ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Компьютерная лингвистика»

Зайдельман Людмила Яковлевна

Моделирование речевого поведения носителя миноритарного языка РФ в социальной сети

Native Speaker Speech Behaviour Modelling (Based on Russian Federation Minor Languages in a Social Network)

Выпускная квалификационная работа студента 2 курса магистратуры группы МТР141

Академический руководитель
образовательной программы
канд. филологических наук, доц.
А. А. Бонч-Осмоловская

Научный руководитель

канд. филологических наук, доц.
Б. В. Орехов

Москва 2016

Оглавление

1. Введение

Интернет представляет собой широкое поле для лингвистических исследований, его можно считать самым большим корпусом живого языка. Подавляющая его часть – это тексты на больших языках, таких как испанский, английский, русский. Однако есть сегменты интернета, в которых есть тексты и на миноритарных языках. Одним из таких сегментов являются социальные сети, ставшие удобной платформой для создания групп по интересам. В такие группы (или сообщества) носители миноритарных языков и объединяются для общения друг с другом.

В этой работе мы рассмотрим 43 миноритарных языка России с числом носителей от 2 тысяч до 4 миллионов. В качестве материала для исследования будут использованы данные, полученные из более чем 1600 сообществ социальной сети Вконтакте (vk.com), в которых говорят на миноритарных языках России.

В рамках этой работы мы попытаемся выяснить, как устроено речевое поведение носителей разных миноритарных языков. Мы постараемся нарисовать портрет типичного носителя миноритарного языка в социальной сети, проанализировав такие данные о пользователях, как возраст, место жительства, а также собственно речевое поведение: участие в диалогах, использование или неиспользование миноритарного языка, количество сообществ, говорящих на миноритарном языке, членом которых является конкретный пользователь. Кроме того, объектом нашего исследования станут группы носителей, то есть сообщества. Мы проверим, есть ли зависимость между числом пользователей и числом сообществ, и выясним, насколько правомерно называть такие сообщества «говорящими на миноритарных языках».

Результатом работы будет общая модель поведения типичного носителя миноритарного языка, если в процессе работы мы выясним, что носители разных миноритарных языков действительно похожи друг на друга. Однако если исследование покажет, что речевое поведение носителей разных языков различное, мы постараемся выделить дифференцирующие признаки, которые позволят нам разделить языки на несколько групп.

1.1. Описание данных

В 2014 году на базе Школы лингвистики ВШЭ стартовал проект «Языки России», целью которого было собрать коллекции текстов на миноритарных языках России. В результате были собраны коллекции двух типов: тексты из интернета (региональ-

ные СМИ, сайты местных органов управления, блоги и пр.) и тексты из социальной сети Вконтакте. В этой работе мы будем использовать данные из сети Вконтакте, полученные в результате работы этого проекта (*Сайт проекта «Языки России»* б.г.).

Все данные, используемые в работе, относятся к миноритарным языкам России, поэтому важно объяснить, что именно мы понимаем под этим термином. В (Cormack, Hourigan 2007) миноритарные языки определяют как языки, подавляемые окружающими их крупными языками. В нашей работе мы тоже рассматриваем языки, подавляемые большим языком – русским, но кроме этого мы накладываем на миноритарные языки еще два ограничения: во-первых, эти языки используются на территории Российской Федерации, а во-вторых, они не являются при этом государственными языками других государств. По этому определению татарский язык, являющийся государственным языком республики Татарстан, является миноритарным, а украинский язык – нет, поскольку это государственный язык другого государства – Украины. Всего в России около ста миноритарных языков (Katzner 2002).

Кратко опишем процесс получения данных. Для каждого языка был составлен список уникальных для этого языка слов-маркеров. Такие слова искались вручную в грамматиках, словарях и разговорниках миноритарных языков. Поиск текстов велся с помощью Яндекс.XML – сервиса, позволяющего посылать автоматические запросы поисковой машине Яндекса. Для получения списка страниц, на которых есть тексты на интересующем нас языке, в запросе необходимо указать адрес сайта, на котором будет вестись поиск (в нашем случае это адрес социальной сети Вконтакте: https://vk.com) и слова-маркеры. Поскольку слова-маркеры однозначно указывают на язык текста (это одно из необходимых условий, учитывавшихся при поиске), в результате таких запросов мы получали страницы с текстами на нужном языке. Для того, чтобы получить большее число страниц, в качестве слов-маркеров выбирались частотные для языка слова, то есть служебные части речи и местоимения. В сети Вконтакте есть два типа страниц: персональные страницы и страницы сообществ. Из всех полученных ссылок мы оставили только ссылки, ведущие на страницы сообществ. Мы приняли такое решение, посчитав, что пользователь скорее будет использовать для записей на своей персональной страницице русский язык, поскольку среди его друзей, вероятно, есть не говорящие на известном ему миноритарном языке. В сообщества же пользователи объединяются благодаря наличию у них общих интересов, таким объединяющим интересом может быть и язык. Данные со страниц сообществ мы получали с помощью API Вконтакте – инструмента, позволяющего работать со страницами социальной сети, не задумываясь об их внутренней структуре. Важное преимущество данных из социальных сетей перед данными с других интернет-ресурсов в том, что кроме собственно текстов у нас

есть возможность получить большое количество метаинформации. Перечислим данные о сообществах, которые мы сохраняли: название сообщества; количество его участников; список всех постов, опубликованных в сообществе, включая комментарии; кроме того, для каждого поста мы сохраняли дату его создания и всю доступную нам информацию о его авторе: пол, дату рождения, город проживания. Несмотря на то, что все сообщества были найдены по словам-маркерам, тексты в них оказались не только на интересующем нас языке, но также и на русском. Поэтому с помощью метода буквенных n-грамм (Cavnar, Trenkle 1994), адаптированного для языков, не имеющих своих корпусов, мы разработали инструмент, определяющий язык текста. С помощью этого инструмента все тексты были дополнены языковой пометой: rus – для текстов на русском языке, *iso-код миноритарного языка* – для текстов на миноритарных языках, trash – для «мусорных текстов», содержащих ссылки, смайлы и эмодзи или не содержащих текста вообще (например, такое возможно, если оригинальный пост состоял исключительно из иллюстраций). Мы попытались скачать данные для 43 языков, для 31 из них нам удалось найти хотя бы одно сообщество. Всего было выкачано и обработано 1630 сообществ. Все данные были получены в период с 18.01.2015 по 29.04.2016. Процесс получения данных по шагам см. (Zaydelman [и др.] in print).

Все данные, которые используются в этой работе, открытые и доступны всем пользователям интернета. Такая личная информация как имена, фамилии и пользовательские id были заменены на искусственно созданные номера, которые позволяют различать пользователей между собой, но при этом не дают возможности восстановить исходные данные о пользователях.

1.2. Исследования о миноритарных языках

Тот факт, что языки вымирают, неоспорим, язык жив, пока на нем разговаривают, а разговаривают все больше на мажоритарных языках – языках большинства. Однако с возрастающей популярностью интернета растет число исследований, связанных с представленностью миноритарных языков в интернете, а также растет и уверенность некоторых исследователей в том, что именно интернет может спасти миноритарные языки от неминуемого исчезновения.

В (Prado 2012) анализируется сложившаяся в мире ситуация с миноритарными языками и их представленностью в интернете. В статье автор пытается выяснить, какие причины препятствуют более широкому распространению миноритарных языков в интернете. В частности, он упоминает поисковые системы и системы машинного перевода, которые пока так плохо работают с миноритарными языками, являющимися, порой бо-

лее многочисленными, чем некоторые мажоритарные языки. Кроме того, Прадо неоднократно заявляет, что изначально созданный для английского языка интернет до сих пор не приспособился к системе письменности некоторых языков, лишая носителей этих языков возможности общаться на них в интернете. Автор предполагает, что логичным развитием таких языков, а также языков, не имеющих письменности, может стать появление большого числа медиаконтента – звуковых и видеофайлов. Однако сам же автор признает проблемы, которые непременно возникнут на пути к возникновению нетекстового сегмента интернета: создание и даже поиск такого контента требует большей компьютерной грамотности и большей технической оснащенности. Борцмейер в (Bortzmeyer 2012) рассматривает существующие стандарты, которым должно соответствовать все ПО в интернете. В статье он поднимает вопрос об адаптированности этих стандартов по отношению к многоязычному интернету.

А. Корнай в (Kornai 2013) изучает и оценивает жизнеспособность языков в интернете, для этого он вводит четырехступенчатую шкалу. Эта шкала является некоей адаптацией для интернет-среды рейтинга EGIDS (Expanded Graded Intergenerational Disruption Scale), который имеет 13 уровней жизнеспособности (Lewis, Simons 2010). Противопоставляя живые языки мертвым, Корнай, тем не менее, разделяет языки на «более живые» и «менее» и, соответственно, «более мертвые» и «менее», это различие отражено в его шкале: thriving (T) – процветающие языки, на которых много общаются в интернете, причем не только носители, такие языки также поддерживаются крупными IT-компаниями, vital (V) – живые языки, тоже достаточно популярные и распространенные, но в меньшей степени, heritage (H) – наследованные языки, материалы на этих языках можно встретить в интернете, однако это будут не результаты живого общения, а скорее архивы, и последний уровень – still (S) – неподвижные языки, такие языки никто не использует в интернете, и вероятность встретить материалы на них крайне мала. По подсчетам, сделанным авторами работы, число процветающих языков (V) составляет лишь 5% от неподвижных (S), то есть 95% всех языков не представлены в интернете.

1.3. Исследования о социальных сетях

В (Benevenuto [и др.] 2009) авторы анализируют поведение пользователей четырех социальных сетей: Orkut, MySpace, Hi5, и LinkedIn. Авторы предлагают в качестве данных использовать не видимое поведение пользователей, а скрытое: в работе используется информацию о «кликах» 37,024 пользователей, собранную за 12 дней. Эта информация включала в себя частоту подключений к социальным сетям, время проведенное

пользователями в каждой из сетей, а также информацию о различных действиях, которые пользователи могут выполнять в социальных сетях: поиск других пользователей и посещение их страниц и страниц сообществ, написание личных или публичных сообщений, публикация фотографий и видеороликов и др. Одним из выводов, которые делают авторы, является то, что в основном пользователи социальных сетей ведут себя очень пассивно: отслеживают обновления страниц своих друзей, просматривают персональные страницы других пользователей – такая пассивная деятельность составляет 92% всех кликов. Оставшиеся 8% приходятся на активные действия: написание сообщений и постов, публикацию и редактирование медиаматериалов.

2. Пользователи

2.1. Количество пользователей

Когда речь заходит о численных данных для языков, в первую очередь имеют в виду количество носителей. Данные о числе носителей известны благодаря переписи населения, регулярно проводимой государством. В этой работе мы исследуем не просто языки, а языки в интернете, а значит кроме числа носителей нас интересует также число интернет-пользователей. Интересно, есть ли зависимость между числом носителей и числом пользователей. Логичной и закономерной выглядит гипотеза: чем больше носителей, тем больше пользователей. В наших данных о сообществах Вконтакте есть информация о числе участников для каждого сообщества. Но возникает вопрос: можно ли считать, что число участников во всех сообществах одного языка и есть число пользователей этого языка. Чтобы ответить на этот вопрос, нужно сделать важное замечание. Членство в сообществе, говорящем на миноритарном языке, не делает человека носителем этого языка. Очевидно, что не все члены сообщества являются активными участниками, создающими посты или комментарии. Мы с уверенностью можем утверждать, что человек является пользователем миноритарного языка, только если нами было зафиксировано, что он пользовался этим языком, то есть написал как минимум один пост или комментарий, используя миноритарный язык. Несмотря на то, что у нас нет возможности анализировать поведение «читателей», нельзя недооценивать их участия в жизни сообществ. Тексты в социальных сетях существуют до тех пор, пока их читают, причем важно не столько качество читающих, сколько их количество (Филь 2012).

