ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

Обнаружение тематической связности новостной повестки и наивной поэзии

Detection of topical connections between the news agenda and naive poetry

Студентка 3 курса
группы № 182
Панасюк А. В.

Научный руководитель
Орехов Б. В.
доцент
кандидат филологических наук

Москва, 2021 г.

Оглавление:

1. Введение

С ростом числа пользователей русскоязычного интернета исследователям стали доступны огромные массивы текстов, прежде остававшихся невидимыми для литературной традиции. До этого исследованием неопубликованных (или опубликованных малым тиражом) литературных работ занимались скорее фольклористы, изучавшие локальные сюжеты [Лурье 2001]. Сейчас же есть как количественные исследования [Бонч-Осмоловская, Орехов 2017, Орехов 2013], так и качественные [Князева 2015, Петров 2015, Югай 2016], и многие из них используют материалы, опубликованные на сайте stihi.ru.

Говоря о таких текстах, не являющихся частью профессиональной литературной традиции, используют термин “наивный” в понимании [Неклюдов 2001]. В этой курсовой работе мы обращаемся к наивной поэзии в контексте количественных исследований. Наследуя [Бонч-Осмоловская, Орехов 2017], мы считаем стихотворения, опубликованные на сайте stihi.ru, наивной поэзией, и говорим о ней, как о “творческой продукции непрофессиональных литераторов, создающих свои произведения с оглядкой на высокие образцы, но неспособных сочинить что-то равновеликое им по качеству”.

1

В рамках этой курсовой мы также обращаемся к порталу stihi.ru. Ресурс был создан в марте 2000 года[1], и с тех пор продолжает неуклонно расти. На сайте отсутствует модерация, и любой человек может опубликовать свое произведение мгновенно. На данный момент (май 2021), на сайте опубликовано 52 660 279 записей[2]. Ресурс живет активной жизнью, в 2013 году публиковалось 7-9 тысяч записей в день.

Многие наивные авторы откликаются на новостные темы. Традиция написания стихотворных откликов на актуальные события восходит в России как минимум к XVIII веку («Ода на Взятие Хотина» М. В. Ломоносова) [Лейбов, Орехов (in print)]. Целью этой курсовой работы является создание инструмента, позволяющего проследить современное состояние традиции. Из-за объемов текста это возможно только с применением компьютерных методов обработки естественного языка. Результатом такой работы является список обнаруженных наивных новостных стихотворений, список новостных тем, на которые они были написаны, и описание распределения тем в наивном поэтическом пространстве. В рамках курсовой работы для отработки метода мы рассматриваем стихотворения, опубликованные в 2013 году.

Задачами работы являются:

1) сбор и обработка корпусов новостей и стихотворений;

[1] О портале Стихи.ру [Электронный ресурс]. URL: https://o.stihi.ru (дата обращения: 23.05.2021).
[2] Стихи.ру [Электронный ресурс]. URL: https://stihi.ru (дата обращения: 23.05.2021).

2) создание фильтра, способного выбрать из огромного массива текстов новостные стихотворения;
3) оценка фильтра и описание отобранных стихотворений.

Такая работа представляет собой шаг к описанию поля современной наивной поэзии, представляющий интерес для теории литературы. Также созданный инструмент может быть интересен ученым, занимающимся социальными исследованиями. В дополнение к этому, полученная выборка стихотворений — это специфическая история 2013 года в стихах, написанная совместно сотнями авторов.

2. О коллекциях текстов

Для того, чтобы найти новостные стихотворения во всем массиве текстов, необходимо два корпуса. Первый из них — корпус новостей года. Он необходим для того, чтобы найти новостные события в корпусе стихотворений. Второй — корпус стихотворений.

Новостной корпус был составлен нами на основе двух сетевых изданий: Lenta.ru (Лента) и Vz.ru (Взгляд). Корпус новостей “Ленты” был взят у пользователя DmitryYutkin с сайта kaggle.com[3]. В датасете собраны новости из категории ленты “news” (большие
2 аналитические материалы не включены в датасет). Корпус новостей “Взгляда” собран в ходе работы над курсовой. В него попадали все новости с сайта, чьи id находились в промежутке между первой новостью 1 января выкачиваемого года и последней новостью 31 декабря выкачиваемого года[4]. Корпус был собран из нескольких источников для увеличения репрезентативности выборки новостей и уменьшение биаса в сторону одного из изданий. Архив стихотворений был предоставлен Б. В. Ореховым, коллекция скачана посредством краулинга.

Объем новостного корпуса за 2013 год составил 88557 записей (45260 “Взгляд”, 43297 “Лента”), объем стихотворного: 3 025 460. В среднем, в 2013 году на stihi.ru ежемесячно публиковалось около 252 342.5 записей (значение медианы), около 8466.75 записей в день (значение медианы). Новостные корпуса представляли собой tsv-файлы со всеми новостями за год, в архиве стихотворений каждое стихотворение хранилось отдельным txt-файлом в папке соответствующего месяца, что требовало дополнительной предобработки и конвертации архива в tsv-файлы для каждого месяца.

На stihi.ru существует рубрикация по темам (любовная лирика, гражданская, пейзажная, городская, религиозная, философская, мистика и эзотерика), но в рамках

[3] DmitryYutkin (2019). News dataset from Lenta.Ru [Электронный ресурс]. URL: https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (дата обращения: 23.05.2021).

[4] Например, первой новости 2013 года соответствует id 614494 (vz.ru/news/2013/1/2/614494.html), последней — 666700 (vz.ru/news/2013/12/31/666700.html).

курсовой мы ее не использовали по нескольким причинам. Первая из них: авторы вольны не выбирать никакую метку вовсе, поэтому поиск новостной лирики нужно производить на всем массиве стихотворений. К тому же, возможность рубрикации стихотворений появилась только в 2008 году, что делает ее использование невозможной на стихотворениях, опубликованных раньше.

3. Предотбор стихотворений из архива stihi.ru

Количество публикацией на сайте велико, но не все из них являются стихотворениями. Некоторые пользователи воспринимают stihi.ru как платформу для ведения блога, публикации новостей, прозаических текстов и даже чужих произведений и их переводов [Лейбов, Орехов (in print)]. Такие записи создают ощутимый шум, так как объем их текста превышает объемы стихотворения в несколько раз (в частности, если это публикация литературного произведения.

