«Ищите нас через Яндекс»: методики и проблемы сбора сетевого фольклора

ДАРЬЯ РАДЧЕНКО

АННОТАЦИЯ. Интернет-фольклор представляет собой уникальный материал, открывающий возможности для изучения целого ряда ключевых проблем фольклористики: авторства текстов, истории их распространения, динамики варьирования. Его основной корпус составляют письменные и визуальные тексты, которые относительно легко поддаются поиску с помощью общедоступных технических средств Интернета, а объем получаемого таким образом материала позволяет не только описывать эти тексты на качественном уровне, но и проводить статистически достоверные обобщения. Однако поиск текстов в Сети связан с целым комплексом ограничений: полнотой и репрезентативностью данных, получаемых через поисковые машины, согласованностью и достоверностью автоматизированной статистики, спецификой формулирования поисковых запросов, достоверностью информации о носителях фольклора и т. д. В статье намечаются некоторые проблемы и предлагаются методики собирания интернет-фольклора.

КЛЮЧЕВЫЕ СЛОВА: Интернет, фольклор, полевое исследование, поисковая машина, статистика, методы исследования.

Переход к концепции WEB 2.0 в начале 2000-х годов и формирование рассчитанных на «пользовательский контент» ресурсов (блогов, социальных сетей и т. д.) выдвинули на первый план ценность творческой деятельности членов сетевого сообщества. Культура любительского производства смыслов, подкрепленная технологическими возможностями Интернет-платформ, способствовала размыванию различий между автором и аудиторией (Bell 2001: 165): каждый участник сетевой коммуникации не только потребляет тексты, но и имеет возможность сам создавать и распространять их. В результате Интернет изобилует самыми разнообразными текстами, в том числе отчетливо фольклорного характера.

Сетевой фольклор предоставляет исследователю богатые возможности для изучения процессов, которые не всегда можно проследить на материале классического фольклора. В отличие от устного фольклора, сетевой *a priori* является зафиксированным; более того, такого рода фольклорные тексты

могут быть тематически сгруппированными[1]. Кроме того, зачастую они являются предметом рефлексии самих его носителей, пусть даже на любительском уровне. Собрания таких текстов, нередко весьма обширные, снабженные подробными комментариями и датировками, можно обнаружить как на сайтах, посвященных осмыслению сетевого фольклора[2], так и на тематических площадках (в блогах, на форумах и т. д.). Текст, попадающий в Сеть, теоретически может храниться в ней в течение неограниченного времени. Даже после удаления того или иного сообщения в течение определенного времени его относительно легко восстановить благодаря автоматическому кэшированию страниц и сетевым репозиториям (Berry 2004: 5). При помощи несложных поисковых методов возможно в короткие сроки собрать тысячи релевантных текстов и изображений.

Тем не менее эти доступность, устойчивость и «континуальность» сетевого фольклора во многом иллюзорны. Как сам текст, так и весь сайт, на котором он находится, может в любой момент стать недоступным для постороннего читателя или даже быть уничтожен. Благодаря размерам сетевого сообщества и высокой скорости передачи информации сетевой фольклор быстро становится широко известным, а затем теряет актуальность. «Лентовидная» форма визуальной организации основных коммуникативных ресурсов (чатов, форумов, блогов и пр.) также провоцирует ситуацию, когда сообщение быстро выпадает из поля зрения пользователей и становится неактуальным.

Соответственно, весь комплекс интернет-фольклора оказывается текучим, изменчивым как с точки зрения набора текстов, так и с точки зрения процессов, происходящих с ними. Эту проблему усугубляют популярность и фольклорная продуктивность форумов, чатов и имиджбордов с отсутствием длительного кэширования, на которых тексты автоматически уничтожаются спустя некоторое, часто весьма краткое время. Кроме того, поиск текстов более чем десятилетней давности через обычные поисковые машины малоэффективен в силу ряда технических ограничений. История сетевого фольклора развивается на наших глазах, но это не значит, что спустя одно-два десятилетия сегодняшние тексты будут оставаться легкодоступными. Таким образом, необходима систематическая научная фиксация сетевого фольклора.

В силу временной и пространственной дистанции между исследователем и носителем фольклора методы сбора классического фольклора, апробиро-

1 См., напр., сетевые сборники анекдотов: *http://www.anekdot.ru*; *http://funpark.ru*; *http://xa-xa.org*; *http://www.joke-zone.co.uk*

2 См., напр., *http://www.knowyourmeme.com*; *http://www.lurkmore.ru*; *http://www.netlore.ru*

ванные десятилетиями полевой работы, не всегда применимы. При этом специфика виртуального «поля» и технологические особенности среды открывают принципиально новые возможности для исследования. Можно зафиксировать один и тот же текст в десятках вариантов, проследить тысячи случаев воспроизведения текста, оценить миграцию текста по сети социальных связей, описать носителей исследуемой фольклорной формы и т. п. Однако фиксация этой информации требует разработки специальных методик полевой работы.