На рис. 1 на оси x – миноритарные языки, упорядоченные по уменьшению числа носителей по данным Всероссийской переписи населения в 2010 году (*Портал «Всероссийская перепись населения 2010 года»* 2010). На график попали лишь те языки, для которых нашлось по крайней мере одно сообщество. Синие кружки соответствуют натуральному логарифму от числа носителей для каждого языка. Красные квадратики соответствуют натуральному логарифму от числа пользователей, то есть людей, написавших по крайней мере один текст на соответствующем языке в сети Вконтакте. Поскольку разброс в данных слишком большой, вместо абсолютных величин мы использовали натуральный логарифм от них. По такому рисунку нельзя понять, насколько один показатель больше другого, но факт превосходства одного показателя над другим виден хорошо. Можно заметить, что красный график прерывается на эвенском языке, это объясняется тем, что несмотря на то, что с помощью поисковой машины Яндекса нам удалось найти одно сообщество на этом языке, в нем не оказалось ни одного текста на эвенском, а значит нет ни одного пользователя для этого языка. Вероятно,

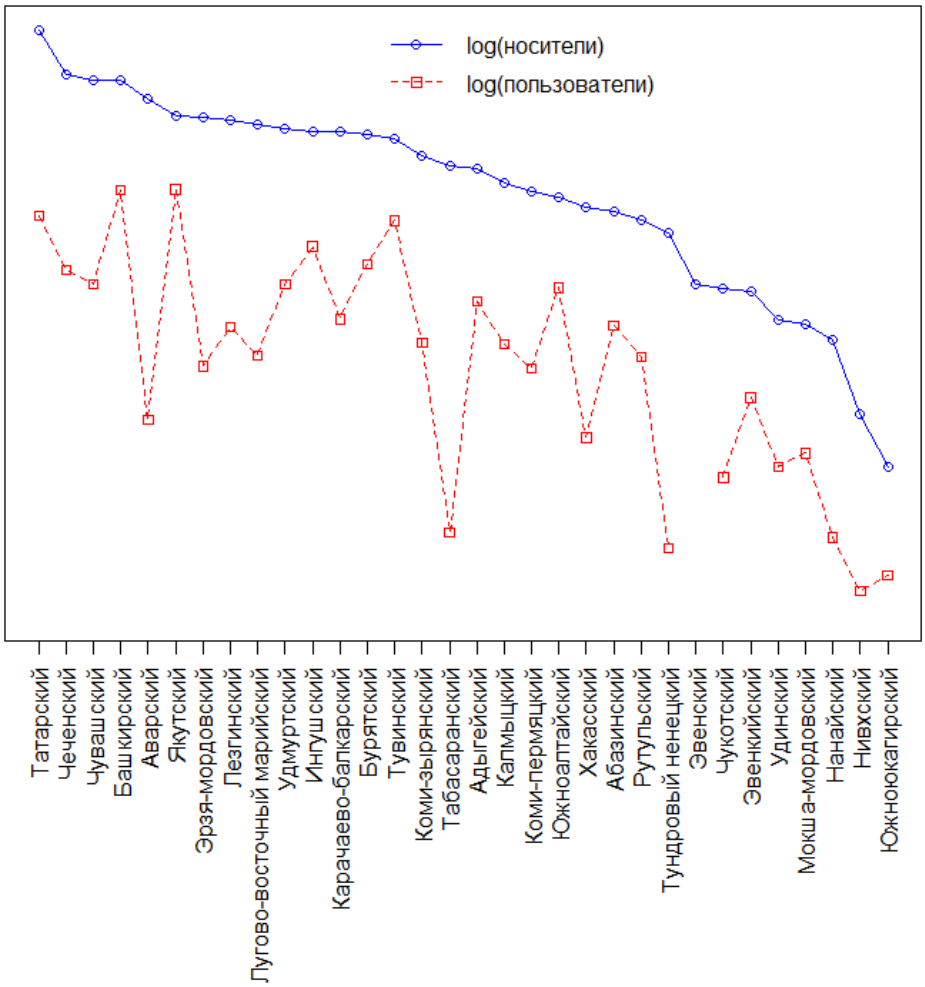

Рис. 1: Изменение числа пользователей и числа носителей в зависимости от языка.

это сообщество попало в выдачу поисковика, потому что в его названии или описании встретилось слово-маркер на эвенском. В любом случае, видно, что красный график, в отличие от синего, не монотонно убывает, а значит первоначальная гипотеза о том, что число носителей и число пользователей линейно зависимы, неверна.

2.2. География пользователей

Как мы уже говорили раньше, важным отличием данных из социальных сетей от данных из просто интернета является дополнительная информация о пользователях, такая как: возраст, пол и родной город. Используя данные о городе пользователя, мы сможем определить точки распространения языка. Мы рассчитываем, что на первом месте по числу пользователей для каждого языка окажется титульный регион, то есть

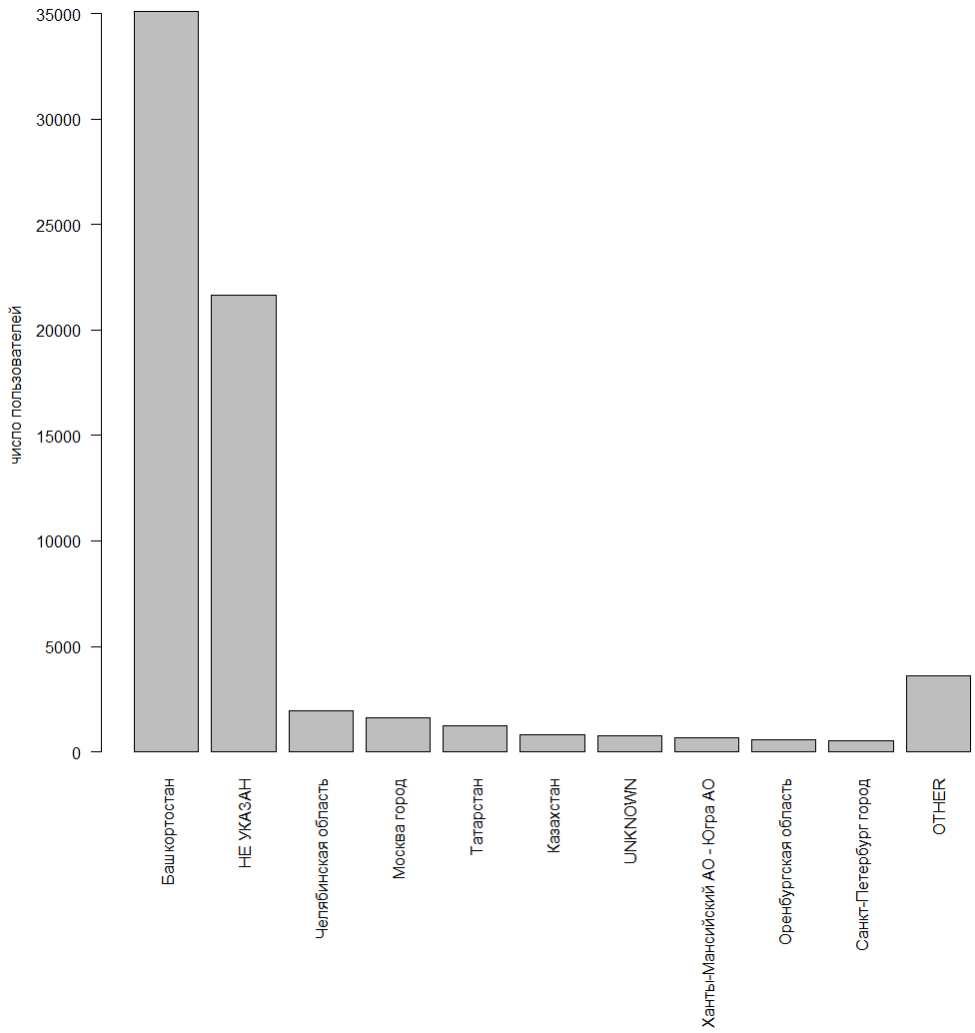

Рис. 2: Распределение пользователей по регионам для башкирского языка.

регион, в котором по данным переписи населения проживает наибольшее число носителей этого языка. Чтобы проверить, так ли это, а также попробовать определить какие-нибудь неожиданные закономерности, мы сопоставили городам, указанным на личных страницах пользователей, регион. Для городов России регионом считается регион РФ, а для иностранных городов за регион мы принимали название государства. Это сопоставление мы произвели с помощью базы городов (*База стран и городов citieslist* б.г.). Для каждого языка мы нарисовали столбчатую диаграмму. Поскольку географическое разнообразие очень велико, мы решили не пытаться поместить на рисунок все регионы, а для каждого языка изобразить топ-10 регионов, одиннадцатый столбик OTHER вместил в себя все регионы, не попавшие в первую десятку. На рис. 2 и рис. 3 показано распределение пользователей для башкирского и чеченского языков. Всего мы построили 31 такой график для каждого языка, имеющего хотя бы одного пользователя. На обоих представленных графиках видно, что среди первых десяти столбиков есть два не отно-

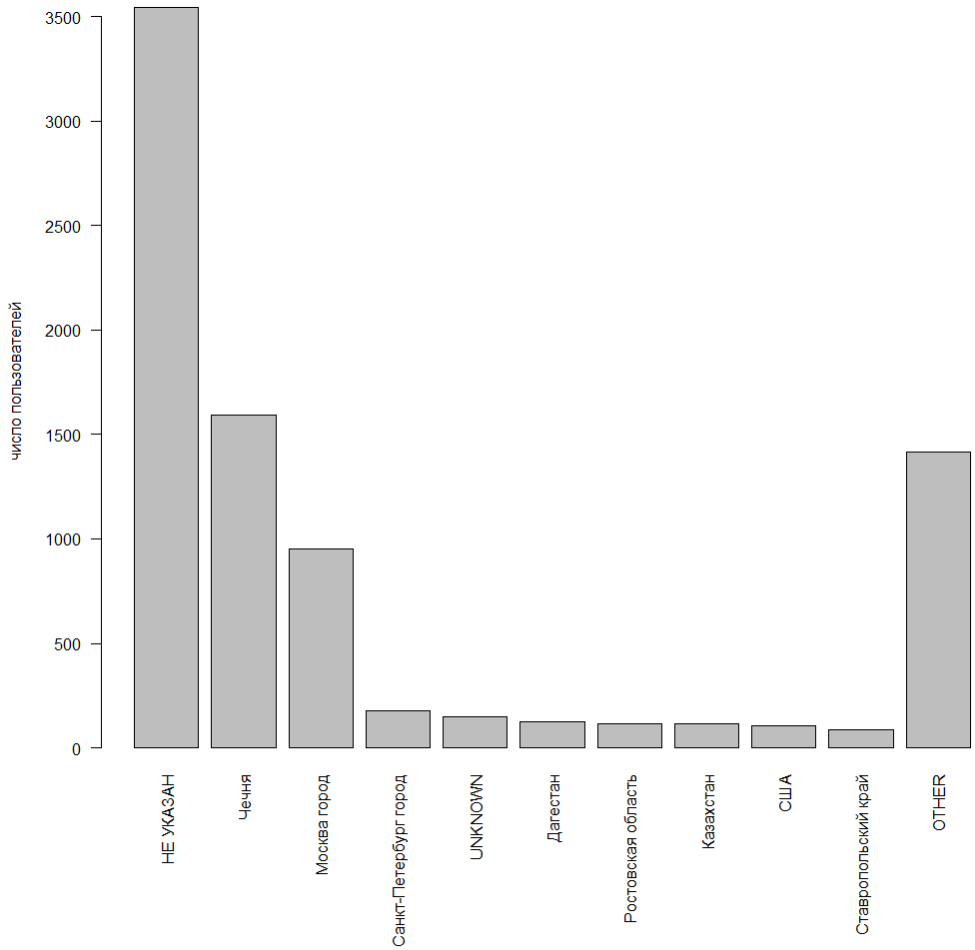

Рис. 3: Распределение пользователей по регионам для чеченского языка.

сящихся к регионам – это НЕ УКАЗАНО и UNKNOWN. Пометку НЕ УКАЗАНО вместо информации о регионе получали пользователи, не указавшие информацию о месте проживания. Любопытно, что среди пользователей чеченского языка таких пользователей оказалось не просто большинство, а почти половина (42,3%). По данным, собранным в 2011 году для всех пользователей Вконтакте (Демьяненко 2011), город не указали 69,45% пользователей социальной сети. Пометка UNKNOWN вместо названия региона проставлялась пользователям, чьих городов не оказалось в базе данных, и мы не смогли достоверно определить регион. Возвращаясь к рис. 2 и 3, заметим, что, если сосредоточиться на восьми «настоящих» регионах, то для обоих языков регионом-лидером стал именно титульный регион, как мы и ожидали. На рис. 4 мы разместили все языки, имеющие носителей, каждому языку соответствует сразу три столбика: титульный регион (зеленый), Москва (синий), Санкт-Петербург (красный). На графике показаны не абсолютные величины, а доли от общего числа пользователей. Языки упорядоче-

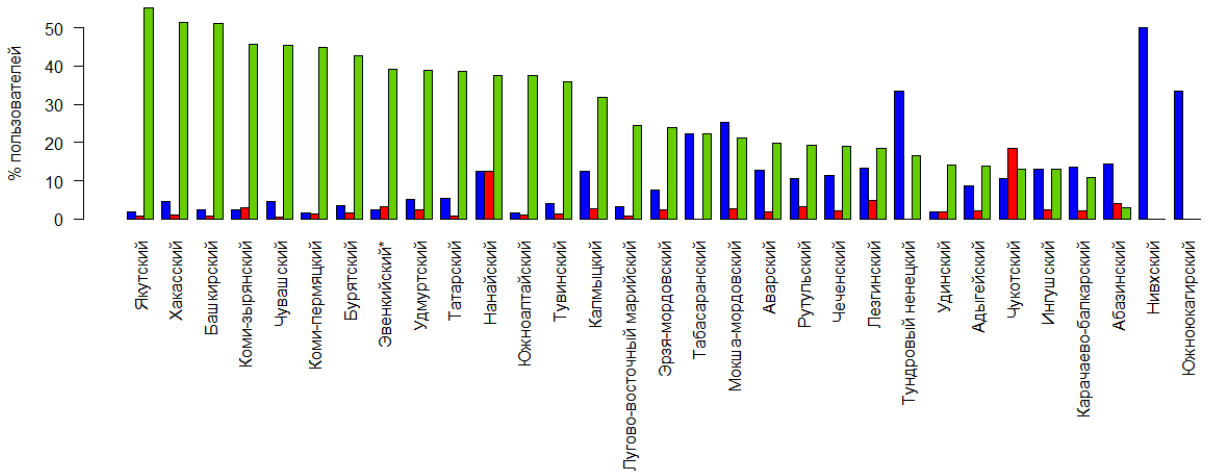

Рис. 4: Соотношение между пользователями, проживающими в титульных регионах (зеленый), в Москве (синий) и в Санкт-Петербурге (красный).

ны по доле проживающих в титульном регионе пользователей в порядке убывания. Отобразить на графике кроме числа проживающих в титульном регионе еще и число проживающих в двух крупнейших городах России мы решили, поскольку для языков, где лидирует не титульный регион, именно эти региона оказались наиболее популярными. Всего из 31 языка у 23 лидирует именно титульный регион, среди оставшихся 8 языков у 7 на первом месте оказался город Москва и у одного – Санкт-Петербург (18,4% или 7 пользователей чукотского языка из 38 оказались из Санкт-Петербурга).