Помимо этого, встречаются записи, в которых со стихотворением соседствует прозаический отрывок — комментарий к стихотворению. В контексте новостных стихотворений, это может быть цитата из СМИ, которая проясняет новостное событие.

Любые нестихотворные строки являются шумом, поэтому перед нами встала задача фильтрации стихотворных строк. Для этого мы проверяли каждую строку на длину. Если в строке было больше 16 гласных русского языка, такая строка отбраковывалась. Граница в 16 гласных была выбрана исходя из традиций русского стихосложения. Размерами силлабо-тонического стиха являются хорей, ямб, дактиль, амфибрахий и анапест [Гаспаров 1993], а наивные поэты в большей мере стремятся создавать стихотворения, похожие на традиционные, а не экспериментировать с формами [Лейбов, Орехов (in print)].

Если искать в поэтическом подкорпусе Национального корпуса русского языка [Гришина и др. 2006] стихотворения, написанные двусложными размерами, в которых стоп больше восьми, то таких стихотворений будет 357 (меньшее количество стоп наблюдается в 73 696 стихотворениях). Для трехсложных размеров в корпусе находится 13 684 стихотворений, в которых в какой-то из строк стоп больше пяти (против 66 730 стихотворений, в которых стоп меньше). Таким образом, строки, в которых больше 16 гласных, являются стихотворными с меньшей вероятностью. Но даже если особенно длинная стихотворная строка будет отбракована, остальные все также войдут в выборку.

Также мы стремились отфильтровать строки, написанные не на русском языке, и отбраковывали строки с латиницей, в которых не было символов русского алфавита. Если же в строке была и латиница, и кириллица, то такая строка попадала в выборку. В

добавление к этому, мы убирали строки, в которых были специальные символы белорусского или украинского языка.

Также в исходном архиве были указаны ники пользователей, написавших стихотворения, и ссылки на рецензии на это стихотворение. Эти строки также были убраны.

После такой фильтрации мы считали количество получившихся строк. Если оно было не больше 100, запись попадала в выборку. Если строк было больше, это указывало на то, что запись является нестихотворным произведением (публикация дневниковой записи, списка участников конкурса и т.д.) или сборником стихотворений. В цифре 100 мы ориентируемся на [Лейбов, Орехов (in print)].

Таким образом отфильтровывалось 8-10 тысяч записей за месяц, отфильтрованный корпус stihi.ru составил 2 666 408 записей. Подчеркнем, что пускай речь идет о 8-10 тысячах записей в месяц (что при медиане в 252 тысяч может показаться незначительным), это были наиболее объемные произведения. К примеру, только за 8 января 2013 года на сайте была опубликована авторская повесть в трех частях, оцифрованная книга, сборник стихотворений, авторская книга в стиле фэнтези в стихах (общее количество строк в этих четырех файлах — 25 394). Отфильтрованные записи сохранялись в tsv-файл соответствующего месяца, время фильтрации занимало около 6 часов.

4. Лемматизация корпусов

После этого мы лемматизировали новостной и поэтический корпус при помощи консольного анализатора mystem [Segalovich 2013]. Мы сохраняли информацию о годе, месяце и дне публикации новостей и стихотворений, в корпусе ленты также была доступна информация о рубрике новости и ее тегах. Также мы сохраняли источник новости: новостное издание Лента.ру или Взгляд. Каждой записи в корпусе соответствовала одна строка.

5. Поиск терминов, указывающих на новостные сюжеты

Следующей задачей стал поиск новостных сюжетов, которые были бы представлены как в новостном корпусе, так и в поэтическом. Сюжеты могут быть как крупными (падение Челябинского метеорита), так и локальными (встреча Виталия Милонова со Стивеном Фраем). В рамках курсовой нас интересуют поиск всех новостных сюжетов, вне зависимости от величины темы, поэтому наш инструмент должен одинаково хорошо работать для сюжетов любого уровня. По этой причине мы не использовали тематическое моделирование: новостных сюжетов в год тысячи, и исключительно по выдаче модели определить сотни тем, не зная о них заранее, представляется трудновыполнимой задачей.

Исследование распределения частотностей токенов по дням также кажется несообразным масштабу задачи: униграмм в год десятки тысяч, биграмм — больше миллиона (не говоря о более длинных n-граммах). Некоторые токены встречаются в корпусе ежедневно (напр. “президент”), и из повышения их частотностей не следует появление какой-либо конкретной темы.

О появлении темы свидетельствуют токены, которые встречаются в корпусе не регулярно, а только в контексте своей темы. По этой причине мы создали словарь бинарных матриц длиной в год (365 для 2013-го года), в которых было бы указано, в какие дни униграмма или биграмма встречалась в корпусе, а в какие нет. В целях сокращения объема памяти и вычислений, токены, которые встречались в новостном корпусе только в одну из дат, не сохранялись в json-документ, в котором мы хранили матрицы вхождений. Также важно заметить, что из поэтического корпуса были обработаны только такие n-граммы, которые встречались в новостном корпусе. Это также сделано для сокращения объемов обрабатываемой информации, так как n-граммы, которых нет в новостном корпусе, не могут свидетельствовать о появлении какой-либо новостной темы. Уже на этом этапе объем обрабатываемой информации для 2013 года составил 7566 мегабайт.

Для поиска всплесков мы работали с этими json-документами, в которых хранились словари бинарных матриц. Униграммы и биграммы хранились раздельно, с разделением по корпусу-источнику.

Посмотрим на первые сто дней для токена “метеорит”:

0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1

Челябинский метеорит упал 15 февраля 2013 года. С 15 февраля “метеорит” упоминался в новостном корпусе каждый день на протяжении 14 дней. После этого встречаемость термина снизилась, но все равно оставалась выше, чем до момента его падения. 3 из 45 дней “до” инцидента (0.06) против 16 из 41 дней после инцидента (0.39).

Но стихотворения отличаются от новостей тем, что их пишут не для того, чтобы проинформировать о происходящем в мире, а для того, чтобы выразить собственную реакцию на эти события и поделиться личным опытом, который имеет ключевое значение для автора. По этой причине публикация стихотворений часто “задерживается”: одни авторы откликаются на новостную повестку моментально, но другим нужно время: как на создание произведения, так и на формирование собственной оценки происходящего.