ЧТО, ГДЕ И КОГДА: ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ ПОИСКОВЫХ СИСТЕМ

Определение оптимальных методов поиска фольклора в Сети связано с характером исследуемых текстов. Целый ряд жанров бытует в основном в закрытых для постороннего доступа каналах, например, циркулирует по электронной почте. Поэтому значительное число исследований сетевого фольклора (в том числе «писем счастья», «нигерийских писем», ложных предупреждений и пр.[3]) построено на текстах, которые собираются исследователем через сеть своих личных контактов.

Другие тексты распространены в открытом сетевом пространстве. Значимую роль в поиске и изучении этих форм играют поисковые системы. Рассмотрим в качестве примера развитие юмористического текста «Гусиная фигня», посвященного проблеме «свиного гриппа», эпидемия которого продолжалась в течение нескольких месяцев 2009 году. Приведем отрывок из этого текста[4]:

> Пациент в Зимбабве: Доктор, у меня отваливаются уши.
>
> Врач в Зимбабве: Да это какая-то х**ня! А что вы делали?
>
> Пациент: Е**л гусей.
>
> Доктор: Так у вас гусиная х**ня!
>
> Пациент: Спасибо, доктор! (Умирает)
>
> Доктор (записывает в журнал): Пациент умер от гусиной х**ни.
>
> Пресс-служба минздрава Зимбабве: За прошедшую неделю в Зимбабве умерло две с половиной тысячи человек от голода, пять тысяч четыреста от отравления протухшими бананами и один человек от гусиной х**ни.
>
> Журналист (записывает): Гусиная х**ня вошла в список трех главных причин смертности в Зимбабве.

3 См., напр., Orasan, Krishnamurthy 2002; Kibby 2005; Ланская 2009.

4 В цитируемых сетевых текстах сохранена орфография и пунктуация оригинала.

Новостное агентство: В Зимбабве участились случаи заболевания неизлечимой гусиной х**ней.

Телеканал: Неизвестное ранее заболевание гусиной х**ней выкашивает население Зимбабве. Министерство здравоохранения Зимбабве призывает не паниковать.

Научное светило А: Да, гусиная х**ня не известна науке и в этом ее главная угроза.

Авиакомпания Конго Эйр: Мы прекращаем все полеты в Зимбабве до разрешения эпидемии гусиной х**ни.

Научное светило Б: Власти скрывают! На самом деле гусиная х**ня уже проникла в Европу – в Амстердаме видели чихающего негра с гусем под мышкой.

Пресса (публикует фотографии гусей): Гагакающие убийцы рядом!

Это один из ключевых сетевых текстов рассматриваемого периода, иронически описывающий предполагаемый механизм развития медиапаники вокруг эпидемии. Данный текст только в некоторой степени подпадает под определение фольклора – достаточно легко прослеживается его авторство, с высокой вероятностью приписываемое пользователю сервиса *LiveJournal* под ником *segal*[5]. Однако немедленно после публикации текст разлетается по кириллическому сегменту Интернета, за первые два дня выходит за пределы блогосферы и утрачивает авторство. Словосочетание «гусиная фигня» становится основным мемом Интернета в период эпидемии (в это время фиксируется около тысячи упоминаний, цитаций и воспроизведений в блогосфере и, по данным различных поисковых систем, от шести до восьми тысяч ссылок в кириллическом сегменте Интернета в целом). Употребление мема развивается параллельно динамике активности блогосферы, для которой отмечаются два пиковых периода за период обсуждения проблемы «свиного гриппа» – в мае и октябре–ноябре 2009 года. Следует отметить, что активность блогосферы не связана напрямую с деятельностью профессиональных журналистов Рунета – последние включаются в обсуждение «свиного гриппа» лишь спустя четыре месяца после того, как блогосфера обратила внимание на эту проблему (этот временной разрыв, по-видимому, обусловлен присутствием в кириллическом сегменте блогосферы значительного числа пользователей, находящихся в не-российском информационном пространстве, которое оказалось насыщено информацией об эпидемии намного раньше). После завершения активного обсуждения темы в блогосфере популярность текста резко падает, но его бытование продолжа-

5 См. Сергей Галёнкин. *Медиапаника*, запись в блоге 28.04.09, URL: <http://segal.livejournal.com/704669.html>, [дата обращения: 10.05.2010].

