ontologies. Number 5 (Proceedings of the XXIV Joint International Conference "Internet and Modern Society, IMS-2022.

***Keywords*— natural language processing, identification of Russian words, Yakut language, n-grams.**

REFERENCES

[1] N. Cortegoso-Vissio and V.P. Zakharov, "Vydeleniye russkikh zaimstvovaniy v yakutskikh tekstakh». Komp'yuternaya lingvistika i vychislitel'nyye ontologii", in *Vypusk 5 (Trudy XXIV Mezhdunarodnoy ob"yedinennoy konferentsii «Internet i sovremennoye obshchestvo, IMS-2022)*, SPb: Universitet ITMO, 2022 (in press).

[2] P.A. Sleptsov, Russkiye leksicheskiye zaimstvovaniya v yakutskom yazyke. Izdates'tvo, Nauka, 1975.

[3] L.N. Kharitonov, Sovremennyy yakutskiy yazyk. Chast' pervaya: fonetika i morfologiya. Nauchno-Issledovatel'skiy Institut yazyka, literatury i istorii YAASSR, Yakutsk: Gosizdat YAASSR, 1947.

[4] N.M. Vasil'yeva, "K voprosu o pravopisanii zaimstvovannykh slov sovremennom yakutskom yazyke", *Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. AI Gertsena*, no. 131, pp. 166-169, 2011.

[5] W.B. Canvar and J.M. Trenkle, "N-Gram-Based Text Categorization", In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.

[6] D. Goldhahn, T. Eckart and U. Quasthoff, "Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages", in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 769-765 [Online]. Available: http://www.lrecconf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

[7] Sakhamediya. Setevogo izdaniya «sakhamedia.ru» [Online]. Available: https://sakhamedia.ru/gazeta-saha-sire/.

[8] A language identification classifier to extract Russian imports from Yakut texts [Online]. Available: https://github.com/nicolascortegoso/sakha_loanwords.