2.3. Возраст пользователей

Ценность данных о дате рождения гораздо ниже: по данным (Демьяненко 2011) полную дату рождения (день, месяц и год) указали лишь 17,47% всех пользователей. Для наших данных этот показатель оказался выше: 30,2% пользователей-носителей миноритарных языков опубликовали свой возраст. Для двух языков – нивхского и южноюкагирского – этот показатель составил 0%, это языки с очень небольшим числом носителей – общее число их пользователей равно 2 и 3, соответственно. Самый высокий процент полных дат рождения у тундрового ненецкого языка – 66,67%, этот язык также является довольно небольшим по числу пользователей, их 8, из них четверо указали свою дату рождения. Средний возраст пользователей всех языков оказался в промежутке от 19,6 до 31,3 лет. Социальная сеть Вконтакте позволяет указать в качестве года рождения любой год в период с 1901 по 2002 годы. Рискуя потерять часть достоверной информации, мы тем не менее, определили минимально допустимую дату рождения, а все даты меньше нее считали подозрительными. Такой датой стало 1 января 1940 го-

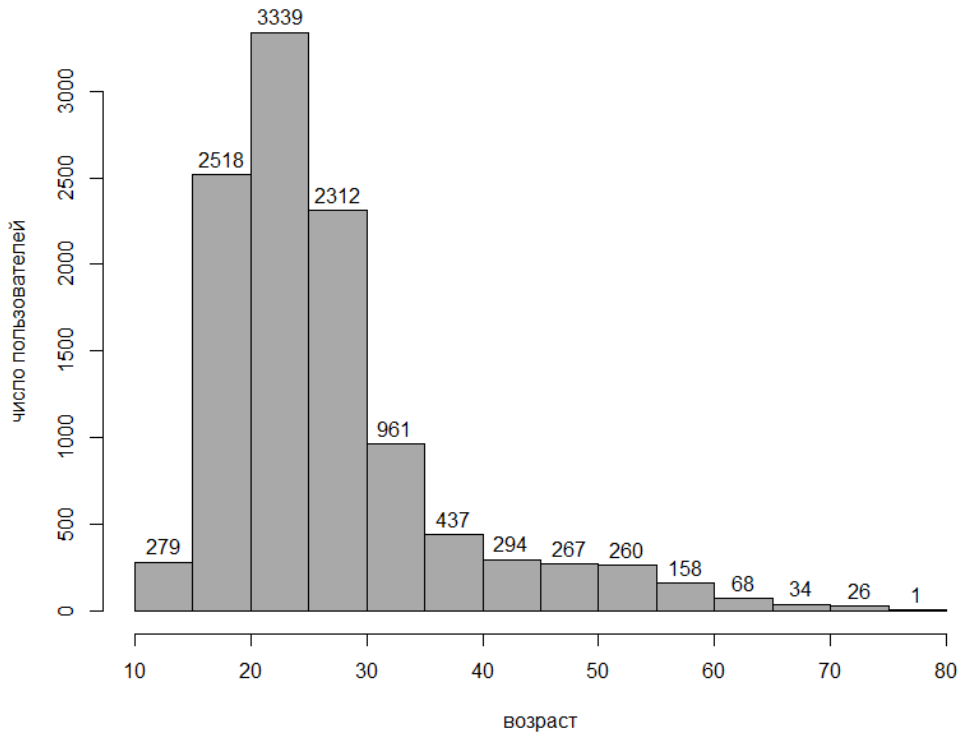

Рис. 5: Гистограмма распределения возрастов пользователей татарского языка.

да. Всего среди наших данных оказалось 1,89% подозрительных дат рождения, которые мы не учитывали при нахождении среднего возраста. На рис. 5 и 6 показаны распределения возрастов пользователей для двух языков: татарского и карачаево-балкарского. Эти графики очень похожи: оба распределения близки к нормальному распределению, больше половины всех пользователей на них находятся в возрасте от 15 до 30 лет (74,6% для татарского языка и 85,3% для карачаево-балкарского). Подобным образом выглядят распределения и для остальных языков.

2.4. Языковые способности пользователей

Все предыдущие вычисления мы производили на основании данных о пользователях, однако графики и выводы по ним мы делали не для пользователей, а для языков. Это объясняется тем, что для каждого пользователя мы определили, носителем какого языка он является, создав для него языковой профиль. Языковой профиль – это n-мерный вектор, где n равно числу языков, в нашем случае – 32: 31 миноритарный язык, имеющий хотя бы одного носителя, плюс русский, каждая компонента вектора равна нулю или единице в зависимости от того, владеет ли человек языком, соответствующим этой компоненте. Мы считали, что пользователь владеет языком, если он написал

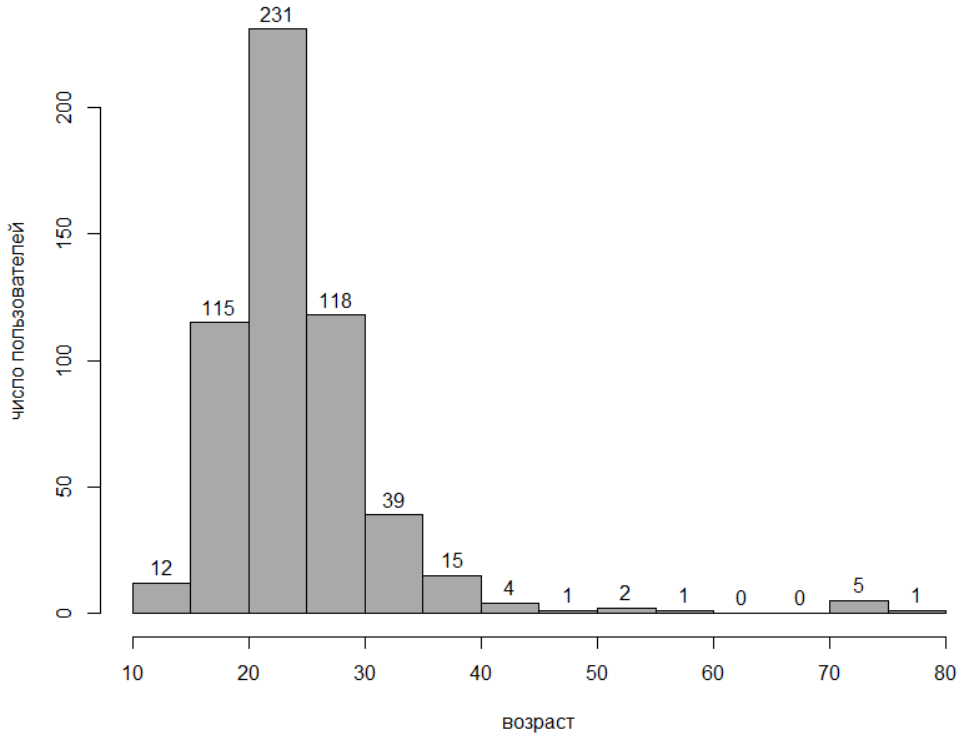

Рис. 6: Гистограмма распределения возрастов пользователей карачаево-балкарского языка.

на нем по крайней мере один текст. Неудивительно, что среди пользователей оказались такие, кто писал на двух или даже трех миноритарных языках. Такие пользователи попадали в число носителей каждого используемого ими языка. На основе языковых профилей пользователей мы составили матрицу совместной встречаемости для языков. Столбцами и строками этой матрицы стали языки, а в ячейки, образованные в результате пересечения двух языков мы поместили число, равное количеству человек, владеющих этими двумя языками. Если какой-то пользователь владел тремя миноритарными языками (случаев с четырьмя и большим количеством языков в наших данных не встретилось), то каждая пара из этой тройки получала очко встречаемости. Например, если пользователь владеет языками X,Y и Z, в нашей таблице числа в шести ячейках должны увеличиться на один, это ячейки XY, YX, XZ, ZX, YZ и ZY. Из 31 языка, имеющих пользователей, только абазинский язык оказался единственным для всех его носителей. Каковы возможные причины того, что два языка имеют общих носителей? Во-первых, схожесть языков, их близость с точки зрения генетической классификации языков. Во-вторых, географическая близость или даже пересечение регионов, в которых эти языки распространены. В-третьих, индивидуальные причины конкретного пользователя, среди таких причин может быть лингвистический интерес, переезд в другой регион и др. И наконец, в-четвертых, ошибки в данных. К сожалению, мы не можем гарантировать,

что наши данные абсолютно чистые: все сообщества мы получаем автоматически с помощью слов-маркеров, которые, безусловно тщательно проверяются на уникальность, но тем не менее могут оказаться «шумными», то есть относиться сразу к нескольким языкам.

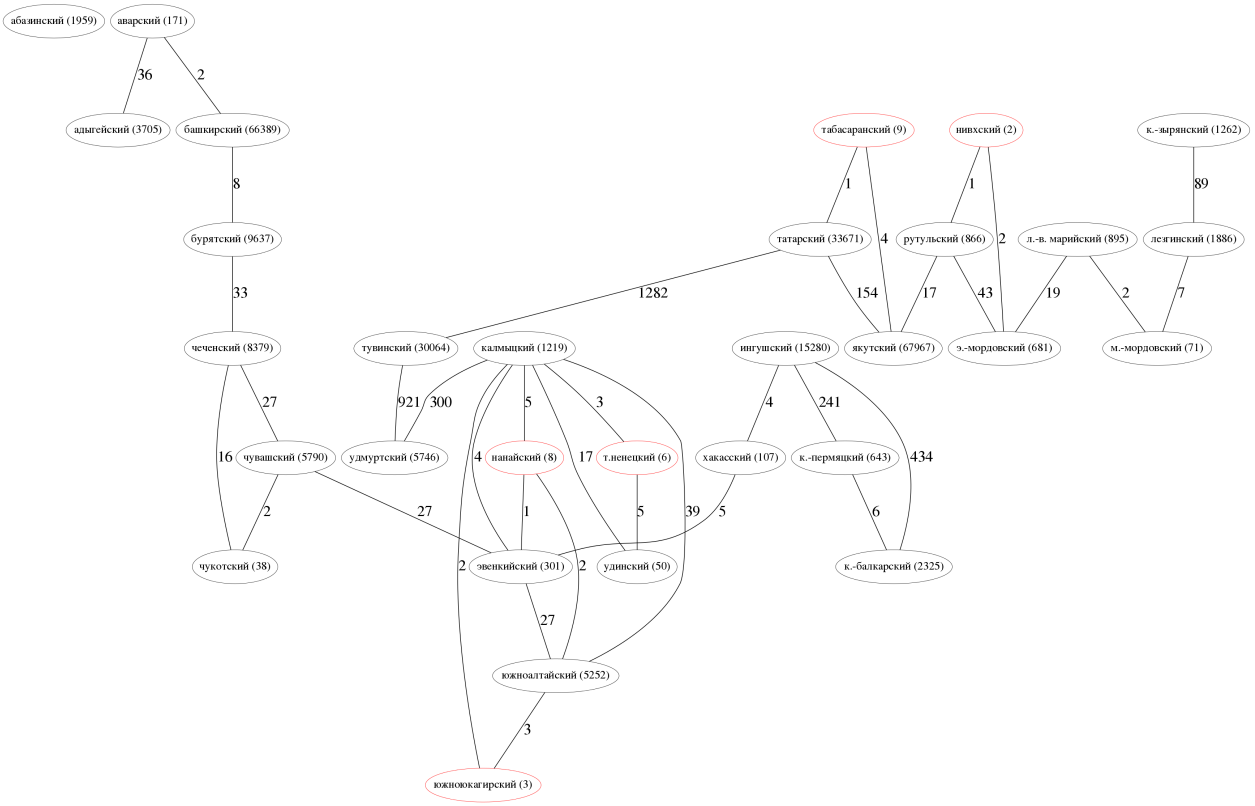

Рис. 7: Граф персечения языков по числу общих пользователей.

На рис. 7 изображен граф, иллюстрирующий эту таблицу смежности. Узлы графа – это языки, имеющие пользователей, для компактности нам пришлось сократить названия некоторых языков, надеемся, что это не вызовет у читателей неудобств, в скобках после названия языка указано число пользователей-носителей для этого языка. Ребро между двумя вершинами-языками означает, что есть какое-то количество людей, владеющих обоими языками, вес ребра показывает это количество. Наибольшее пересечение носителей в этой таблице получилось у татарского и тувинского языков – 1282 человека владеют обоими языками. 1282 человека – это 4,3% от общего числа пользователей тувинского языка и 3,8% от числа пользователей татарского. Попробуем предположить, с чем связано такое большое пересечение. 1) Лингвистическая близость: оба языка принадлежат тюркской языковой семье, то есть являются достаточно похожими. 2) Географическая близость: регионы распространения этих языков – республика Татарстан и республика Тыва – находятся далеко друг от друга и входят в разные федеральные округа; по данным переписи (*Портал «Всероссийская перепись населения 2010 года»* 2010), татары составляют 0,11% населения Тывы, о тувинцах, проживаю-

щих на территории Татарстана информации нет. 3) Личные мотивы: число двуязычных носителей достаточно велико для того, чтобы считать его случайным, объясняющимся исключительно личными причинами носителей. 4) Ошибочные данные: среди тувинских сообществ одно, например, могло оказаться татарским (или наоборот). Таким образом, не исключая вероятность ошибки в данных, мы полагаем, что такое пересечение объясняется родственностью рассматриваемых языков. Может показаться, что ответом на вопрос: есть ли в данных ошибка, может стать информация о месте проживания пользователей из пересечения. Однако, как мы увидели раньше, титульный регион является не единственным местом проживания носителей, а для некоторых языков даже не основным (см. рис. 4). В действительности определить, являются ли пользователи двуязычными, можно только с помощью экспертов по одному из языков. На рис. 7 некоторые вершины мы выделили красным цветом, языки, им соответствующие, имеют меньше десяти пользователей, при этом часть этих пользователей (или даже все) являются также носителями других языков. Языки, отмеченные красным, мы считаем наиболее «подозрительными», возможно, в действительности в сети Вконтакте нет материалов на этих языках.