ется (пользуясь термином Терезы Хейд, текст входит в период «стазиса»; см. Heyd 2008: 73). Итак, использование поисковых систем позволило нам выявить автора текста, проследить развитие и оценить уровень востребованности фольклорной формы сообществом, а также наметить соотношение между «пользовательской» и «профессиональной» информационной активностью.

Однако работа с поисковыми машинами предполагает целый ряд проблем. Поисковые машины предлагают пользователю ряд инструментов, которые привлекают внимание исследователей легкостью использования. Так, ряд исследователей при отборе текстов для исследования берет за основу собрания текстов, размещенные на «топовых» для данного запроса (и, как предполагается, наиболее посещаемых) сайтах. Ольга Евгеньевна Фролова при отборе текстов для исследуемого корпуса основывается на данных сайтов-каталогов, облегчающих отбор сетевых площадок, на которых размещены фольклорные тексты (Фролова 2009: 121). Лимор Шифман предлагает сплошной или выборочный анализ сайтов соответствующей тематики, имеющих самый высокий поисковый рейтинг в нескольких поисковых системах (который свидетельствует об их популярности и востребованности), а если такой лидер не выявляется, метод ранжирования. Наиболее популярные по версиям нескольких поисковиков сайты ранжируются по критерию отсутствия ограничений на содержание (цензуры), разнообразия текстов, присутствия сайта в лидирующей совокупности более чем одного поисковика и высокий рейтинг *Google*. Затем отбирается некоторое количество лидеров, и методом случайной выборки вычленяются тексты для анализа (Shifman 2007: 192).

Однако рейтинги страниц, автоматически создаваемые поисковыми машинами, могут строиться на принципиально различных основаниях: как количества посетителей (отражающего востребованность сайта), так и количества ссылок на сайт (отражающего его влиятельность). Самый посещаемый сайт может оказаться далеко не самым значимым звеном в цепи распространения текста – и наоборот, отдельный текст может многократно воспроизводиться на различных ресурсах, но не отражаться в статистике поисковых запросов. Так, поиск по блогам *Yandex* выявляет, в каком количестве текстов было встречено то или иное слово; аналогичный поиск *Google* – сколько раз создавался поисковый запрос с этим словом. Таким образом, если первая система показывает количество воспроизведений текста, то вторая – только интерес к нему. Соответственно, интерпретация рейтинга сайта или текста в разных поисковых системах не может осуществляться без учета механизма формирования рейтинга в каждом конкретном случае.

Затем встает вопрос о достоверности автоматизированной статистики, предлагаемой поисковыми машинами. Легкость получения этой статистики может ввести в заблуждение, поскольку количество найденных по запросу адресов нестабильно: повторяя поиск несколько раз и переходя между страницами выдачи поисковой машины, можно получить самые разные значения. Это говорит о необходимости ручной или автоматизированной проверки количества полученных результатов запроса.

В связи с этим возникает вопрос о полноте и репрезентативности данных, получаемых через поисковые машины. Огромный сегмент Сети в принципе недоступен для индексации, поскольку информация находится на сайтах, закрытых для внешнего (незарегистрированного) пользователя. Кроме того, для поискового робота могут оказаться недоступными сайты, не связанные гиперссылками с другими сайтами. Для решения этой проблемы был создан ряд инструментов (напр., *Yahoo Subscriptions*), позволяющих осуществлять поиск по некоторой (сильно ограниченной) части скрытого контента. Специфические проблемы возникают при поиске файлов, расположенных на ftp-серверах, поскольку напрямую они не индексируются, а специализированные поисковые системы (напр., *Filesearch.ru*) не учитывают при поиске содержимого файлов. Кроме того, индексация отстает по времени от появления новых адресов, что ставит под сомнение возможности синхронного исследования. Анализ блогов затруднен также тем, что при отсутствии постоянной индексации крупными поисковыми системами получить удаленные по времени данные достаточно сложно; разработка специализированных машин для поиска по блогам, микроблогам и новостным группам / телеконференциям *Usenet* (напр., *Blogs.yandex.ru*; *Archivist.visitmix.com*; *Groups.google.com*) несколько улучшила ситуацию, но, тем не менее, область поиска ограничена во времени (если для первого сервиса это последние 8–10 лет, то для второго – несколько недель). При этом при работе с блого-ориентированными поисковыми машинами необходимо учитывать высокую вероятность неполной индексации удаленных и скрытых записей, а также высокий уровень спама – автоматически сгенерированных записей, созданных для влияния на рейтинги. Наконец, база любой отдельно взятой поисковой системы не охватывает всего открытого сетевого пространства. Таким образом, чтобы получить максимально полные данные, необходимо использовать несколько машин, релевантных для исследуемого языкового сегмента (или использовать статистические методы корректировки выборки).