По этой причине мы приняли решение разбить новостное пространство на интервалы в 10 дней (10 дней — ⅓ месяца) и считать скользящее среднее по трем периодам для суммы

дней, в которые упоминаются термин в течение периода. Мы считали скользящее среднее, учитывая два предыдущих периода, также с мыслью, что 3 периода составляют один месяц. Если значение для периода превышало скользящее среднее для предыдущего периода на 2 единицы (максимальное значение — 10, длина периода), то такое вхождение попадало в список подъемов частотностей. Разница в 2 единицы выбрана ввиду того, что она указывает на более значимое отличие, чем разница в 1 значение, и поэтому лучше отделяет новостные всплески от обычного распределения n-граммы в корпусе.

Если повышение встречаемости наблюдалось и в новостном, и в поэтическом корпусах, информация об этом записывалась в документы для всплесков для униграмм и биграмм. Если повышение наблюдалось только в новостном корпусе, такие n-граммы записывались в отдельные файлы. Если же n-грамма отсутствовала в поэтическом корпусе, то информация о всплеске также записывалась отдельно. В дальнейшем мы работали только с терминами, которые были и в новостном, и в поэтическом корпусах.

В документе были записаны списки n-грамм, дававших всплески, и периоды, в которые наблюдались эти всплески, каждая n-грамма с новой строки. Заметим, что мы хранили данные только о периоде, а не о датах, в которые встречалась n-грамма. При поиске все даты, которые было необходимо проверить, вычислялись через библиотеку datetime. Уже на этом этапе можно было проводить исследование стихотворений, написанных на какую-либо определенную тему: можно было считать документ с всплесками, записать данные в словарь и найти по термину соответствующие стихотворения.

N-грамма: вирус грипп Период: 2013-02-20

1.

ЛЮДИ В МАСКАХ

Гриппа карнавал в осенних красках –
в городе повсюду люди в масках.

Всё скрывают лица друг от друга,
а глаза у них полны испуга.

Не узнать: боится заразиться
или заразить других боится?

Точно, что больны без масок лица,
коли не боится заразиться.

Так испуган пассажир трамвая,
щипача в другом подозревая.

Вирус гриппа стал бациллой страха,

доводящим общество до краха.

Заменяют серые повязки
лицемерной вежливости маски.

03.11.09

Стихотворение присутствует в архиве, но в данный момент удалено с ресурса, поэтому ссылка не приведена.

2.

Вирус подкосил меня коварный,
вирус гриппа, подлостью опасный.
Пятый день борюсь я с ним отважно,
но исход борьбы пока не ясен.

И врача мне вызывали на дом,
к назначеньям приступил я сразу,
и режим постельный – всё, как надо,
но упорна мерзкая зараза.

Вирус гриппа действует паскудно:
размножается в моих же клетках…
Я держусь, держусь ещё покуда,
песенка моя ещё не спета.

И я всё-таки преодолею
вирус мерзопакостный, ничтожный.
Homo sapiens, краса Вселенной,
я сильнее должен быть – и точка.

Рамзан Назиров[5]

Эпидемия вируса гриппа — локальный сюжет. Ежегодные эпидемии вирусов зачастую не выделяют в главные события года. Так, в материале “РИА Новости” “Главные события 2013 года. Выбор РИА Новости”[6] свиной грипп упоминается один раз в контексте отставки Геннадия Онищенко от должности руководителя Роспотребнадзора. Тем не менее, фильтр ловит и такой сюжет.

Как указывают [Лейбов, Орехов (in print)], значение имеет не только дата написания стихотворения, но и дата публикации. Это подтверждает дата написания первого стихотворения: 03.11.09. В 2009 году бушевал свиной грипп. Стихотворение было написано в 2009, но опубликовано во время эпидемии 2013 года, так как оно было актуальным.

Но наиболее интересной задачей был автоматический поиск новостных тем.

[5] Вирус гриппа (Рамзан Назиров) URL: https://stihi.ru/2013/02/25/3803

[6] Главные события 2013 года. Выбор РИА Новости - РИА Новости, 01.03.2020 [Электронный ресурс]. URL: https://ria.ru/20131217/984604250.html (дата обращения: 23.05.2021).

6. Автоматическое выделение новостных тем

Для того, чтобы автоматически выделить новостные темы, которые соответствуют каждому периоду, мы инвертировали датасет с n-граммами. Если раньше каждой n-грамме соответствовали периоды, в которые она повышалась ее встречаемость, то теперь каждому периоду соответствовали все n-граммы, которые давали всплеск.

До:

диссертация 2013-01-31, 2013-02-20, 2013-05-21, 2013-09-08, 2013-11-27

После:

2013-03-02 0 0, 00 мск, 08 30, 09 00, 1 март, 1 марта, 1 создавать, 1 ук, 10 март, 10 марта, 11 класс, 11 март, 12 март, 13 00, 13 7, 13 апрель, 13 март, 14 апрель, 14 март, 15 май, 16 30, 16 март, 164, 17 апрель, 18 30, 1969, 1969 год, 1971, 1971 год, 1973, 1973 год, 1982, 2 го, 2 март, 20 00, 20 2, 20 март, 2001 год, 21 апрель, 21 март, 22 40, 22 год, 22 ноябрь, 23 00, 239, 25 март, 26 4, 27 февраль, 28 май, 28 февраль, 290, 291, 3 8, 3 март, 30 4, 30 тысяча, 373, 39 год, 4 10, 4 6, 4 март, 4 марта, 420, 5 12, 5 8, 5 август, 5 апрель, 5 март, 502, 6 12, 6 март, 6 марта, 65 год, 68, 7 12, 7 март, 7 марта, 7 январь, 750, 8 март, 8 марта, 9 апрель, 9 март, generation, home, man, org, top, vii, авс, австрия, актер режиссер, актер сыграть, актерский, алания, александр...

Как видно из этого списка, в него попадают 1) числа, даты; 2) англоязычные слова и части слов, которые малоинформативные (home, man, org); 3) частотные имена (Александр), которые заведомо встречаются в больших контекстах, чем n-граммы, соответствующие новостному сюжету.

По этой причине мы приняли решение создать фильтр для n-грамм, которые попадут в список периодов. Мы отсеивали слова, которые встречаются в большом количестве контекстов: числительные, служебные слова, частотные наречия, глаголы, существительные с широкой сферой употребления (“гораздо”, “мочь”, “работа” и так далее). Также мы отсеивали часто встречающиеся имена[7], униграммы на латинице, и униграммы, чья длина была меньше трех символов. Ссылка на полный список стоп-листа в приложении.