Вернемся к рис. 7. Степенью вершины в неориентированном графе называется число ребер, выходящих из этой вершины. Мы посчитали среднюю степень для нашего графа, она равна 2,52, это значит, что в среднем у каждого из 30 исследуемых языков есть общие носители с $2-3$ языками. В действительности же обычно степень отдельно взятой вершины может сильно отличаться от среднего значения, чтобы понять разброс значений, находят минимальную и максимальную степени. Минимальной степенью в нашем случае, очевидно, является 0 – именно столько ребер выходит из вершины, соответствующей абазинскому языку. Максимальная степень у вершины, соответствующей калмыцкому языку, он связан с семью другими языками, три из которых являются «подозрительными». По данным UNESCO, калмыцкий язык находится на третей ступени сохранности из шести: его статус «под угрозой исчезновения». Такой статус получают языки, на которых поколение дедушек и бабушек еще разговаривает, а уже поколение родителей язык только понимает и не использует его для общения друг с другом и со своими детьми (Мозли 2010). Несмотря на то, что третья ступень из шести – это довольно высокая оценка для миноритарного языка, правительство республики Калмыкия ведет активную политику популяризации калмыцкого языка (Денисова 2015), чтобы предотвратить его исчезновение. Мы можем предположить, что такая популяризация могла побудить калмыков, проживающих не на территории Калмыкии, а в регионах, являющихся титульными для других миноритарных языков, начать больше общаться на калмыцком языке. Другим возможным следствием мог стать интерес жителей Калмыкии к другим миноритарным языкам России, находящимся под угрозой

исчезновения.

3. Сообщества

3.1. Количество сообществ

Следующей иерархической единицей после пользователей являются сообщества, в которые пользователи объединяются. В предыдущей главе мы выяснили, что неправильно утверждать, что чем больше у языка носителей, тем больше у него пользователей Вконтакте. Поскольку мы изучаем представленность языков именно в социальной сети Вконтакте, будем упорядочивать языки не по числу их носителей, а по числу пользователей. Теперь, если мы будем говорить о самом большом языке, мы будем иметь в виду не татарский, число носителей которого по оценкам переписи 2010 года равняется 4280000 человек, а якутский, на котором, по нашим данным, 67967 человек написали хотя бы один текст в якутоговорящих сообществах сети Вконтакте.

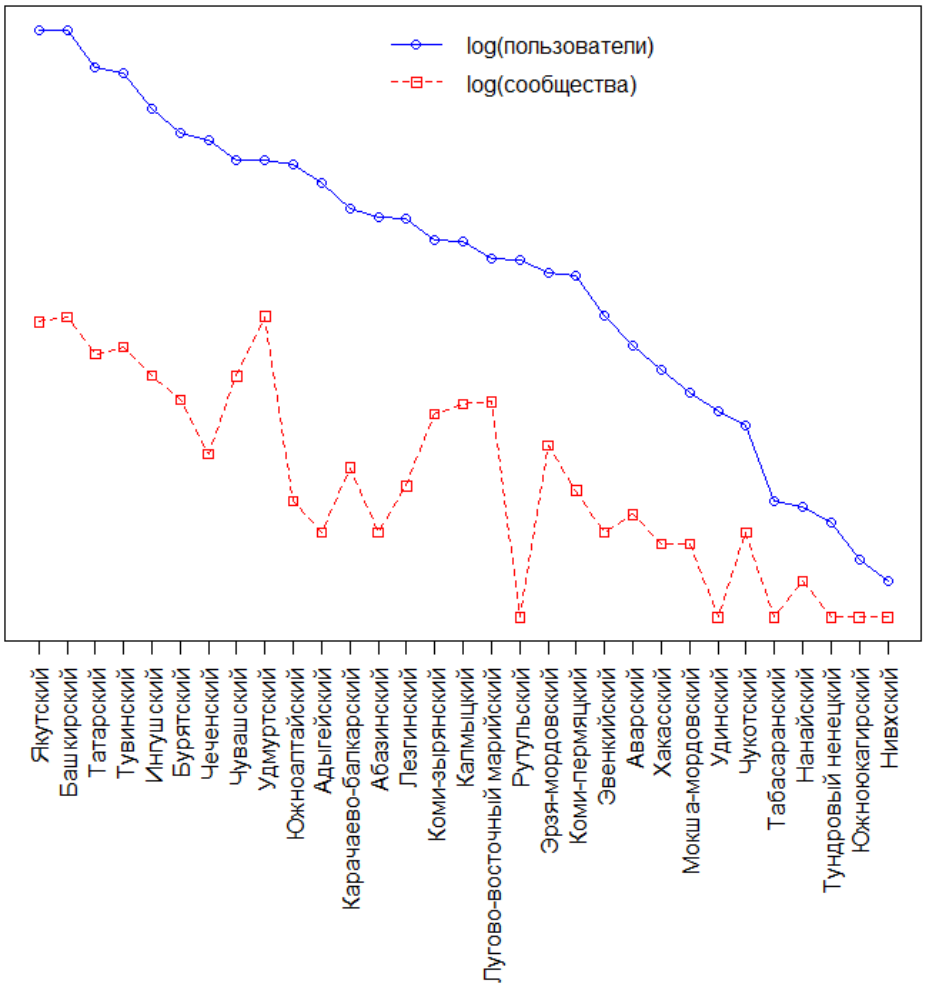

Рис. 8: Изменение числа пользователей и числа сообществ в зависимости от языка.

Прежде всего мы проверим еще одну «наивную» гипотезу, аналогичную гипотезе о зависимости числа пользователей и носителей языка (см. главу Пользователи). Гипотеза звучит так: мы предполагаем, что от числа пользователей зависит число сообществ. Гипотеза основана на логичном предположении: чем у языка меньше пользователей, тем сильнее они ощущают уникальность и ценность и поэтому стараются держаться вместе, чтобы сохранить язык. Пользователи больших языков не чувствуют опасности исчезновения их языка, поэтому не считают необходимым объединяться в группу исключительно по языковому принципу, они делятся на более мелкие подгруппы, объединенные одним языком, но различные по тематике. Для проверки этой гипотезы и при последующем анализе данных мы не будем учитывать сообщества, в которых не было обнаружено ни одного текста на миноритарном языке.

На рис. 8 на оси x по-прежнему миноритарные языки, их стало на один меньше: мы исключили эвенский, в единственном сообществе которого не оказалось ни одного текста на эвенском. Теперь языки упорядочены по уменьшению числа пользователей. Как и на рис. 1, вместо абсолютных величин мы использовали натуральные логарифмы. Синие кружки соответствуют натуральному логарифму от числа пользователей для каждого языка. Красные квадратики соответствуют натуральному логарифму от числа сообществ. По рисунку видно, что снова наивная гипотеза оказалась неверной. Интересен пик у удмуртского языка, говорящий о том, что сообществ на удмуртском сильно больше, чем мы могли бы ожидать исходя из числа пользователей у этого языка и соотношения между числом пользователей и сообществ для других языков. Если посмотреть в наши данные, то окажется, что удмуртский язык занимает первое место по числу сообществ, их у него 298, и лишь девятое место по числу пользователей, их 5746. Разнообразие удмуртского сегмента социальной сети Вконтакте отмечает К. Пишлегер в (Пишлёгер 2013). Среди прочего он выделяет сообщества о событиях в «удмуртском мире» и удмуртской культуре, сообщества, являющиеся официальными и неофициальными СМИ, сообщества с карикатурами, комиксами и мемами на удмуртском, сообщества для изучения удмуртского языка, а также сообщества для изучения других языков с помощью удмуртского.

Глядя на рис. 8 и особенно на удмуртский пик, мы задались вопросом: как распределены люди по сообществам. Интересно, сколько людей участвует в жизни нескольких сообществ (как и прежде, мы считаем, что человек – активный участник сообщества и носитель языка, если он написал не меньше одного поста на миноритарном языке), а сколько – только в одном. На рис. 9 показано, какая часть пользователей пишет только в одно сообщество (здесь, естественно, имеются в виду сообщества, говорящие на миноритарных языках, а не сообщества вообще). Любопытно, что для каждого языка (где

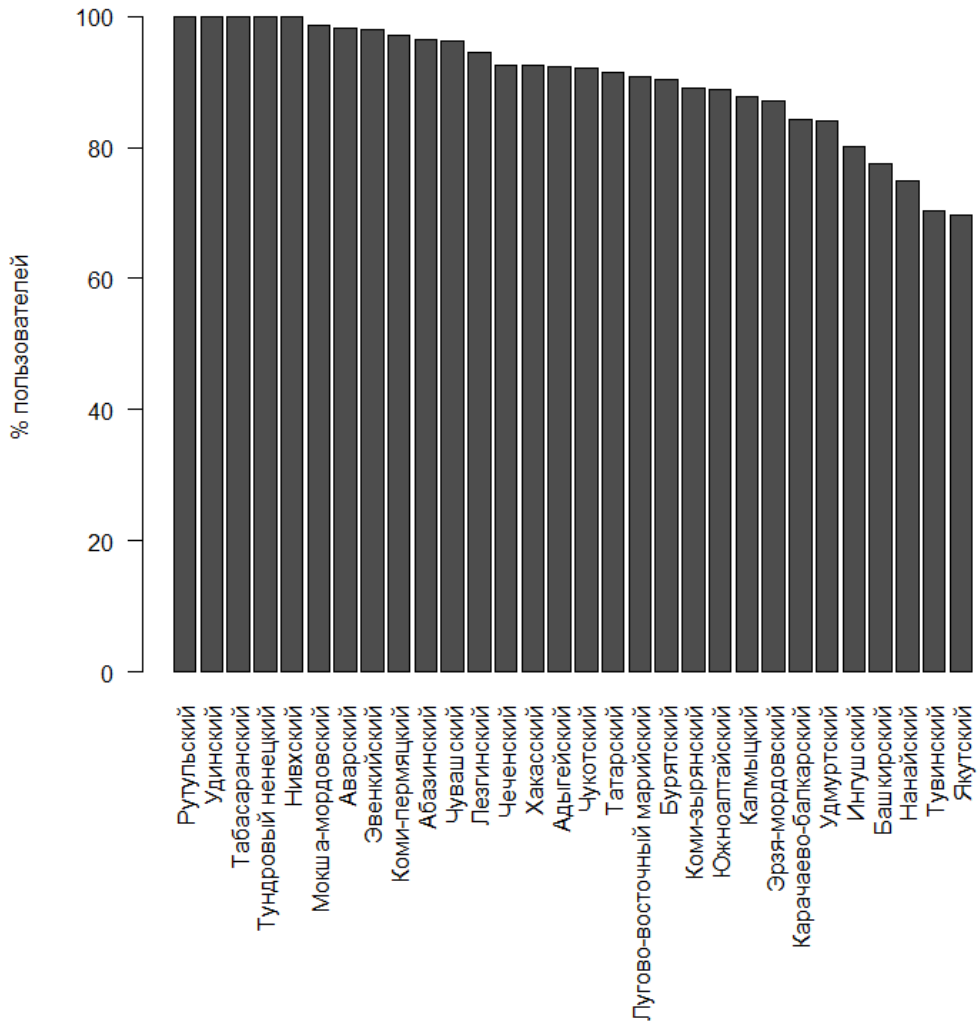

Рис. 9: Доля пользователей, принимающих участие в жизни только одного сообщества.

это возможно) нашелся хотя бы один человек, который участвует в жизни нескольких сообществ – 100% на этом графике имеют лишь языки с одним сообществом. Но таких людей мало, мы видим, что абсолютное большинство активных участников сообществ в действительности являются активными участниками только одного сообщества. Самый низкий показатель у якутского языка – 69,6%, что вполне закономерно, поскольку именно якутский является самым большим по числу пользователей языком. Мы также вычислили, как много пользователей написали всего один текст на миноритарном языке. Самые высокие и самые низкие показатели у языков с малым числом пользователей – 100% для нивхского и 25% для нанайского, в среднем же 57,4% пользователей-носителей написали только один пост на миноритарном языке. Но безусловно, есть и другая крайность: на рис. 10 показано максимальное число сообществ, в которых состоит и принимает активное участие один человек. Этот график не предназначен для сравнения языков между собой, поскольку для его построения мы использовали абсолютные

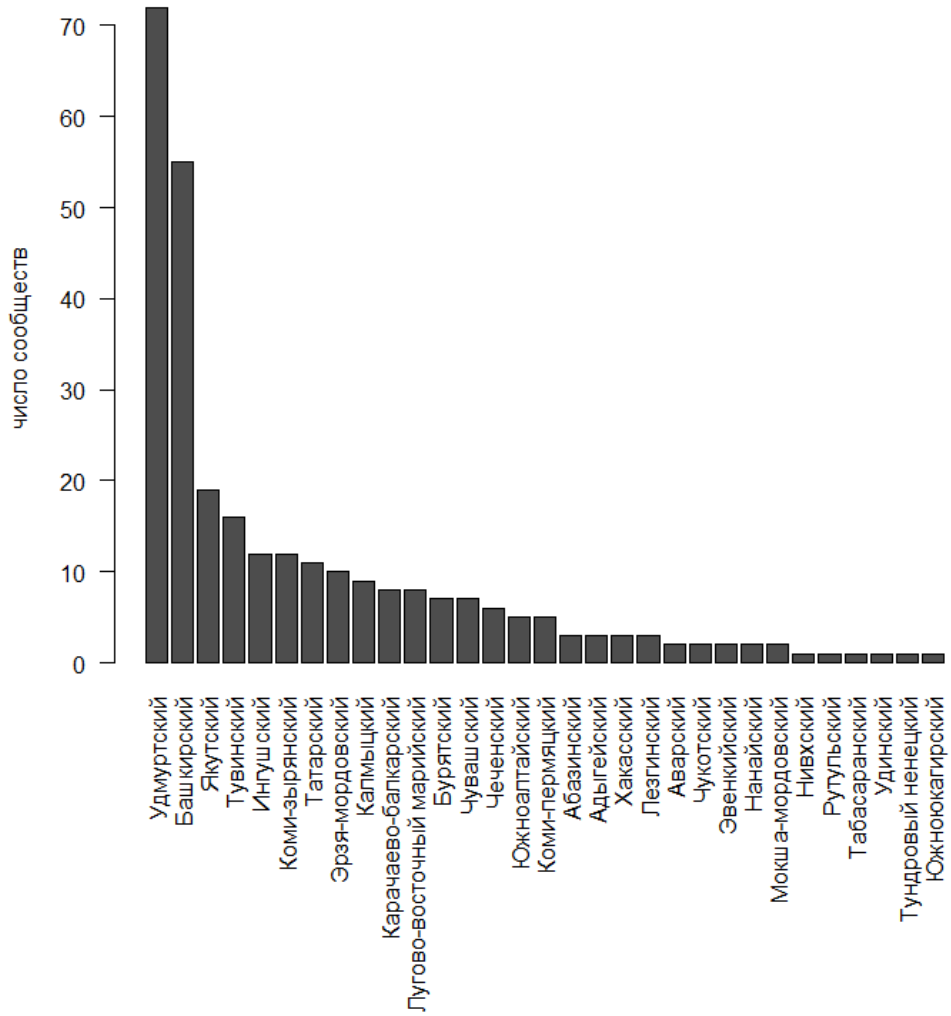

Рис. 10: Максимальное число сообществ у одного пользователя.