ТЕКСТ В ГИПЕРТЕКСТЕ: ПОИСК ФОЛЬКЛОРА В СЕТИ

Отдельная задача при проведении исследования – формулирование поисковых запросов (и их последующее отражение при публикации результатов). В исследованиях интернет-фольклора нередко встречается ситуация, когда в качестве поискового запроса задается название жанра (напр., «анекдот») или группы текстов (напр., «школьный фольклор»; см. Самоделова 2009: 181). Такая исследовательская стратегия в принципе небесполезна. Так, Филип Гралиньский путем анализа результатов простых поисковых запросов провел работу по анализу речевых формул, «сцепленных» с воспроизведением городских легенд в сети. В результате выяснилось, что формула «третьего лица» («знакомый знакомого», «подруга знакомой» и т. п.) связана с легендами в 2–5% случаев. Фактор устности («слышал такую историю», «слышал похожую историю») значительно повышает шансы на обнаружение городской легенды – до 25%. Достаточно часто используются указания на подлинность информации («подлинный случай», «правда было», «история из жизни») – до 22%. Ожидаемо, самый высокий результат дало прямое указание на жанровую природу текста («городская легенда», «это просто городская легенда» и т. п.) – до 50% случаев (Graliński 2009: 258–259).

Тем не менее, следует обратить внимание на то, что запросы, указывающие на жанр, обычно приводят исследователя либо на сайт-сборник соответствующих текстов (что само по себе полезно, особенно на начальном этапе исследования, однако не дает ни малейшего представления о естественном бытовании этих текстов), либо на сайты, где выложены научные работы по данной проблематике. Только в редких случаях прямой поиск по жанровому наименованию показывает площадки, где воспроизведение текста повлекло за собой обсуждение его носителями. Поэтому наиболее адекватным представляется поиск по ключевым словам, который может быть ограничен по языку текста, дате размещения в Интернете и т. п. Здесь соблюдается общее правило поиска в Интернете: запрос должен быть достаточно специфичен, чтобы дать максимальное количество релевантных результатов, но достаточно широк, чтобы из поля зрения исследователя не выпало значительное количество не вполне соответствующих строгому запросу текстов. Проблема формулирования поискового запроса и методологии поиска поставлена, в частности, в работах Михаила Дмитриевича Алексеевского (2009) и Марии Вячеславовны Ахметовой (2011). Правильно сформулированный запрос дает возможность не только найти максимальное количество случаев воспроизведения текста, но и его варианты, а также оценить динамику бытования текста как в общем в Сети, так и в отдельных ее сегментах. Такие исследования предпринимались, в частности, Биллом

Эллисом (Ellis 2002), Гиселиндой Киперс (Kuipers 2002), Виолеттой Кравчик-Василевской (Krawczyk-Wasilewska 2003) и др.

Другой подход – поиск конкретного текста, к примеру, ранее выявленного исследователем на сайте «архивного типа». Для поиска вариантов текста вводить в поисковое поле его название или первую строку недостаточно – это название может изменяться, утрачиваться и т. д. Для выявления максимального количества вариантов и количественного анализа трансмиссии фольклорного текста необходим также поиск по отдельным ключевым фрагментам текста.

В этом отношении очень показателен текст «Гостевая книга сайта *России.Ру*». Авторство текста приписывается группе КВН «Уральские пельмени» (2003 г.)[6], однако он значительно варьировал после попадания в сетевое пространство. Приведем отрывок из этого текста в наиболее распространенной в Интернете форме:

> *862 год:* Добро пожаловать на наш сайт! Рюрик, Трувор и Синеус.
>
> *9 век:* К нам заходит половецкий князь Кончак: Я ща вас всех!!!
>
> upd: Администратор сайта князь Игорь банит половецкого князя.
>
> *10 век:* Княгиня Ольга - админу сайта: мне кажется, язычество исчерпало себя на Руси...
>
> comment Владимир Ясно Солнышко - княгине Ольге: когда кажется - креститься надо.
>
> *11 век:* Всем прюветы. Ыскренне ваши. Кирилл и Мефодий. Есть клевые шрифты с ижицей.
>
> *13 век:* Реклама на сайте: Приглашаем на экскурсии по Золотому кольцу России.
>
> comment Тевтонские рыцари: Отличная экскурсия. Отличные озера. Спасибо, Александр Невский. От лица всех рыцарей, Карл и Йохан - те, которые ехали в обозе с пенопластом.
>
> *14 век:* Сервер обрушен татаро-монгольскими хакерами. Надолго.
>
> *15 век:* Иван Грозный выложил на сайте статью о воспитании детей. Подробности на Репин.ру.
>
> *16 век:* Реклама на сайте: Экскурсии по золотому кольцу России.
>
> comment Поляки: Отличная экскурсия. Отличные люди. Thxs проводнику. Отличные болота. Поели мухоморов. Кшиштоф женился на царевне-лягушке. Ищите нас через Яндекс. <...>[7]