После этого был получен более короткий и информативный список:

2013-03-02 авс, австрия, актер режиссер, актер сыграть, актерский, алания, александрович, алименты, альпинист, алюминиевый, американо, анжи, аномалия, антарктида, антарктический...

[7] Русские имена от А до Я [Электронный ресурс]. URL: https://тайна-вашего-имени.рф/russkie-imena.html (дата обращения: 23.05.2021).

После получения таких списков мы итерировались по лемматизированному поэтическому корпусу и искали стихотворения, которые относились к одному из периодов и в которых встречались термины, давшие всплески за этот период. Но текстов, в которых хотя бы один раз встречается одна из n-грамм (в списке всем периодам соответствует 48 828 n-грамм), очень много. По этой причине мы выбирали те стихотворения, в которых встречалось хотя бы три токена из списка всплесков за этот период. Биграммы не проверялись.

Также мы добавили фильтр для записей, освещающих конкурсы и сборники на сайте. В таких записях встречается много фамилий, часть которых совпадает с фамилиями из новостного списка, и дает шум. Фильтр выглядел так: если в отобранном стихотворении находилось хотя бы три термина из “черного” списка, такое стихотворение не передавалось на следующий этап и записывалось в специальный файл для отбракованных стихотворений. Это было сделано для того, чтобы контролировать работу фильтра и проверять, не попадают ли в него верные стихотворения.

Черный список:

здравствовать, произведение, ведущий, конкурс, балл, стихобалл, взнос, приз, б, пзс, зрительский, симпатия, литературно-художественный, антология, прочтение, заявка, внимание, рецензия

При помощи этого списка была отбраковано 34 записи. Все их них были записями, связанными с проведением конкурсов или публикацией антологий.

Пример отбракованного текста. В третьей колонке — найденные слова из черного списка, в четвертой — токены из списка всплесков, которые встретились в записи. В тексте они выделены подчеркиванием, пустые строки убраны в целях экономии места.

2013 01 24 ['балл', 'прочтение', 'внимание'] ['соловьев', 'обосновывать', 'объективный', 'химический']

Дорогие авторы!
С вами Эмили Шантэ
*
Объявляем голосование по конкурсу "ИДЁТ ВОЛШЕБНИЦА-ЗИМА"
ВНИМАНИЕ!!!
Судейство.
==========
У ВАС ЕСТЬ 3 БОНУСНЫХ БАЛЛА.
ВЫ МОЖЕТЕ ИХ
1) ОТДАТЬ ОДНОМУ СТИХОТВОРЕНИЮ (3)
ПРИМЕР ШОРТ-ЛИСТА
.....
1,3, 4(3), 5, 6, 7 .....

2)РАСПРЕДЕЛИТЬ МЕЖДУ ДВУМЯ (2) И (1)
ПРИМЕР ШОРТ-ЛИСТА

.....
1,3, 4(2), 5, 6, 7(1) .....
3) ИЛИ РАСПРЕДЕЛИТЬ МЕЖДУ ТРЕМЯ (1),(1), (1)
ПРИМЕР ШОРТ-ЛИСТА
.....
1(1),3, 4(1), 5, 6, 7(1) .....
ЖЕЛАТЕЛЬНО ОБОСНОВАТЬ СВОЙ ВЫБОР)
ПРОШУ БЫТЬ ОБЪЕКТИВНЫМИ! ИТАК,ГОЛОСУЕМ!
1.Валентина Калёва
"Гайдай - навсегда!"
2.Николай Ватулин
«Голая правда»
3.Вера Да Юра
Эхо семидесятых
4.Соловушка 2
Мои стихи
5.Виктор Колесников 6
За любимую женщину выпить
6.Татьяна Василевская2
Владимиру Высоцкому
7.Виктор Граков
Помня Высоцкого
8.Юрий Кожанов
"Володю вспоминая"
9.Александр Копп
"Коньяк теперь мне только снится"
10.Геннадий Смирнов2
"Вулкан и Река"

11.Чернухо Игорь.
В.С. Высоцкому.
12.Валерий Таиров
Владимиру Высоцкому.Бег иноходца"
13.Евгения Чмут.
"Ушёл поэт"
14.Елена Соловьева 3.
"Охота на волка".
15.Фёдор Резник
"В.С. Высоцкий"
16.Павел Конча
ОБЫКНОВЕННЫЙ, ПРОСТОЙ, НЕ ХИМИЧЕСКИЙ...
17.Лееле Потапова
18.Галина Панюшкина
В.Высоцкому
19.Михаил Стихоплётов
Вспоминается мне...Владимиру Высоцкому
20.Мария Лещинска
Спой мне, Гитара, песню…
21.Самчук Александр
"Задыхаюсь я здесь... Без свободы"
22.Лев Неф
SOS. И будут с вами случаться чудеса.
23.Анатолий Постников
ЛЕГЕНДА
24.Юрий Семецкий Меня сегодня дама посетила
25.Ксения Виниченко
"Сказка про Музу и школьные годы"
26.Пётр Кантарович.

К Владимиру Высоцкому обращаюсь…
27.Желамский.
Возвращая Высоцкого.
28.Андрей Кошмаров.
"Зарисовка".
29.Василий Толстоус
Высоцкий
30.Эдмар
На 75-летие Высоцкого
ЖЕЛАЮ ПРИЯТНОГО ПРОЧТЕНИЯ![8]

По списку терминов можно было бы предположить, что этот текст является новостным (соловьев, обосновывать, объективный, химический), но токены из черного списка указывают на обратное.

В результате такой фильтрации мы получили 2204 стихотворений. Среди них большинство стихотворений были “гражданскими”, но встречались и не привязанные к новостному пространству стихотворения.

poem#376-1871 2013 10 30 ['втрое', 'обстрел', 'сектор', 'советник']

-1-
Сделай сыну деревянный меч,
С ним сойдись в потешном поединке.
Подыграй, и чтоб не юзнул с плеч
Придержи тихонько за ботинки.
<...>
-5-
Врали, врали весь двадцатый век,
Врали так, что верили и сами,
В то, что рай построит человек
Сам. Среди земли. Под небесами.

В то, что обретя на все ответ,
Станут люди вдруг счастливей втрое.
Что на Марсе яблоневый цвет
Цвет бинта запекшегося скроет.
<...>
-9-
И поляжет оголтелых цепь,
Запертая в секторе обстрела.
Вдалеке уснет живая степь,
Окна пригашая неумело.