величины: пользователь языка с тремя сообществами просто физически не может проявлять активность в четырех сообществах, говорящих на этом языке. Являясь лидером по числу сообществ, удмуртский стал лидером и на этом графике: существует по крайней мере один пользователь-носитель удмуртского, являющийся активным участником 72 сообществ, говорящих на удмуртском. В среднем же на одного пользователя-носителя приходится около четырех текстов на миноритарном языке.

3.2. Тексты в сообществах

Мы уже несколько раз говорили, что сообщества объединяют людей. Но главной лингвистической ценностью в сообществах являются не пользователи, а порожденные ими тексты. Посмотрим на исследуемые нами языки с новой стороны – со стороны

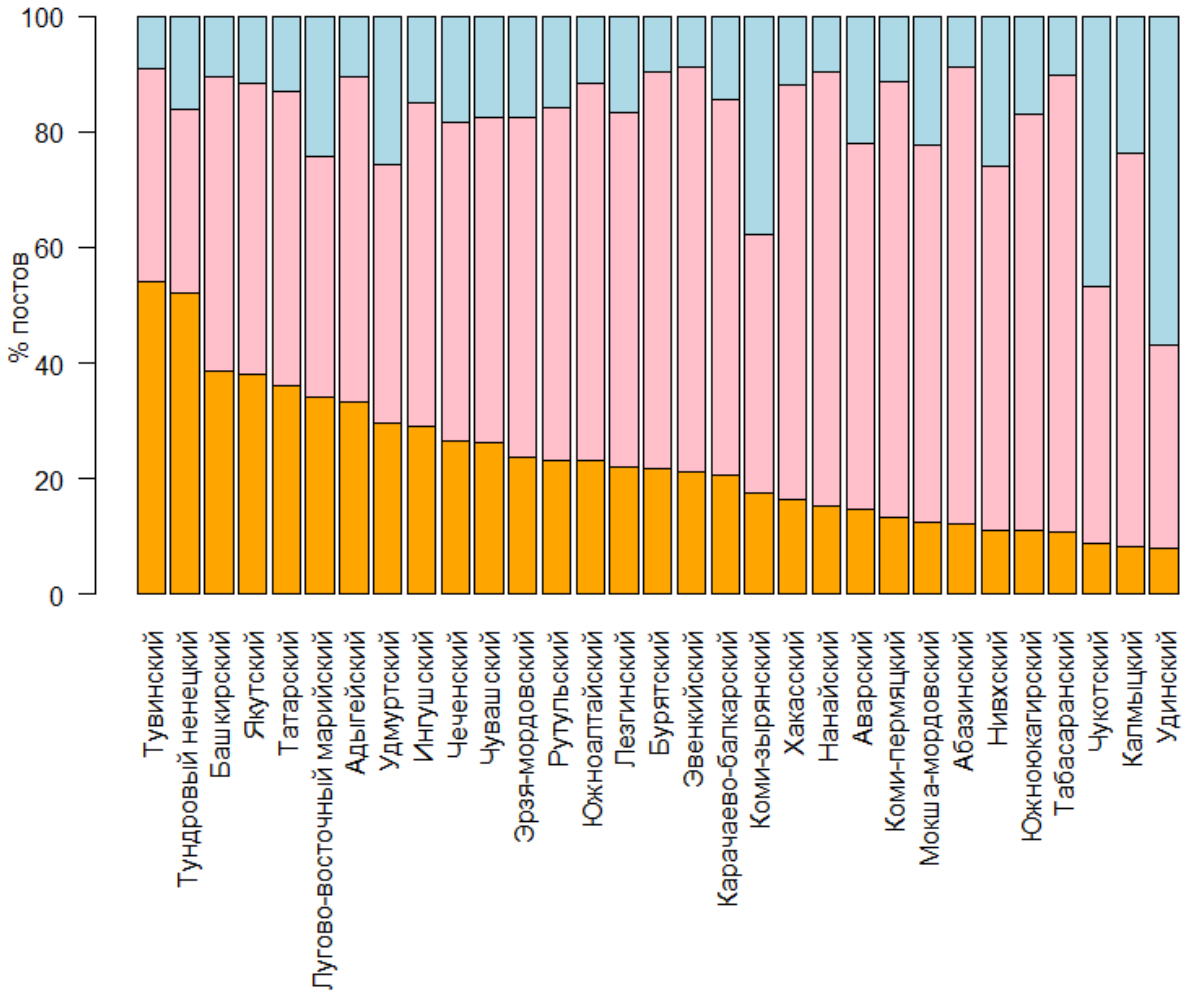

Рис. 11: Соотношение между постами, написанными на миноритарном языке (оранжевый), постами, написанными на русском языке (розовый), и мусорными постами (голубой).

текстов, на них написанных.

При описании данных мы мельком упомянули, что не все посты в исследуемых нами сообществах являются текстами на миноритарных языках: часть постов написана на русском языке – поскольку большинство носителей миноритарных языков являются также и носителями русского, а часть постов не являются текстами вовсе (исходный пост мог состоять из иллюстрации или видеоролика, которые при получении данных мы игнорировали, или же текст поста состоит исключительно из ссылок или смайликов). Все тексты были автоматически распределены на три группы: русский язык, миноритарный язык, мусор. Важным открытием для нас стало то, что в большинстве сообществ, которые мы изучаем и которые мы называем «языковыми», то есть говорящими на миноритарных языках, текстов на русском языке оказалось гораздо больше,

чем текстов на миноритарных языках. На рис. 11 показано, какую долю от всех постов для каждого из языков занимают посты на русском языке, посты на соответствующем миноритарном языке и мусорные посты. Видно, что только для двух языков – тувинского и тундрового ненецкого – постов на миноритарном языке больше, чем постов на русском. Получается, что бóльшая часть контента языковых групп в действительности состоит из русскоязычных текстов. В главе «Пользователи» мы неоднократно делали оговорку: мы работаем только с пользователями-носителями, написавшими по крайней мере один текст на миноритарном языке. Однако в каждом сообществе есть участники, которые пишут исключительно на русском, доля таких участников может достигать 70% (в сообществах удинского языка из 168 активных участников лишь 50 можно с уверенностью назвать носителями). В этой главе мы пытаемся разобраться со структурой сообществ, поэтому подобных допущений делать не будем. Так, на рис. 11 учтены абсолютно все посты в сообществах вне зависимости от языковых способностей их авторов.

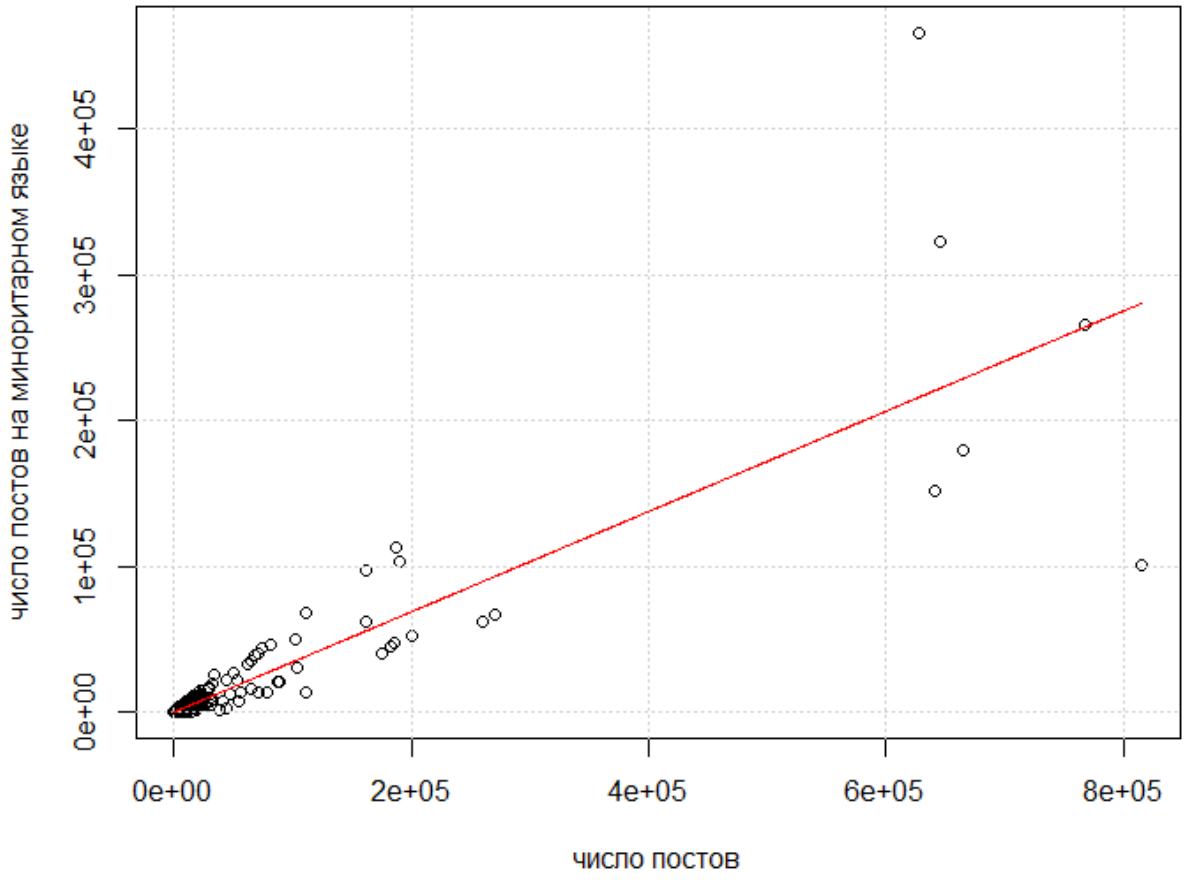

Рис. 12: Зависимость между числом постов на миноритарном языке и общим числом постов в сообществе.

Мы выяснили, что число постов на миноритарных языках меньше общего числа

постов, интересно проверить, есть ли зависимость между этими данными. Для этого мы построили модель линейной регрессии для двух переменных: число постов на миноритарн $88,47735 + 0,3440812 \cdot$ число постов. Коэффициент при переменной «число постов» значим на уровне p-value $< 0,001$, а само значение p-value равно для него $2e^{-16}$, значение p-value для свободного коэффициента модели равно 0,693 – это очень большой показатель, говорящий о том, что если этот коэффициент будет равен нулю, качество регрессионной модели не изменится. Получившаяся модель достаточно хорошая – коэффициент детерминации, характеризующий ее качество, равен 0,7625. Таким образом, мы выяснили, что зависимость между числом постов, написанных на миноритарном языке и общим числом постов в сообществе есть: примерно треть всех постов в изучаемых нами сообществах – на миноритарных языках. На рис. 12 эта зависимость проиллюстрирована. Точки на этом графике – это сообщества, координатами каждого сообщества являются два числа: общее число постов в этом сообществе (координата по оси x) и число языковых постов (ось y). Красная линия построена по формуле, соответствующей полученной нами модели: $y = 0,3440812x$. Рисунок подтверждает неидеальность полученной модели: многие точки оказались достаточно далеко от эталонной линии.