6 URL: <http://www.youtube.com/watch?v=JGym-AvGs38>, [дата обращения: 08.02.2010].

7 URL: <http://forum.zelek.ru/t7631-1000-letie-interneta-na-rusi-iz-kvn.html>, [дата обращения: 07.01.2013].

Несмотря на то, что внешне «Гостевая книга» представляет собой замкнутую последовательность сюжетов, организованную рамочной конструкцией, она обладает удивительно высокой вариативностью для текста, распространяемого методом «копи-пейст». Развитие данного текста происходит нелинейно, в процессе трансляции он как сокращается, так и расширяется. За время его бытования в интернет-среде возникло не менее 14 версий, отличающихся как на уровне синонимических замен, «вибрирования» (Чистов 2005: 73), так и количеством структурных элементов (от 29 до 54).

В многочисленных вариантах текста может отсутствовать любой элемент, включая первую строку. Более того, в значительном количестве случаев воспроизводится именно вариант, начинающийся со строки «Всем прюветы». Изменения могут касаться также текстов отдельных элементов. В этой ситуации (весьма характерной для сетевого фольклора) проблема выделения ключевых фрагментов для формирования поисковых запросов выходит на первый план.

Как правило, сетевые тексты достаточно легко членятся на смысловые элементы, не превышающие длину стандартного поискового запроса. Это может быть, например, элемент «перечня» или пуант анекдота. При этом первичный грубый поиск (по первой строке) нередко дает более одного варианта. Сопоставляя частотность, с которой отдельные элементы текста встречаются в разных вариантах, можно выделить ключевые элементы, которые с высокой вероятностью будут присутствовать даже в максимально сокращенных вариантах. В случае «Гостевой книги» это, например, фрагменты «Всем прюветы. Ыскренне ваши. Кирилл и Мефодий. Есть клевые шрифты с ижицей», «Сервер обрушен татаро-монгольскими хакерами. Надолго» (присутствуют во всех 14 версиях текста) и т. д. В качестве следующего шага предпринимается поиск по каждому из ключевых элементов, причем оптимален в этом случае строгий поиск по фразе. Такая стратегия позволяет выявить максимальное количество вариантов. При строгом поиске необходим ввод фразы с сохранением особенностей правописания оригинала. Особенно ярко эта необходимость проявляется при работе с текстами с элементами т. наз. «олбанского языка». Так, при поиске вариантов мема «я криведко» поисковый запрос должен сохранять все искажения общепринятой орфографии: по запросу «я креветка» поисковые машины выдадут абсолютно нерелеватные результаты.

Для реализации задачи сплошного обследования применяется автоматизированный сбор данных (использование поисковых роботов, разработанных под конкретную исследовательскую задачу). Так, разработанная Юре Лесковецом, Ларсом Бэкстромом и Джоном Клейнбергом (Leskovec [et al.]

2009) база данных под названием *MemeTracker* позволяет выявить сети распространения информации в Интернете, определить авторство сетевого текста или временную последовательность его распространения, выявить паттерны распространения информации на основе массива в десятки миллионов документов. Ф. Гралиньский в уже упомянутой работе (Graliński 2009) демонстрирует возможности автоматизированного поиска городских легенд в сети. Робот (аналогичный поисковым машинам) собирает записи в блогах, комментарии к новостным статьям, записи на форумах и в группах *Usenet* (т. е. площадки, на которых наиболее часто встречаются городские легенды), затем программа-классификатор (аналогичная спам-фильтру) «процеживает» эти тексты в поисках легенд на основе заданных формальных критериев, например, употребления определенных речевых клише («знакомый знакомого», «слышал такую же историю», «это правда было» и т. д.). В задачу исследователя при этом входит отсев нерелевантных текстов и постоянное дополнение базы критериев новыми на основании анализа уже выявленных текстов. Тем не менее, автоматизированный поиск независимо от того, выполняется ли он общедоступной или кастомизированной поисковой машиной, упускает значительное количество документов. Он подходит для статистической обработки больших массивов, когда задачей является определение наиболее вероятных (но не истинных) механизмов и процессов (Adar, Adamic 2005).