Если же обступят в дымный час -
Встанете к спине спиною с другом.

11

[8] Голосование Идёт Волшебница-Зима (Конкурс Самовыражение 2) URL: https://stihi.ru/2013/01/24/9683

Правда ополчится посредь вас.
Зов орлиный грянет по округам.
-10-
Сделай сыну деревянный меч.
Не спеши купить ему планшетник.
В жизни можно чем-то пренебречь -
Здесь уж разум - всякому советник,

Только знай, что время сочтено:
Сколько дней пробуксовать осталось?
Все оно для дела придано -
Духом возмужать - совсем не малость.
28-30.10.2013

Иванов Антон Геннадьевич[9]

Для того, чтобы в выборку попадали новостные стихотворения, нужно, чтобы термины были связаны между собой.

7. Цепочки терминов

1. Чтобы проверить, насколько стихотворение соответствует какому-либо новостному сюжету, мы взяли выдачу прошлого этапа (дата, термины, текст стихотворения) и через новостной корпус получили для каждого стихотворения все новости, в которых упоминался хотя бы один из терминов, встретившихся в стихотворении.

Примеры новостей, соответствующих одному из найденных стихотворений. Приводится первое предложение новости и по меньшей мере одной предложение, в котором встретились термины.

2013 02 06 ['липовый', 'диссертация'] Дмитрий Медведев заявил, что количество «липовых» кандидатов и докторов наук в России «зашкаливает за все возможные пределы». Таким образом премьер отреагировал на скандал с плагиатом в диссертациях. <...>

2013 02 07 ['докторский', 'диссертация'] Газета «Комсомольская правда» связала имя задержанного главы Высшей аттестационной комиссии (ВАК) Минобрнауки Феликса Шамхалова со строительством корта, конюшни и вилл под Сколково. <...> В частности, изданию удалось выяснить, что Шамхалов защитил докторскую диссертацию в 30 лет, а уже через год стал заслуженным деятелем науки РФ, что является рекордом среди живущих ученых-экономистов. <...> Задержание Шамхалова по делу о мошенничестве было произведено на фоне скандала, связанного с плагиатом в диссертациях и случаями необоснованного присвоения ученых степеней. <...>

[9] Сделай сыну деревянный меч (Иванов Антон Геннадьевич) URL: https://stihi.ru/2013/10/30/7771

2. После этого мы считали все встреченные в новостях термины за каждый период отдельно. Если в одной из новостей упоминалось несколько терминов ('липовый', 'диссертация'), и какой-то из терминов встречался в другой новости ('диссертация', ‘докторский’), то термины обеих новостей объединялись в термины одной новостной темы ('липовый', 'диссертация', ‘докторский’). Так был сформирован список тем и соответствующим им терминов за каждый период.

После проверки новостей из всех стихотворений за период 2013-01-31 был получен такой список терминов для темы: вако, диссертация, докторский, липовый, плагиат, химик.

3. Получив такие списки, мы вновь перепроверили полученные стихотворения. В финальную выборку попадали стихотворения, в которых хотя бы три термина принадлежали одной из выделенных на предыдущем этапе теме.

Посмотрим на одно из стихотворений. Терминов много, но к какой-либо из тем относятся только три, и эта тема — “скандал с липовыми диссертациями”.

poem#69-470 2013 02 08 ['менеджер', 'липовый', 'докторский', 'отблагодарить', 'диссертация', 'социалистический', 'чехов']

Глядишь на Россию и диву даёшься!
Иной раз плачешь и тут же смеёшься...
Только недавно гордились страной!
Самой образованной в мире становится строй.
Тракториста, токаря в ней не видать!
Ушли в «отсталую» социалистическую даль...
Сейчас в демократической нашей стране!
Образование на «высшем» уровне, как и в Кремле.
Юристы Экономисты Менеджеры Артисты!
Все диссертации в стране защитили...
О чем Медведеву на совещании заявили!
Количество «липовых» кандидатов и докторов...
Зашкаливает за пределы, будь здоров.
Кругом одни,- профессора наук!
Расширился научный Российский круг...
Даже наш премьер здесь удивился!
Своим профессионализмом всегда гордился.
А тут, почти докторские все защитили!
Хотя нигде и ничего здесь не учили...
А дипломы с отличием все получили!
Преподавателей своих отблагодарили.

Теперь стремятся пост большой занять!
Чтобы хорошие зарплаты получать...
Обидно премьеру за страну стало!
В детстве играл с друзьями мало.
Науку грыз зубами, сутками изучал
И очень часто он по ночам рыдал!
Его не пускали во двор играть...
Заставляли Чехова много читать.
Сейчас немного при платился...
Диплом в карман тебе и свалился!
Ты в верхах с дипломом сидишь...
На место премьера уж глядишь.
Как его подвинуть быстренько там
И приступить к большим трудам!
Народ русский в лапти обувать...
Страну родную быстрее угроблять.

Александр Котельников[10]

Стихотворение действительно новостное, и можно однозначно определить тему, по которой оно написано.

8. Оценка полученных результатов

В финальную выборку попало 492 стихотворения из полученных на предыдущем этапе 2204 (из 2 666 408). Нашей целью было получение наиболее соответствующих новостной повестке стихотворений. Многие новостные стихотворения не проходили последующие уровни фильтрации[11], поэтому вопрос о том, как оценить recall, остается открытым. Оценка количества ненайденных стихотворений представляется сложной задачей, потому что, чтобы сказать однозначно, является ли стихотворение действительно новостным, его нужно прочитать и установить соответствующие ему новости.

Для того, чтобы оценить работу фильтра, нами были прочитаны все отобранные стихотворения. В 74.4% случаев выбранные стихотворения действительно были новостными (precision). Если они не были новостными, то они принадлежали к одной из категорий:

[10] Скандал вокруг липовых диссертаций (Александр Котельников) URL: https://stihi.ru/2013/02/08/6116

[11] Например, еще на стадии отсеивания стихотворений, в которых было меньше трех терминов из списка всплесков, были потеряны два из двух стихотворения о наводнении в Зее и одно из двух стихотворений об одном из эфиров Ивана Урганта программы «Смак» . Также не была выделена тема с днем рождения В. С. Высоцкого, на которую было написано множество стихотворений. Это связано с тем, что стихотворения о нем публикуют регулярно, и поэтому дата его дня рождения не попала в список терминов, так как тема не являлась редкой для stihi.ru.