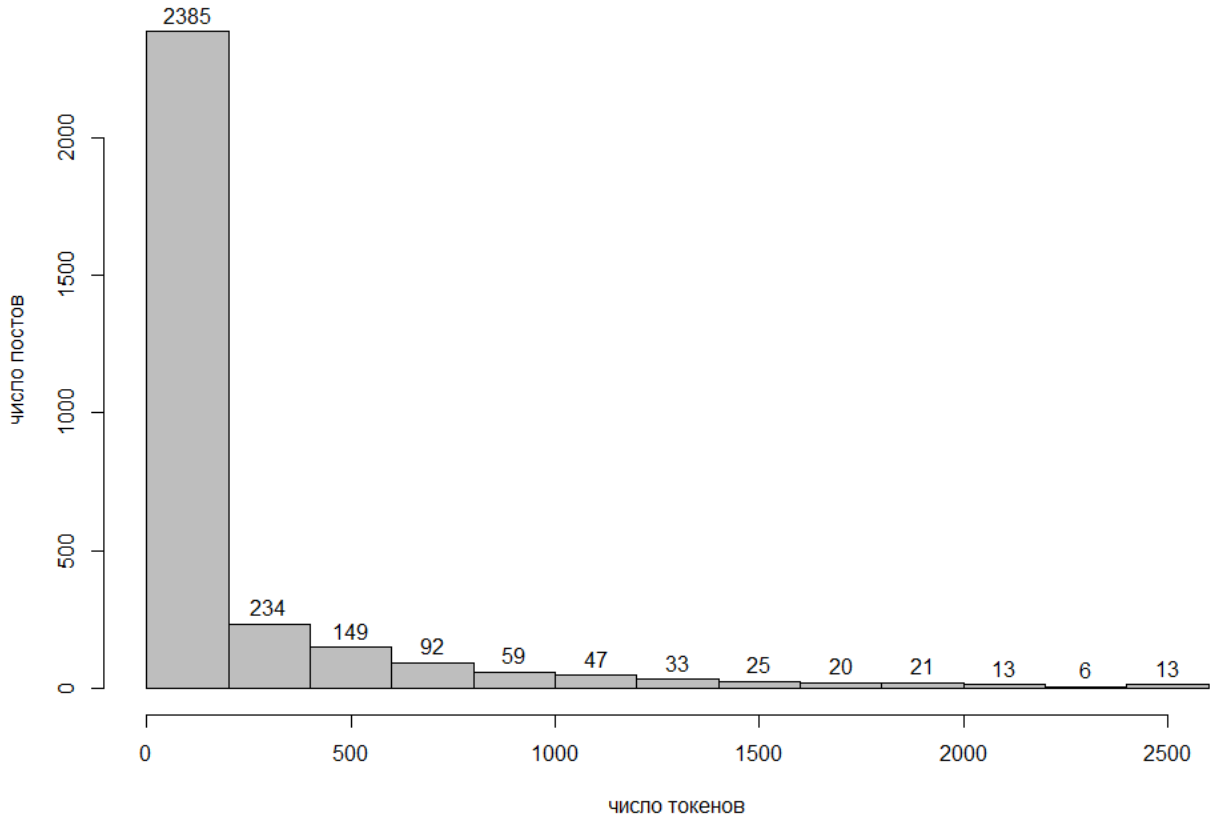

Рис. 13: Гистограмма распределения длин постов на русском языке в сообществах, говорящих по-аварски.

Важной характеристикой текста, кроме его языковой принадлежности, является

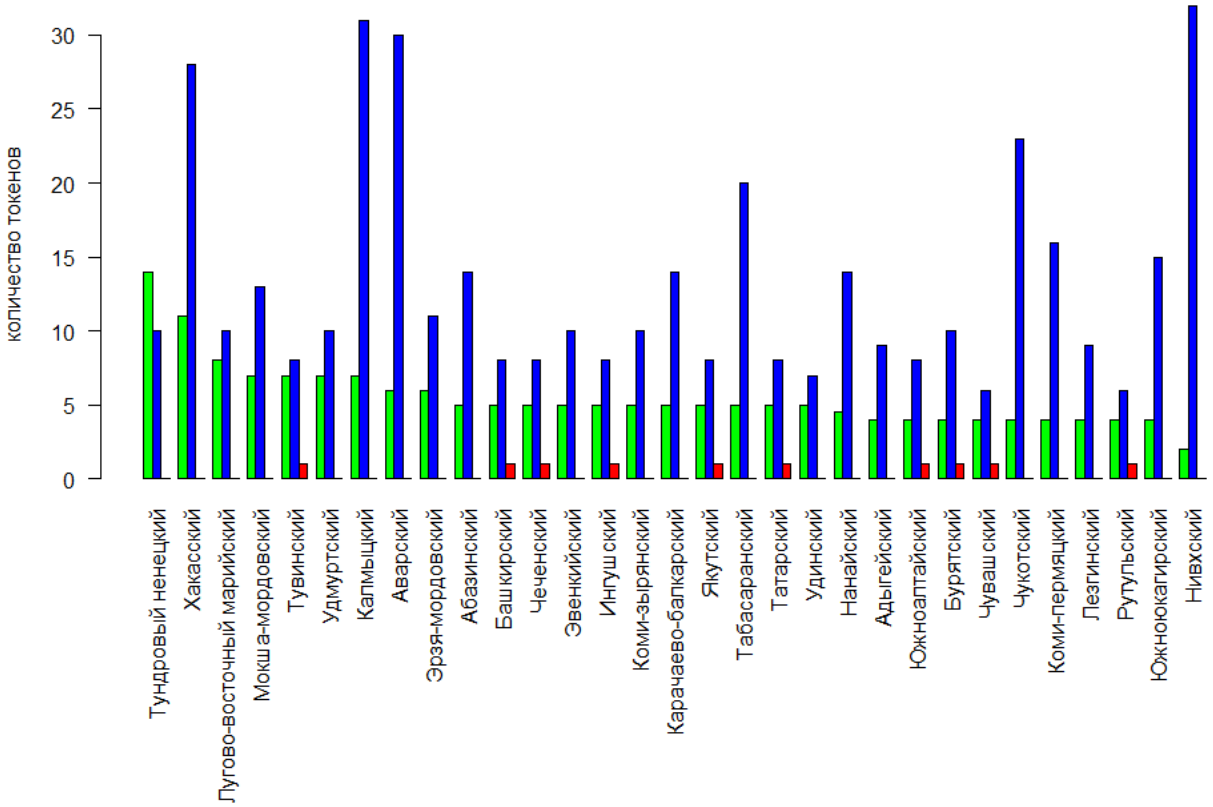

Рис. 14: Медианные значения длин постов трех типов: постов на миноритарном языке (зеленый), постов на русском языке (синий) и мусорных постов (красный).

его размер. В компьютерной лингвистике принято измерять размер текста в токенах – фрагментах текста, ограниченных пробелами. При том, что тексты в социальных сетях обычно довольно короткие, мы предполагаем, что тексты на миноритарных языках будут еще короче, чем тексты на русском языке. Такое предположение основано на сделанных ранее наблюдениях о достаточно молодом контингенте языковых сообществ и о большой доле пользователей, которые, вероятно, понимая язык, не способны к порождению на нем текстов. Чтобы сравнить длину текстов на русском и миноритарном языках, нужно собрать данные по всем текстам всех сообществ этого языка и усреднить их. Здесь нужно заметить, что распределение данных о длине постов, в отличие, например, от распределения данных о возрасте пользователей (см. рис. 5 и рис. 6), не является нормальным – см. рис. 13, на нем показано распределение постов на русском языке в аварских сообществах. Для распределений, отличных от нормального, в качестве среднего обычно берут не среднее арифметическое, использованное нами для оценки среднего возраста пользователей-носителей миноритарных языков, а медиану. На рис. 14 показаны медианные значения длины постов. Поскольку в категорию «мусорные посты» попадали в основном посты, не содержащие текста вовсе, не удивительно, что абсолютно для всех языков красный столбик, соответствующий таким

постам, на этом рисунке имеет меньшую высоту по сравнению с двумя другими столбиками, а для некоторых языков его практически не видно. Также на рисунке видно, что тексты на миноритарных языках действительно оказались короче текстов на русском языке для всех языков, кроме тундрового ненецкого. Тундровый ненецкий уже оказывался в меньшинстве: ранее мы отмечали, что в сообществах этого языка текстов на русском языке меньше, чем текстов на тундровом ненецком (см. рис. 11). Посмотрим на данные, имеющиеся у нас для этого языка: он имеет только одно сообщество Вконтакте, в котором всего 25 постов, из них 8 на русском языке, 4 мусорных поста и 13 постов на миноритарном языке. Вероятно, среди сообществ других исследуемых нами языков есть похожие сообщества: с большим числом длинных текстов на миноритарном языке, однако в процессе усреднения эти сообщества потеряли свое преимущество под давлением менее языковых, но бóльших по числу постов сообществ. Возвращаясь к рис. 14, заметим, что значения, отмеченные на оси ординат и соответствующие медианному значению длины текста, в целом очень небольшие: около половины всех текстов на всех миноритарных языках короче 5 слов.

3.3. Тематическое моделирование

Результаты, представленные в этом разделе, были получены при содействии И. Крыловой, участницы проекта «Языки России» (*Сайт проекта «Языки России»* б.г.).

Кажется, что для определения темы любого разговора, как минимум, необходимо понимать язык, на котором этот разговор ведется. Мы постараемся без знания миноритарных языков попробовать разделить сообщества на несколько групп так, чтобы каждой группе соответствовала определенная тематика. Для каждого языка мы построим граф, в вершинах которого — сообщества, говорящие на этом языке, дуги между вершинами должны будут показывать, насколько сообщества, соединенные одной дугой, тематически близки друг к другу.

Каждое сообщества языка было представлено нами в виде частотного списка токенов (или словоформ), составленного по всем текстам, написанном в этом сообществе на миноритарном языке. Ожидая, что в верхушке таких списков окажутся частотные для языка слова, абсолютно не влияющие на тематику текстов (обычно лидерами в частотных списках для разных языков являются предлоги, союзы и местоимения), мы удалили 10% самых частотных слов из каждого списка. Для сравнения сообществ между собой нам необходимо перейти от представления в виде частотных списков к векторному представлению. Каждое сообщество мы представим в виде вектора, длина этого вектора должна быть равна сумме всех токенов во всех сообществах. В компонентах вектора,

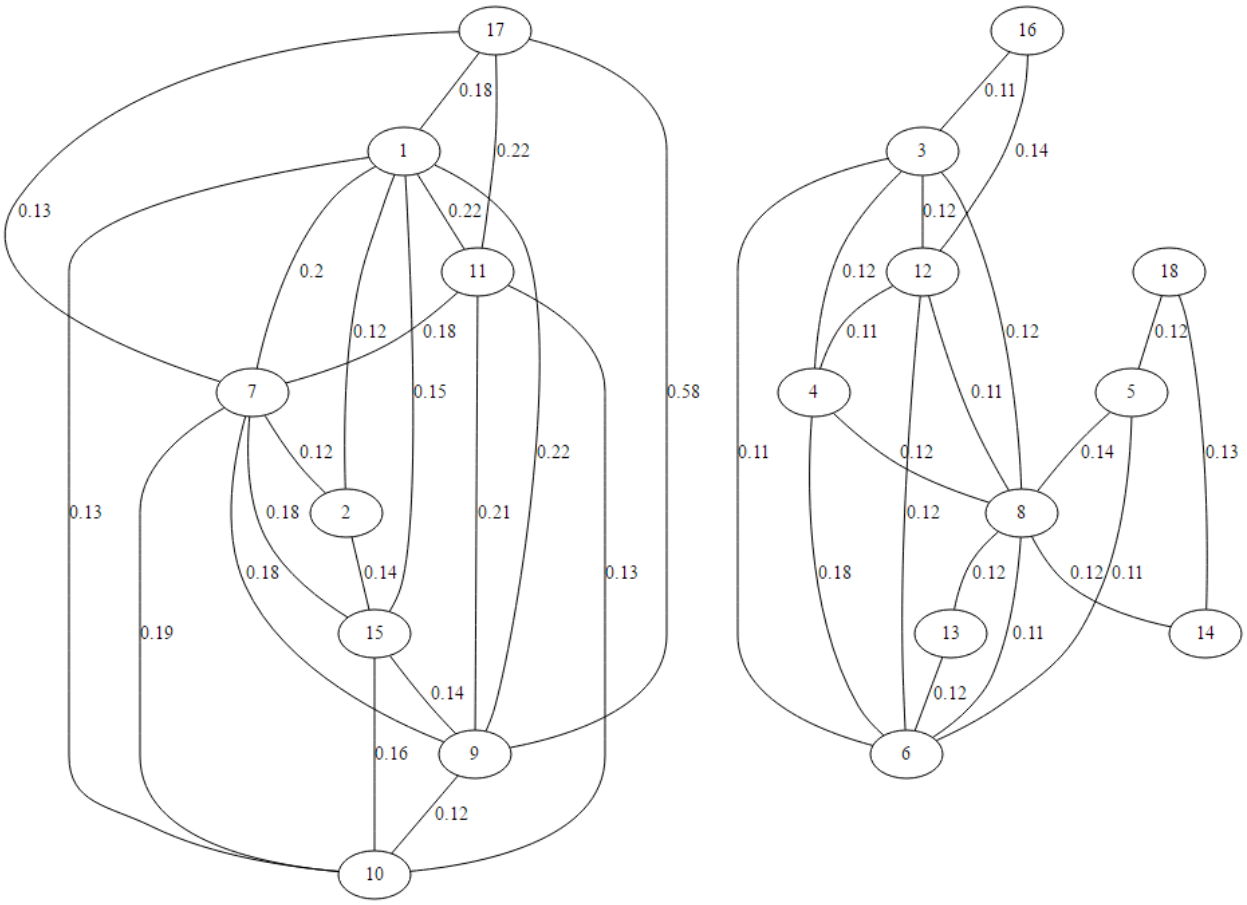

Рис. 15: Граф распределения по тематике сообществ, говорящих по-чеченски.