Всё сказанное касается прежде всего текстовых файлов. Выявление графических и видеофайлов до недавнего времени ограничивалось поиском по ключевым словам или названию файла, которые далеко не всегда отражают содержание изображения. Это создавало значительные трудности для исследования визуального фольклора: поиск нужного текста или его варианта требовал просмотра сотен изображений, описанных теми же ключевыми словами, что и искомый файл. Отчасти эту проблему решило внедрение сервиса *Tineye.com*, который успешно применяется при необходимости поиска аналогов незначительно измененного изображения (отличающегося размером, разрешением, некоторыми деталями и т. п.). Эта поисковая машина особенно актуальна для работы с фотожабами[8] и демотиваторами[9], т. к. помогает обнаружить как исходные изображения, так и их варианты. Кроме того, *Tineye.com* помогает оценить распространение в Сети явлений интернет-фольклора, прежде всего англоязычного.

8 Заведомо модифицированное / коллажированное изображение, иногда с добавлением вербальной составляющей.

9 Изображение в черной рамке, снабженное подписью иронического или абсурдного характера.

Поскольку поисковые машины неизбежно упускают определенный процент текстов, автоматизированный поиск необходимо сочетать с ручным: сплошным или выборочным обследованием тематических сайтов и сбором текстов методом «снежного кома». Тексты определенных типов обычно концентрируются на тематических сайтах (на форумах, сайтах анекдотов, в блогах, сборниках текстов для использования в социальных сетях[10]) и т. п. При использовании метода «снежного кома» проводится обследование выявленных зон бытования нужных текстов и поиск упоминаний о них. Эти упоминания помогают как найти тексты при помощи перехода по гиперссылкам, так и собрать их непосредственно у носителей. Так, при поиске текстов «писем счастья» этот подход оказался весьма эффективным: люди, упоминавшие пришедшее им письмо счастья в блоге или на странице социальной сети, обычно позитивно отзывались на просьбу исследователя прислать им этот текст. Доверие, возникающее в результате принадлежности исследователя и информанта к одной социальной сети, вполне достаточно для установления такого рода контакта.

НОСИТЕЛИ ФОЛЬКЛОРА И ОПРЕДЕЛЕНИЕ ЛИЧНЫХ ДАННЫХ КОММУНИКАНТОВ

Возможность зафиксировать социокультурный профиль автора или «републикатора» текста в Сети также поднимает ряд вопросов. Необходимость учета данных о коммуниканте, публикующем фольклорный текст, отмечалась в работах М. Д. Алексеевского (2009: 83), П. А. Бородина (2007), Д. А. Радченко (2012). В настоящее время степень деанонимизации многих сетевых площадок достаточно высока: многие блогеры на персональной странице сообщают о себе множество личных данных – имя, место проживания, возраст; та же ситуация на целом ряде форумов[11]. Достоверность личной информации, предоставляемой коммуникантами в Интернете, не может не вызывать сомнений; тем не менее преднамеренное искажение информации фиксируется относительно редко: коммуниканты, как правило, не считают нужным скрывать такие данные о себе (в отличие от подлинного имени, контактных данных, сведений о семейном положении и пр.). Исключение

10 См., напр., *Продолжение сообщений-цепочек для ВКонтакте*: 04.05.2009 // Life Vkontakte. URL: <http://life-vkontakte.com/20-soobshheniya-cepochki-dlya-vkontakte-2.html>, [дата обращения: 11.02.2011].

11 Для упрощения такого поиска по социальным сетям и открытым источникам типа *Usenet* разработан ряд специализированных машин (*Jibros.com*; *Yoname.com*; *Wink.com*; *Yahoo.com* и т. д.).

обычно составляют узкоспециальные площадки сетевой коммуникации, где искажение информации о поле и возрасте может принести участнику некоторую выгоду или позволить избежать негативных последствий (например, сайты знакомств, форумы, посвященные сексуальной тематике, и т. п.).