Тема	Кол-во	Процент
Новости, записанные прозой / другая информация в прозе	18	3.66%
Неотфильтровавшиеся записи с информацией о конкурсах	4	0.809%
Неновостные стихотворения	68	13.82%
Гражданские стихотворения про Россию	29	5.89%
Гражданские стихотворения не про Россию	3	0.61%
История в стихах на гражданскую тематику (описание события, которого никогда не было, или фантазия на тему реального события)	3	0.61%

Заметим, что даже если в стихотворении упоминались общие с новостной темой действующие лица или места действия, если в нем не упоминалось событие новости, то такое стихотворение относилось к категории “неновостные стихотворения”.

poem#310-1515 2013 08 13 ['вагончик', 'ай-петри', 'мисхор']

Стальные нити уносили мысли ввысь,
к творению природы, скал немых синьор.
Скрипит вагончик подо мной, антрепренёр.
мазком невидимым к Ай-Петри прикоснись…
- тяни меня к пленеру, вверх фуникулёр!

Обворожительный здесь вид, с высот обзор,
стволы деревьев искорёжил изувер.
Полёт фантазий в слой озона стратосфер,
нас провожает к высоте теплом Мисхор,
в тенистых пальмах спит загадочно Дюльбер.*

Чарует, манит красота с подножья гор,
- мы на галёрке у театра парадиз.*
«Большая Ялта» исполняет свой каприз,
рукой, касаясь облаков, птиц разговор,
- я слышу сосен шум, надменный Кореиз.
13 августа 2013г.

Юрий Слащев[12]

[12] Большая Ялта, неизвестная глава (Юрий Слащев) URL: https://stihi.ru/2013/08/13/2994

В августе 2013 года произошла авария на канатной дороге “Мисхор — Ай-Петри”. Но в стихотворении об этом не упоминается, поэтому оно было отнесено в категорию “неновостные стихотворения”.

Посмотрим на распределение новостных стихотворений по темам. В скобках дается абсолютное количество стихотворений на тему и их доля в корпусе.

	Тема	Кол-во	%
1	Майдан	87	17.68%
2	Смена Папы в Ватикане (отречение-избрание-папство)	32	6.5%
3	Челябинский метеорит	28	5.689%
4	Смерть Уго Чавеса	26	5.28%
5	Сноуден	21	4.27%
6	Убийство Егора Щербакова на овощебазе в Бирюлево-бойня-мигранты	14	2.85%
7	Война на Ближнем Востоке	14	2.85%
8	Смерть Калашникова	10	2.03%
9	Спорт - хоккей	10	2.03%
10	Смерть (болезнь) Нельсона Манделы	9	1.83%
11	Наводнение на Дальнем Востоке	8	1.63%
12	Взрывы на Бостонском марафоне	8	1.63%
13	Политика государства, депутаты, законы	7	1.42%
14	Офшоры Кипр	7	1.42%
15	Закон Магнитского - дело Димы Яковлева	5	1.02%
16	Опальные политики	5	1.02%
17	Катастрофа боинга в Казани	5	1.02%
18	Финансы	4	0.809%
19	Дни памяти - Блокада	4	0.809%
20	Протон не взлетел	4	0.809%
21	Дело Оборонсервиса-Сердюков	4	0.809%
22	Освобождение Ходорковского, Лебедева	4	0.809%
23	Бойня в Белграде	4	0.809%
24	Дни памяти - Холокост	3	0.61%
25	Смерть Березовского	3	0.61%
26	Спорт - атлетика (Исинбаева)	3	0.61%
27	Спорт - бокс (Поветкин-Кличко)	3	0.61%
28	Липовые кандидаты наук - критика системы образования	3	0.61%

29	Казнь Чан Сотхэка (дядя Ким Ченына)	3	0.61%
30	Думские фракции	2	0.41%
31	Скандалы с Роснано (Чубайс) - Росатом	2	0.41%
32	Смерть Тэтчер	2	0.41%
33	Спорт - биатлон (Домрачева)	2	0.41%
34	Спорт - футбол (Халк)	2	0.41%
35	Гражданская лирика о роли поэзии (с новостями)	2	0.41%
36	Самоубийство Долматова	2	0.41%
37	Саммит в Давосе	2	0.41%
38	Скандал с кониной в полуфабрикатах	2	0.41%
39	Бозон Хиггса (открытие и награждение нобелевской премией)	2	0.41%
40	Евровидение	2	0.41%
41	Универсиада Казань	2	0.41%
42	Прослушка телефона Меркель	2	0.41%
43	Теракт в Волгограде	2	0.41%
44	Статья Мос. Комсомольца о "политической проституции"	1	0.2%
45	Дни памяти - Сталинградская битва	1	0.2%
46	Дни памяти - день рус. водки (диссертация Менделеева)	1	0.2%
47	Смерть деда Хасана	1	0.2%
48	Смерть Алексея Германа-старшего	1	0.2%
49	Смерть Золотухина	1	0.2%
50	Смерть Пола Уокера	1	0.2%
51	Спорт - теннис (Шарапова)	1	0.2%
52	Мятеж Квачкова	1	0.2%
53	Взрыв на шахте Воркутинская	1	0.2%
54	15-ый съезд компартии	1	0.2%
55	Абрамович ремонтирует замок в Лондоне	1	0.2%
56	Поднятие тарифов ЖКХ	1	0.2%
57	Депардье подарили квартиру в Саранске	1	0.2%

	58	Встреча Милонова со Стивеном Фраем	1	0.2%
	59	Задержание Новгородской ОПГ	1	0.2%
	60	Губернатор Ульяновск. обл. заменил текст Дины Рубиной на Тотальном диктанте	1	0.2%
	61	Неудачная шутка Урганта про Украину	1	0.2%
	62	Витас сбил велосипедистку	1	0.2%
	63	Неудачная вербовка ЦРУ	1	0.2%
	64	Легализация однополых браков	1	0.2%
	65	Французский писатель застрелился против однополых браков	1	0.2%
	66	Торнадо в Оклахоме	1	0.2%
	67	Съезд Общероссийского Народного Фронта	1	0.2%
	68	Сочи-2014	1	0.2%
	69	Пиджак президента Польши испачкали яйцов во время визита на Украину	1	0.2%
	70	ДТП под Подольском	1	0.2%
19	71	Убийство пастыря Павла Адельгейма	1	0.2%
	72	Годовщина войны-2008 в Грузии-Саакашвили	1	0.2%
	73	Смена пола Брэдли Мэннинга	1	0.2%
	74	Саммит Двадцатки в Стрельне	1	0.2%
	75	Ройзман стал мэром Екатеринбурга	1	0.2%
	76	Пожар в Новгород. обл. психоневр. интернате	1	0.2%
	77	Задержание экологов гринпис на территории России	1	0.2%
	78	Курбан-Байрам	1	0.2%
	79	Поимка Полонского в Камбодже	1	0.2%
	80	Гололедица	1	0.2%
	81	Обрушение крыши над супермаркетом в Риге	1	0.2%
	82	Митинги против Эрдогана в Турции	1	0.2%
	83	Митинги против Мурси в Египте	1	0.2%