соответственно, отражен вес каждого токена в сообществе. В качестве меры близости между полученными векторами мы использовали косинусную меру сходства. На рис. 15 показан граф, построенный нами для сообществ, говорящих по-чеченски. В вершинах идентификаторы сообществ, веса на ребрах графа – полученная косинусная мера близости между сообществами. На рисунке мы изобразили все ребра с весом, большим, чем 0,1, и все сообщества, имеющие по крайней мере одно ребро. Видно, что все чеченскоговорящие сообщества разделились на две группы, как мы рассчитываем, объединенные одной тематикой. Поверхностного изучения контента сообществ из каждого кластера оказалось достаточно, чтобы определить, что сообщества, оказавшиеся в правом кластере, посвящены религии, в то время как основным контентом сообществ из левого кластера являются шутки и «мемы». Мы построили подобные графы для всех исследуемых языков. Все графы оказались двух типов: состоящими из двух кластеров (подобно графу для чеченского языка) и состоящими из одного кластера. По-настоящему понять, действительно ли в сообществах на миноритарных языках России говорят об одном и том же, можно с привлечением специалистов по каждому из исследуемых языков.

4. Языки

Все, о чем мы говорили до этого, в той или иной степени сводилось к миноритарным языкам: мы исследовали географию пользователей и их возраст, объем написанных текстов и количество посещаемых сообществ, но после этого мы каждый раз сравнивали между собой языки, пытаясь понять, насколько они похожи друг на друга. Пришло время объединить всю полученную к этому моменту информацию воедино. Но прежде постараемся собрать такие данные о миноритарных языках, которые нельзя получить, анализируя исходную коллекцию сообществ из социальной сети Вконтакте. В действительности, таких чисто языковых данных очень мало.

Как мы уже упоминали раньше, в нашем распоряжении есть число носителей для каждого из языков. Эти данные были опубликованы после проведения последней Всероссийской переписи населения в 2010 году, предыдущая перепись проводилась в 2002 году, и ее результаты также опубликованы (*Портал «Всероссийская перепись населения 2002 года»* 2002). Число носителей языка в 2002 году кажется гораздо менее ценной информацией – для миноритарных языков 14 лет – довольно большой срок, за это время какие-то языки могли полностью лишиться носителей. Однако из этой информации можно попробовать извлечь пользу. Для каждого языка мы вычислили процент, на который увеличилось число его носителей за 8 лет, прошедших между двумя последними переписями. Это важный показатель того, насколько быстро растет язык. К сожалению, оказалось, что подавляющее число языков наоборот – уменьшились. Так, из 43 анализируемых нами языков у восьми языков не оказалось нужной нам информации, за один из двух годов или за оба. Среди оставшихся 35 языков лишь семь имеют положительный прирост числа носителей. Любопытно, что шесть языков из этих семи являются языками Кавказа, исключение составляет тувинский язык, являющийся официальным языком в республике Тыва, он получил прирост числа носителей в 4,4%.

Полезной информацией о языке были бы данные о числе выпускаемых на языке книг, журналов, научных статей и пр. Такие данные можно получить из электронной энциклопедии Википедии (*Электронная энциклопедия Википедия* б.г.), к сожалению такая информация есть не для всех языков и далеко не для каждого года. Нам удалось получить: для 27 языков количество выпускаемых на языке книг за 2009 год, для 17 языков количество журналов, выпускаемых в 1996 году и количество выпускаемых газет за этот же год – для 20 языков. Кроме того, для 18 языков нам известно количество детских садов для детей-носителей миноритарного языка. Еще одним важным параметром является наличие у языка своей Википедии и количество статей в ней. А. Корнай в (Kornai 2015) утверждает, что Википедия – необходимое, но не достаточное

условие для того, чтобы язык процветал в сети, а потом добавляет: «no wikipedia, no survival» – нет Википедии, нет шансов выжить. В дальнейшем нам предстоит подтвердить или опровергнуть этот тезис. Кроме описанной выше, никакой другой информации о языках у нас нет.

Однако есть ряд параметров, не относящихся к языкам напрямую. Это параметры, характеризующие регионы РФ. Строго говоря, взаимно однозначного соответствия между миноритарными языками и регионами России нет: как в одном регионе могут пользоваться несколькими языками, так и один язык может быть распространен на территории нескольких регионов. Тем не менее, мы постарались присвоить каждому языку только один регион, в котором проживает наибольшее число носителей. Про каждый регион мы можем получить данные о населении, в том числе об уровне рождаемости и смертности; о доходах региона в целом и зарплатах его жителей; об интернетизированности и компьютеризированности населения.

4.1. Преобразование данных

После того, как мы объединили все имеющиеся данные, у нас получилась таблица из 65 колонок, соответствующих разным характеристикам языков. Но работать с ними в «чистом виде» нельзя, часть из них требуют дополнительной обработки и преобразований.

Во-первых, в наших данных довольно много «белых пятен» – то есть пропущенной информации. Есть разные способы избавления от пропусков в данных, зависящие от типа этих данных. Один из способов заключается в замене всех пропусков средним значением по выборке. Такой способ подходит, например, для данных о возрасте носителей. В наших данных сразу два столбца содержат информацию о возрасте: средний возраст пользователей сети Вконтакте и средний возраст носителей языка. В обоих наборах данных есть пропуски, которые мы заменим на средние значения по столбцам. Особенность данных из столбца со средним возрастом пользователей заключается в том, что кроме «хороших данных» и пропусков, в нем есть нули. Пропуски в данных – это не то же самое, что нули: например, информация о том, что для шорского языка нет ни одного сообщества Вконтакте, не является пропуском, а информации о среднем возрасте носителей для этого языка у нас нет – это пропуск. В случае с возрастом пользователей нули в качестве значений получили языки, не представленные в сети Вконтакте, пропуски же оказались у языков, все пользователи которых не указали дату рождения. Также средними значениями мы заменили пропуски в столбцах с информацией о распространенности интернета в регионах РФ и о числе компьютеров на семью

по регионам.

В наших данных много пропусков в графе о числе носителей языков в 2002 году, а как следствие – несмотря на то, что данные о числе носителей в 2010 году есть для всех языков, вычислить прирост числа носителей в период с 2002 по 2010 года мы можем не для всех языков. Чтобы восстановить пропуски в этом случае, мы воспользуемся знанием о числе носителей в 2010 году и методом замены пропущенных значений средним значением его соседей. Для этого необходимо упорядочить данные за 2010 год, а затем каждый пропуск в данных за 2002 год заполнить средним значением от двух его соседей. В таблице 1 показан фрагмент наших данных, видно, что в них есть пропуск – нам неизвестно число носителей кубачи-аштинского языка в 2002 году. С помощью данных о его соседях – Чукотском и Эвенском языках – мы постараемся заполнить этот пропуск: $(7742 + 7168)/2 = 7455$ – именно это число будет вписано в ячейку с вопросительным знаком. Аналогичным образом заполним пропуск в данных о доходе на душу населения в 2002 году.

Таблица 1: Восстановление пропуска методом среднего значения соседей.

Язык	Число носителей, 2002	Число носителей, 2010
Эвенкийский	7584	4800
Чукотский	7742	5095
Кубачи-аштинский	?	5200
Эвенский	7168	5656

Кроме избавления от пропусков, наши данные нуждаются еще в одном типе преобразований. Среди наших данных есть не только числовые признаки, но и признаки, выраженные словами. Например, среди наших данных есть информация о статусе языка: язык может быть государственным, такой статус получают языки, использующиеся на территории республик России наряду с русским языком; язык может быть официальным, этот статус может быть присвоен языкам, использующимся в автономных округах или в местах проживания говорящего на нем народа; у части миноритарных языков РФ нет ни государственного, ни официального статуса (Доровских 2007). Данные о статусе довольно важны: по статусу языка можно понять, насколько он распространен и поддерживаем правительством титульного региона. Характеристику «статус языка» из текстовой мы переведем в шкальную, установили такой порядок превосходства: нет статуса→официальный язык→государственный язык. Еще часть столбцов мы исключим из дальнейшего анализа, например, названия и самоназвания языков – нам достаточно только одного параметра идентифицирующего язык, им будет индиви-

дуальный iso-код.

4.2. Поиск связей

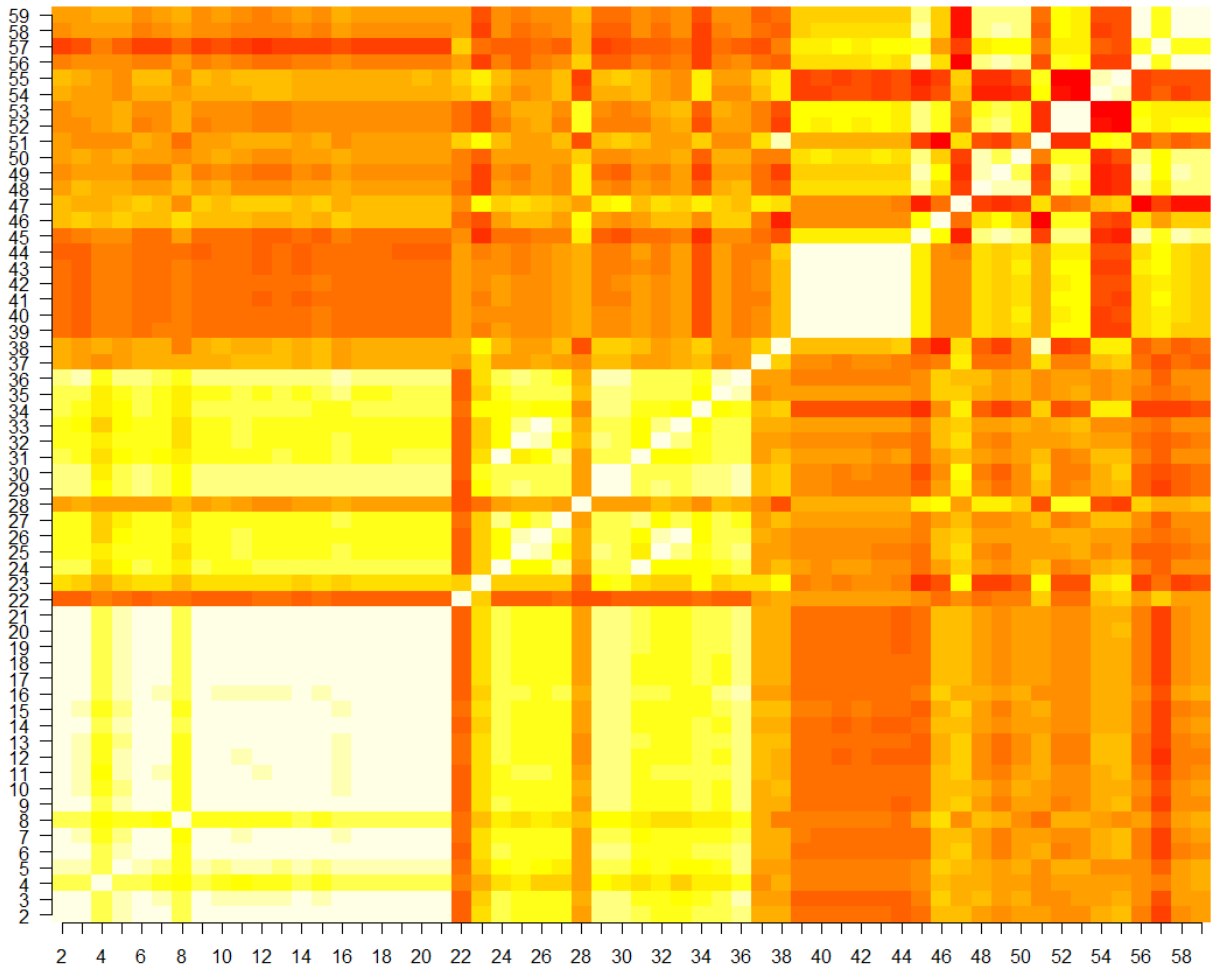

Рис. 16: Тепловая карта коэффициентов корреляции.

После всех преобразований у нас осталось 58 признаков. Это довольно большое число, для начала попробуем окинуть все признаки быстрым взглядом. Для всех возможных пар признаков мы проверили наличие между ними корреляции, или линейной связи. Мы использовали метод корреляции Спирмена, который хорошо приспособлен для данных, распределение которых не близко к нормальному, а кроме того он меньше, чем, например, метод корреляции Пирсона, подвержен влиянию случайных выбросов в данных. На рис. 16 представлена тепловая карта получившихся коэффициентов корреляции. Числа на осях – идентификаторы признаков, цвет ячейки говорит о размере коэффициента корреляции между признаками, соответствующими строке и столбцу, образующим эту ячейку. Тепловые карты устроены так, что числовые значения на них

закодированы цветами, в нашем случае соответствие такое: максимальный коэффициент корреляции, равный единице, закодирован белым цветом (см. побочную диагональ на карте, показывающую корреляцию каждого признака с самим собой), отсутствие линейной зависимости или нулевой коэффициент корреляции закодирован красным цветом. Высокий коэффициент корреляции говорит о силе связи между признаками, но не о природе этой связи: признаки могут зависеть друг от друга, один может определять другой, или же оба признака зависят от третьего признака. На этой карте есть несколько светлых пятен – одно из них находится в левом нижнем углу, видно, что признаки с индексами 2 – 21 очень хорошо коррелируют друг с другом. Именно эти 20 признаков были получены нами в результате анализа выкачанных сообществ из социальной сети Вконтакте. В меньшей степени связанными с другими признаками оказались признак 4, обозначающий для языка, с каким количеством других языков он имеет общих носителей (степень узла графа на рис. 7), и признак 8 – средний возраст пользователей социальной сети Вконтакте. Интересно, что значения этого признака достаточно слабо коррелируют со средним возрастом носителей языка, полученным из переписи населения (признак 28). Второе светлое пятно на карте показывает высокий уровень корреляции между признаками 39 – 44. Эти шесть признаков характеризуют состояние бюджета титульного для языка региона: три из них содержат информацию о доходах за 2010, 2011 и 2014 года, а еще три – о расходах за те же года. Корреляция между этими признаками абсолютно закономерна. Никаких по-настоящему неожиданных и любопытных зависимостей в наших данных не обнаружено.