При исследовании фольклора, распространяемого через электронную почту, персональную информацию возможно установить по косвенным данным. Так, например, в рассылках «писем счастья» часто встречается полная подпись[12], поскольку для пересылки этих текстов многие пользуются корпоративной почтой (которая сама по себе может служить источником сведений о месте работы адресата и отправителя – так, почтовый адрес ***@ *lukoil.ru* означает, что пользователь является работником «Лукойла»). Такая подпись содержит имя, фамилию, должность, название компании, телефон и адрес компании. Тем самым возможно достоверно установить населенный пункт, в котором живет коммуникант (по адресу или по телефонному коду), примерно определить его социальный статус и профессию (по занимаемой должности), а затем по открытым данным и данным социальных сетей установить по фамилии, имени и городу проживания возраст, нередко также семейный статус, или даже довольно достоверно описать жизненный путь коммуниканта. Для проверки полученной таким образом информации используется дополнительное значение, содержащееся в письме. Так, в городе Нальчике могут проживать десятки женщин по имени Марина Петрова (имя и все дальнейшие данные, приведенные в примере, условны), однако вероятность того, что в компании ООО «Азоттрест» работает больше одной Марины Петровой, крайне низка. К тому же пользователи социальных сетей нередко указывают свои прошлые и настоящие места работы (как в презентационных целях, так и для поддержания сложившихся там социальных связей), что позволяет сопоставить дату отправки письма с корпоративной почты «Азоттреста» с периодом работы в этой организации. Однако такая вероятность все же есть, тем более, что должность указывается в личных данных довольно редко. Для исключения такой вероятности используем метод социальных сетей. Допустим, Марина получила «письмо счастья» от Ирины Воронцовой, а отправила его Анне Дементьевой и Михаилу Горбунову. Из проведенных интервью мы знаем, что такие письма редко отправляются лицам, с которыми коммуникант не поддерживает неформальных связей. Теперь, обнаружив в списке контактов Марины в соци-

12 Публикации рассылок см., напр., o-l-e-g-a-t-o-r, 19.04.2007, запись в блоге // URL: <http://o-l-e-g-a-t-o-r.livejournal.com/102334.html#cutid1>, [дата обращения: 17.12.2010]; Vladimir Ivanov, 16.04.2007, запись в блоге // URL: <http://ivlad.livejournal.com/197579.html>, [дата обращения: 17.12.2010].

альной сети Ирину, Анну и Михаила, мы можем с высокой уверенностью идентифицировать коммуниканта. Разумеется, получить такую полноту информации можно нечасто, и при отсутствии некоторых значимых данных достоверность значительно снижается (проблему могут представлять, например, распространенная фамилия, неполные данные в исходном тексте, отсутствие «третьего вводного» и, соответственно, возможности проверки полученной информации, деперсонализированный адрес и пр.). Также важно понимать, что далеко не все коммуниканты указывают в социальных сетях и на личных страницах достоверную информацию о себе. Причиной искажения информации может быть не виртуальная игра с образом или стремление к приватности, а случайные личные предпочтения. Пример такого рода приводит М. В. Ахметова (2011). Поэтому в ситуации, когда личные данные каждого коммуниканта имеют принципиальное значение для исследования, их лучше уточнять непосредственно у него.

Как показывает практика, в случае «писем счастья» можно определить данные лишь около 20% коммуникантов (Радченко 2012) (что на больших массивах все же составляет десятки, а то и сотни человек и поэтому дает возможность если не статистически достоверного анализа, то оценки сообщества, в котором циркулирует текст). Более того, поиск и проверка личных данных – задача крайне трудоемкая, занимающая много времени и далеко не бесспорная с этической точки зрения. Однако фиксация личных данных необходима, иначе мы лишь приблизительно сможем судить о той среде, в которой возникает и распространяется изучаемый материал. Статистика сетевых площадок обычно является довольно достоверной, но дает слишком общее представление, а в социальных сетях пользователи все чаще закрывают страницы для индексации, и данные, собранные после введения даже одной сетью этой пользовательской опции, имеют ограниченное значение для количественных исследований.

Как бы исследователь ни воспринимал среду бытования сетевого фольклора – как виртуальное пространство, в котором взаимодействуют существа, добровольно отказавшиеся от пола, возраста, лица, биографии в пользу свободной игры идентичностей, или как среду взаимодействия конкретных людей, проявляющих в процессе коммуникации личностные качества, которые обусловлены их реальным социокультурным горизонтом, – он всегда обращает внимание на то, как происходит эта коммуникация. В условиях, когда возможность личного интервью ограничена, а массовость практики требует использования соответствующих методов исследования, наиболее доступный ответ на вопросы о причинах и способах бытования текста дает сам процесс коммуникации, возникшей и зафиксированной в текстовой

форме. Детальное понимание контекста бытования текста в Сети и статистический анализ распространения фольклора позволяют ответить на принципиальные вопросы о том, как устроен процесс коммуникации в Сети, как создаются, видоизменяются и распространяются тексты.

ЛИТЕРАТУРА

Adar E., Adamic L. A. 2005. Tracking Information Epidemics in Blogspace, *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, p. 207–214.

Bell David 2001. *An Introduction to Cybercultures*, London.

Berry D. M. 2004. Internet Research: Privacy, Ethics and Alienation – An Open Source Approach, *The Journal of Internet Research*, No. 14(4), p. 5.