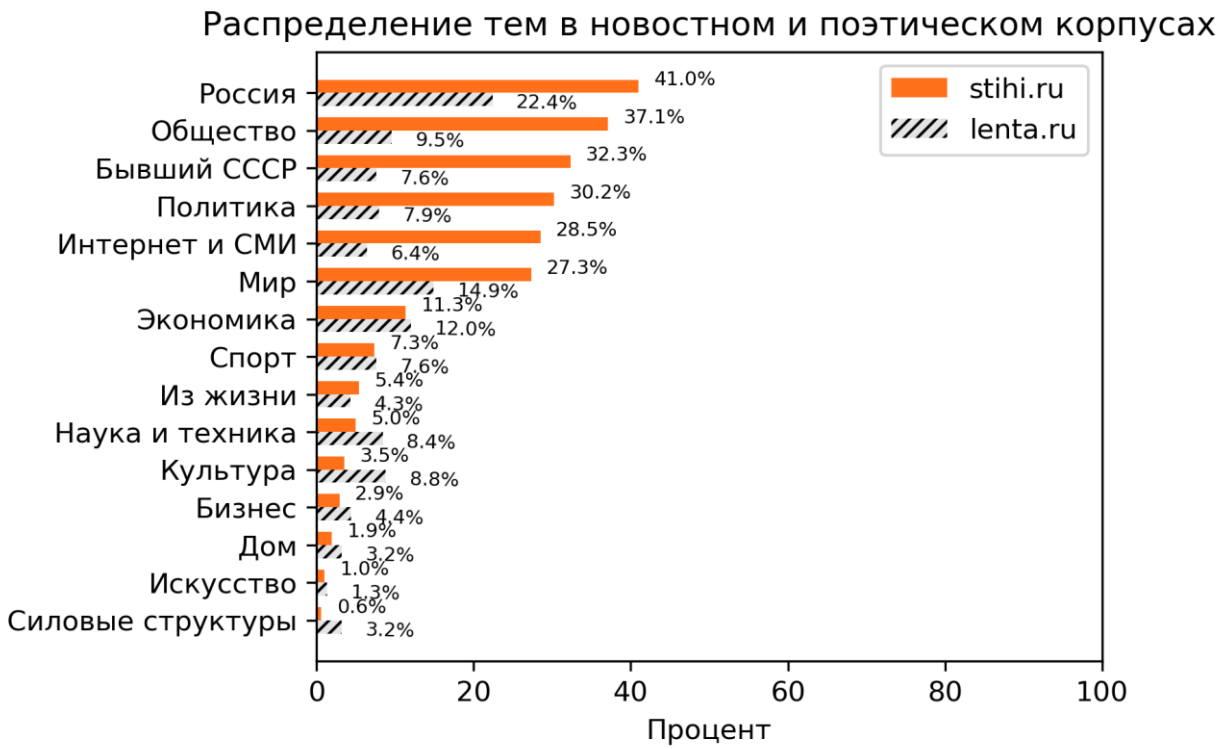
Файлы со стихотворениями (chain_filter2013-3.tsv) и их номерами по темам (topics2013.tsv) доступны в репозитории.

9. Распределение новостных рубрик

Сравним распределение новостных рубрик в найденных новостных стихотворениях и в новостном корпусе в целом для того, чтобы понять, какая тематика преобладает в новостной поэзии наивных авторов.

В датасете с lenta.ru были указаны рубрики, к которым относятся новости. Мы приписали каждому найденному стихотворению все рубрики из новостей, которые ему соответствовали, каждая рубрика по одному разу. В новостях к 12 стихотворениям из 492 не упоминалась ни одна рубрика (вероятно, что этим стихотворениям соответствовали новости только из газеты “Взгляд”). После этого мы посчитали долю представленности новостной рубрики во всех стихотворениях выборки.

Также мы подсчитали распределение новостных рубрик во всем датасете lenta.ru. Мы учитывали основные рубрики “Ленты”, к которому добавили четыре подрубрики: “Политика”, “Общество”, “Бизнес”, “Искусство”. В новостях 2013-го года не представлены существующие сейчас рубрики “Путешествия” и “Ценности”. У 13 новостей из 43297 не были указаны рубрики.



Заметим, что стихотворениям в выборке соответствовало несколько новостей с “Ленты” (медиана — 10), и некоторые из них не относились к новостному сюжету, а описывали похожую тему. Например, во время скандала с “липовыми” диссертациями в России, похожий скандал был в Германии, что в совокупности дало рубрики “Мир”, “Общество” и “Россия”, “Общество”, после объединения были получены теги “Россия”, “Общество”, “Мир”. По этой причине доля некоторых рубрик увеличилась. Тем не менее, по графику видно, что наивная поэзия с сайта stihi.ru пишется преимущественно на общественно-политические темы, а культура интересует авторов в меньшей мере. В центре внимания обоих корпусов находится Россия.

10. Особенности метода и их влияние на результат

1. Подсчет скользящего среднего

Так как в исследовании использовались данные только 2013 года, первый период, для которого скользящее среднее с контекстом 3 могло указать на повышение встречаемости какого-либо термина — 2013-01-21. За этот период в корпусе найдено 66 стихотворений, при медиане в 11. 43 из 66 стихотворений не являются новостными. Для того, чтобы избежать этой проблемы, стоит добавить в корпус последний месяц 2012 года.

2. Региональность (история про “Трактор”)

Во время разметки встретилось любопытное стихотворение.

poem#369-1803 2013 10 17 ['хоккеист', 'болельщик', 'шайба']

„Письмо белого МЕДВЕДЯ- ХК „Трактор„

Дорогие хоккеисты!!!!!!!!!