Выше мы цитировали А. Корная, утверждающего, что без Википедии языку в интернете не выжить. Среди наших данных есть два признака, характеризующих представленность языка в Википедии – это число статей в Википедиии в 2014 и в 2015 годах, на рис. 16 это признаки 35 и 36. Видно, что корреляция этих признаков между собой довольно высокая, но ни один из них не коррелирует достаточно сильно ни с одним из первых двадцати признаков (полученных по данным из Вконтакте). В Приложении 1 представлена таблица для всех языков со следующими признаками: количество сообществ для этого языка, количество постов, написанных на этом языке, число пользователей-носителей, количество статей в Википедии в 2014 и 2015 годах. Мы отсортировали данные по увеличению мощности Википедии в 2015 году. Видно, что ни один из четырех языков, не имеющих статей в Википедии, не представлен в социальной сети Вконтакте. Среди следующих семи языков, имеющих от одной до трех статей в Википедии, только два нам удалось обнаружить Вконтакте.

В (Kornai 2015) приводится таблица с данными о жизнеспособности языков, распространенных на территории бывшего Советского Союза. Оценка жизнеспособности

сформирована на основе количества и качества статей в Википедии на соответствующем языке и выставлена по шкале (Kornai 2013). В своей работе мы не стремились оценить анализируемые языки по шкале Корная, тем не менее, мы можем утверждать, что среди 43 исследуемых языков 12 являются неподвижными в социальной сети Вконтакте. В упомянутой выше таблице постсоветских языков из этих 12 присутствуют 8, причем один из них – даргинский – получил оценку живой (V), в то время как оценка по остальным семи языкам совпала . Любопытно, что к неподвижным языкам Корнай отнес 6 языков, на которых Вконтакте нашлись тексты. В таблице 2 представлены некоторые данные, полученные из сети Вконтакте, для этих языков (более полная таблица в Приложении 1). Звездочкой мы пометили языки, которые в главе «Пользователи» мы назвали «подозрительными», а на рис. 7 – выделили красным цветом. Как видно в таблице, у всех шести языков действительно малое число текстов и пользователей, но мы надеемся, что по крайней мере три языка из шести еще рано причислять к замершим.

Таблица 2: Данные о шести неподвижных языках.

Язык	Число сообществ	Число постов	Число пользователей
Абазинский	5	15297	1959
Хакасский	4	815	107
Удинский	1	188	50
Нанайский*	1	14	8
Табасаранский*	1	19	9
Южноюкагирский*	1	11	3

Мы уже назвали 12 языков из исследуемых неподвижными на основании того, что тексты на них не встретились Вконтакте. Попробуем разделить ставшиеся языки на группы по схожести – кластеры. Мы воспользуемся методом иерархической кластеризации, предложенном Уордом. Все методы иерархической кластеризации построены по такому принципу: все исследуемые объекты изначально – это отдельные кластеры, далее наиболее близкие кластеры объединяются в один кластер и так далее. Различие разных методов состоит в критериях, по которым определяется близость кластеров. По методу Уорда в качестве такой меры близости выступает получаемый при объединении кластеров прирост суммы квадратов евклидовых расстояний от объектов до центров кластеров. В качестве объектов у нас выступают минритарные языки, в качестве признаков – характеристики, полученные нами в результате анализа данных из сети Вконтакте. Мы не использовали другие характеристики языка, чтобы оценить языки с точки зрения их жизнеспособности Вконтакте. Получившаяся в результате класте-

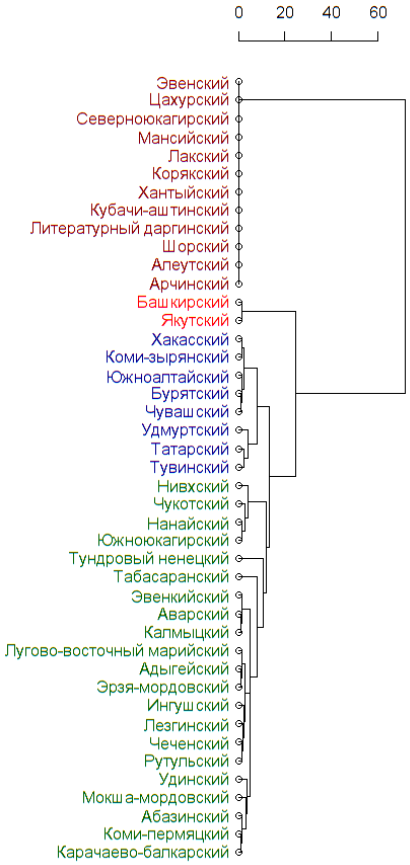

Рис. 17: Результат иерархической кластеризации языков.

ризации дендрограмма – на рис. 17. Получившиеся четыре кластера можно выстроить по шкале жизнеспособности в сети Вконтакте: темно-красный кластер попали 12 неподвижных языков, языки из зеленого кластера Вконтакте есть, но по-настоящему живыми эти языки назвать нельзя (данные о языках см. в Приложении 1), на них мало текстов, у них мало пользователей, часть из них мы уже называли «подозрительными», а в терминах анализа их можно назвать главными кандидатами на перемещение в первый – темно-красный кластер, в синем кластере – большие живые и процветающие языки, так же можно охарактеризовать и языки из красного кластера.

5. Заключение

В этой работе мы рассмотрели 43 миноритарных языка России. В качестве данных для изучения мы использовали коллекции текстов с метаинформацией о самих текстах и об их авторах, собранные из социальной сети Вконтакте. В действительности, у нас была возможность проанализировать лишь 31 из этих языков – именно столько языков оказались представленными в социальной сети.

Оказалось, что с точки зрения языкового поведения пользователи «больших» языков не сильно отличаются от пользователей «малых». Типичный пользователь Вконтакте, носитель миноритарного языка России, – это человек в возрасте от 19 до 31 года, проживающий в титульном регионе для своего языка. Являясь носителем миноритарного языка, намного лучше и чаще говорит на русском. Вероятнее всего, активно принимает участие в жизни лишь одного сообщества, говорящего на миноритарном языке. Для языкового сообщества характерно то, что действительно языковых постов в нем лишь треть, и посты эти преимущественно короткие – около пяти слов.

В России около ста миноритарных языков, около пятидесяти из них находятся в критическом состоянии или в серьезной опасности (Мозли 2010). Некоторые исследователи полагают, что интернет вообще и набирающие популярность социальные сети в частности способны вдохнуть новую жизнь в эти исчезающие языки. Проанализировав небольшой сегмент интернета, говорящего на миноритарных языках, мы с сожалением должны заметить: по-настоящему активно и много пишут носители языков, не находящихся под угрозой исчезновения. Наиболее процветающими языками в социальной сети Вконтакте, по нашим данным, являются Башкирский и Якутский языки, находящиеся на второй ступени сохранности из шести по шкале UNESCO.

5.1. Возможные дальнейшие исследования

Исследование нельзя считать в полной мере оконченным. Мы провели поверхностный анализ имеющихся данных и попытались понять, есть ли разница между тем, как проявляют себя в сети Вконтаке языки с малым числом носителей и языки с большим. В процессе работы, например, было выяснено, что многие носители миноритарного языка в действительности являются носителями нескольких языков, мы выделили четыре возможные причины, это объясняющие. Получившийся в результате подсчета пересечений языков по пользователям граф изображен на рис. 7. Отдельным исследованием мог бы стать детальный анализ этого графа с привлечением специалистов по соот-

ветствующим языкам. Кроме того, в работе остался неохваченным временной аспект – в исходной коллекции есть информация о времени создания постов, используя эту информацию можно изучать всплески активности сообществ разных языков или вычислять их жизненный цикл.

Список литературы

Benevenuto F., *Rodrigues T.*, *Cha M.*, *Almeida V.* Characterizing user behavior in online social networks // Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. — ACM. 2009. — С. 49—62.

Bortzmeyer S. Multilingualism and the internet's standardisation // Net. Lang: Towards the multilingual cyberspace. — 2012. — С. 106—117.

Cavnar W. B., *Trenkle J. M.* [и др.] N-gram-based text categorization // Ann Arbor MI. — 1994. — Т. 48113, № 2. — С. 161—175.

Cormack M. J., *Hourigan N.* Minority language media: Concepts, critiques and case studies //. Т. 138. — Multilingual Matters, 2007. — Гл. Introduction: Studying Minority Language Media. С. 1.

Katzner K. The languages of the world. — Routledge, 2002. — С. 9.

Kornai A. A New Method of Language Vitality Assessment // Linguistic and Cultural Diversity in Cyberspace. — 2015. — С. 132.

Kornai A. Digital language death // PloS one. — 2013. — Т. 8, № 10. — e77056.

Lewis M. P., *Simons G. F.* Assessing endangerment: expanding Fishman's GIDS // Revue roumaine de linguistique. — 2010. — Т. 2. — С. 103—119.

Prado D. Language presence in the real world and cyberspace // Net. Lang: Towards the multilingual cyberspace. — 2012. — С. 34—51.

Zaydelman L., *Krylova I.*, *Orekhov B.*, *Stepanova E.* Languages of Russia: Using Social Networks to Collect Texts // Proceedings of the 9th Summer School in Information Retrieval and Young Scientist Conference (RuSSIR 2015) – Revised and Selected Papers. — in print.

База стран и городов citieslist. — URL: `http://citieslist.ru/`.

Демьяненко И. Статистика по профилям пользователей ВКонтакте // [Электронный ресурс]. — 2011. — URL: `https://habrahabr.ru/post/123856/`.

Денисова А. В. Языковая ситуация в республике Калмыкия // Теория и практика общественного развития. — 2015. — № 24. — С. 92—95.

Доровских Е. М. К вопросу о разграничении понятий «государственный язык» и «официальный язык» // Журнал российского права. — 2007. — 12 (132).

Мозли К. Атлас языков мира под угрозой исчезновения. — [Электронный ресурс] : Париж, Издательство Юнеско, 2010. — URL: `http://www.unesco.org/languages-atlas/en/atlasmap.html`.

Пишлёгер К. Удмуртский и бесермянский языки в социальных сетях // Наука, просвещение, искусство провинции в социокультурном пространстве: Девятые Короленковские чтения: Материалы Международной научно-практической конференции,

посвященной 260-летнему юбилею В.Г. Короленко. (28 октября 2013 года). — 2013. — С. 187—190.

Портал «Всероссийская перепись населения 2002 года». Том 4. Национальный состав и владение языками, гражданство. — 2002. — URL: http://www.perepis2002.ru/index.html?id=17.

Портал «Всероссийская перепись населения 2010 года». Том 4. Национальный состав и владение языками, гражданство. — 2010. — URL: http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm.

Сайт проекта «Языки России». — URL: http://web-corpora.net/minorlangs/.

Филь М. Влияние феномена «социальных сетей» на процессы самоорганизации общества // НИИ Социологии [Электронный ресурс]. — 2012. — URL: http://niisocio.%20ru/press-centr/nauchnye-stati/142-vliyaniefenomena-sotsialnykh-setej-na-protsessy-samoorganizatsii-obshchestva.

Электронная энциклопедия Википедия. — URL: http://ru.wikipedia.org.

6. Приложения

6.1. Приложение 1

Таблица 3: Данные о представленности языков в Википедии и Вконтакте.

Язык	**Сообщества**	**Посты**	**Пользователи**	**Вкипедия 2014**	**Вкипедия 2015**
Алеутский	0	0	0	0	0
Арчинский	0	0	0	0	0
Кубачи-аштинский	0	0	0	0	0
Северноюкагирский	0	0	0	0	0
Нанайский	1	14	8	0	1
Мансийский	0	0	0	0	1
Даргинский	0	0	0	0	2
Эвенский	1	0	0	0	2
Хантыйский	0	0	0	0	2
Цахурский	0	0	0	0	3
Удинский	1	188	50	0	3
Абазинский	5	15297	1959	0	4
Нивхский	1	3	2	0	4
Рутульский	1	20208	866	0	4
Табасаранский	1	19	9	0	5
Корякский	0	0	0	0	9
Южноюкагирский	1	11	3	0	11
Шорский	0	0	0	0	16
Эвенкийский	5	2312	301	0	18
Чукотский	5	73	38	0	47
Южноалтайский	9	44897	5252	0	51
Хакасский	4	815	107	0	54
Тундровый ненецкий	1	13	6	0	65
Адыгейский	5	41927	3705	0	241
Ингушский	97	155585	15280	0	500
Мокша-мордовский	4	127	71	1181	1136
Лакский	0	0	0	1201	1208
Тувинский	168	363143	30064	421	1230
Бурятский	61	68583	9637	939	1430

Продолжение на следующей странице

Язык	**Сообщества**	**Посты**	**Пользователи**	**Вкипедия 2014**	**Вкипедия 2015**
Карачаево-балкарский	17	14758	2325	1932	1997
Аварский	7	729	171	1126	1999
Калмыцкий	57	6628	1219	1871	2017
Эрзя-мордовский	26	5393	681	1642	2368
Лезгинский	12	10168	1886	1937	2619
Коми-пермяцкий	11	3189	643	3429	3421
Удмуртский	298	63639	5746	3389	3624
Коми-зырянский	47	10055	1262	4164	4392
Лугово-восточный марийский	59	4159	895	4007	7510
Якутский	268	1056873	67967	10287	10445
Чувашский	96	72563	5790	28247	33550
Башкирский	294	1113948	66389	32008	34983
Татарский	147	462137	33671	57033	67196
Чеченский	22	49291	8379	23138	100401