Пишет белый ВАМ, медведь!!!!!!!

Я болельщикам Урала,

Не могу в глаза смотреть.

Разогнать пора наверно

Ваш, мышиный серый трест,

У меня от шайб влетевших

Нет живых на теле мест.

Предлагаю в виде бренда

Вместо мишки, КЕНГУРУ!!!!!!!!

Будет ваши РЕЗУЛЬТАТЫ

Прятать в сумку, как в дыру.

Я прошу ОСВОБОДИТЕ!!!!!!!!!

Не губите РЕЙТИНГ мой.

Я сегодня с ВАШИХ маек

В лес уйду к себе ДОМОЙ!!!!!!!!

Олег Бояркин[13]

Это стихотворение напоминало новостные: это реакция на какое-то, очевидно, недавнее событие. Но в новостном корпусе не было никаких упоминаний команды “Трактор”.

В действительности, у этого стихотворения был свой инфоповод: 16 октября 2013 года челябинский “Трактор” проиграл петербургскому “СКА” со счетом 0:7[14]. Но ни в “Ленте”, ни во “Взгляде” об этом не написали. Для того, чтобы определять региональные новости (“Я болельщикам Урала / не могу смотреть в глаза” — автор через лирического героя выражает свою причастность к Уральскому региону), необходимы региональные корпуса.

11. Саморефлексия авторов

Некоторые авторы пишут стихотворения и о том, что они думают о новостной повестке и написании стихотворений.

poem#47-303 2013 01 28 ['багдад', 'кувейт', 'хусейн', 'ирак', 'аравия', 'наземный']

Дома я часто телевизор смотрю,

Лихо по трём программам ручкой кручу.

И очень полезно смотреть телеролики эти:

Вот жаркие споры шли в Моссовете,

А Станкевич Поповым оставлен в ответе.

Сергей с депутатами лихо сражался,

И до хороших решений всё же добрался.

<...>

Маликов Алик Емельянович[15]

---

[13] Письмо белого медведя- хк, ,трактор, , (Олег Бояркин) URL: https://stihi.ru/2013/10/17/3214

[14] Хоккей. Трактор 0:7 СКА - Онлайн трансляция матча - 16 октября 2013 [Электронный ресурс]. URL: https://www.liveresult.ru/hockey/matches/match51453_Traktor-SKA_St_Petersburg-online

[15] *** (Маликов Алик Емельянович) URL: https://stihi.ru/2013/01/28/4854

Писать о том, что происходит сейчас,

Ловить каждый жест проходящих людей.

Писать о себе, о них, о нас,

Как можно больше и как можно быстрей.

Писать о том месте, где родился,

Или о месте, где сейчас живешь,

<...>

Писать о том, как живет страна,

Или писать о том, что такое мир,

Писать, как живет он или живет она,

Как простой человек живет или кумир.

<...>

Писать о дотациях или реальные заводы,

Писать про армию или тех, кто откосил,

Про шаманов писать или прогноз погоды,

О тех кто промазал или кто гол забил.

Писать пока есть силы, или диктовать,

На бумаге ручкой или мелом на асфальте.

Писать пока ты живой, даже когда время выбирать,

Писать пока сам себе не скажешь, хватит.

Сергей Клименко Из Крыма[16]

Заметим также, что новостные авторы зачастую подписываются именем и фамилией (иногда еще отчеством и городом), а не никнеймом.

[16] Писать о том, что происходит вокруг... (Сергей Клименко Из Крыма) URL: https://stihi.ru/2013/02/05/9255

12. Заключение

В результате работы был создан алгоритм поиска новостных стихотворений. Алгоритм требует улучшения, но при его помощи удалось выделить некоторое число новостных наивных стихотворений и в первом приближении описать их специфику.

В будущем необходимо улучшить алгоритм с учетом полученных результатов: понять, какие параметры фильтра оптимальны, как влияет добавление биграмм в фильтр, какие методы машинного обучения и языковые модели могут помочь в автоматической классификации стихотворений.

Также для обобщения выводов о поле новостной наивной поэзии необходимо провести исследование на материале нескольких лет.

Приложение

Ссылка на репозиторий

https://github.com/mjolnika/naive-poetry-news-agenda-detection

Список литературы

1. Бонч-Осмоловская А. А., Орехов Б. В. Корпусно-статистические подходы к наивной поэзии // Корпусный анализ русского стиха: Сборник научных статей. Вып. 2 / Отв ред. В. А. Плунгян, Л. Л. Шестакова. — М: Издательский центр «Азбуковник», 2014. — С. 20—36.
2. Гаспаров М. Л. Русские стихи 1890-х — 1925-го годов в комментариях. — М.: Высшая школа, 1993.
3. Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования //Национальный корпус русского языка. – 2006. – Т. 2008. – С. 71-113.
4. Князева Е. А. Биографический дискурс в" наивной" поэзии Екатерины Зверевой //Филология в XXI веке: методы, проблемы, идеи. – 2015. – С. 282.
5. Лейбов Р. Г., Орехов Б. В. Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике. (in print)
6. Лурье М. Л. О феномене «наивного» сочинительства //Наивная литература»: исследования и тексты. М.: Московский общественный научный фонд. – 2001. – С. 15-28.
7. Неклюдов С. Ю. «Наивная литература»: исследования и тексты //М.: Московский общественный научный фонд. – 2001. – Т. 246.
8. Орехов Б. В. К количественному исследованию наивной поэзии: строки-шаблоны //Историческая поэтика жанра. – 2013. – №. 5. – С. 80-84.
9. Орехов Б. В. и др. Цитирование текстов ВС Высоцкого в текстах современных наивных поэтов //Филологические науки. Вопросы теории и практики. – 2013. – №. 11-2. – С. 141-143.
10. Петров А. А. " Наивная литература" тверского региона: история и перспективы изучения //Вестник Тверского государственного университета. Серия: Филология. – 2015. – №. 3. – С. 267-271.

11. Югай Е. Помянуть стихами: коммеморативная наивная поэзия //Археология русской смерти. – 2016. – №. 3.
12. DmitryYutkin (2019). News dataset from Lenta.Ru [Электронный ресурс]. URL: https://kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta (дата обращения: 23.05.2021).
13. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine //MLMTA. – 2003. – С. 273-280.