Ellis Bill 2002. Making a Big Apple Crumble: The Role of Humor in Constructing a Global Response to Disaster, *New Directions in Folklore*, No. 6, URL: <https://scholarworks.iu.edu/dspace/bitstream/handle/2022/6911/NDiF_issue_6_complete.pdf?sequence=4>, [дата обращения: 16.02.2012].

Graliński Filip 2009. Tropiąc czarną wołgę w sieci. O poszukiwaniu legend miejskich w internecie, in: *Tekst (w) sieci*, t. 2, red. Anna Gumkowska, Warszawa, p. 253–261.

Heyd Theresa 2008. *Email Hoaxes: Form, Function, Genre Ecology*, Amsterdam / Philadelphia.

Kibby Marjorie D. 2005. Email Forwardables: Folklore in the Age of the Internet, *New Media & Society*, vol. 7(6), p. 770–790.

Krawczyk-Wasilewska Wioletta 2003. Post September 11th: Oral and Visual Folklore in Poland as an Expression of the Global Fear, *Consciousness, Literature and the Arts*, vol. 4, No. 3, URL: <http://www.aber.ac.uk/cla/archive/krawczyk.html>, [дата обращения: 16.02.2012].

Kuipers Giselinde 2002. Media Culture and Internet Disaster Jokes: Bin Laden and the Attack on the World Trade Center, *European Journal of Cultural Studies*, vol. 5, No. 4, p. 451–471.

Leskovec J., Backstrom L., Kleinberg J. 2009. Meme-Tracking and the Dynamics of the News Cycle, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD),* New York.

Orasan C., Krishnamurthy R. 2002. A Corpus-Based Investigation of Junk Emails, *Proceedings of LREC-2002*, Las Palmas, URL: <http://www.clg.wlv.ac.uk/papers/orasan-02b.pdf>, [дата обращения: 28.06.2012].

Shifman Limor 2007. Humor in the Age of Digital Reproduction: Continuity and Change in Internet-Based Comic Texts, *International Journal of Communication*, No. 1, p. 187–209.

Алексеевский М. Д. 2009. «Что мне водка в летний зной...»: Проблемы текстологии фольклора в Интернете, кн.: *Интернет и фольклор*, сост. А. В. Захаров, Москва, с. 71–89.

Ахметова М. В. 2011. Интернет как корпус источников по фольклору и этнографии (региональный аспект), кн.: *Фольклор и этнография: К девяностолетию со дня рождения К. В. Чистова*, отв. ред. А. К. Байбурин, Т. Б. Щепанская, Санкт-Петербург, с. 302–311.

Бородин П. А. 2007. Фольклор в Интернете и вне его: попытка сопоставления, кн.: *Folk-Art-Net: новые горизонты творчества. От традиции – к виртуальности*, сост. А. С. Каргин, А. В. Костина, Москва, с. 39.

Ланская Ю. Н. 2009. Американские "Bogus Warnings" («ложные предупреждения об опасности») и российские «письма несчастья», кн.: *Интернет и фольклор*, сост. А. В. Захаров, Москва, с. 158–169.

Научное знание 2011. Научное знание в условиях Интернета. Вопросы редколлегии, *Антропологический форум*, № 14, с. 21.

Радченко Д. А. 2012. „Очень хочется чуда...“: бытование письма счастья в Интернете, кн.: *Фольклористика и культурная антропология сегодня: Тезисы и материалы Международной школы-конференции-2012*, Москва, с. 280–287.

Самоделова Е. А. 2009. Школьный фольклор в Интернете, кн.: *Интернет и фольклор*, сост. А. В. Захаров, Москва, с. 180–193.

Фролова О. Е. 2009. Анекдот как отражение интересов пользователя, кн.: *Интернет и фольклор*, сост. А. В. Захаров, Москва, с. 117–130.

Чистов К. В. 2005. Текст письменный – текст устный, кн.: *Фольклор. Текст. Традиция*, Москва, с. 68–74.

„Ieškokite mūsų *Jandekse*“: interneto folkloro kaupimo metodika ir problemos

DARJA RADČENKO

Santrauka

Perėjus prie „vartotojų turinio“ koncepcijos (*WEB 2.0*), trečiojo tūkstantmečio pradžioje į pirmą vietą iškilo internetinės bendrijos narių kūrybinės veiklos vertė. Dėl to internete gausu pačių įvairiausių, tarp jų – ir aiškiai folklorinių tekstų. Interneto folkloras ryškiai skiriasi nuo klasikinio savo tariamu prieinamumu: išsikėlus, galima sakyti, bet kokį tiriamąjį uždavinį, nesudėtingais paieškos metodais galima sukaupti tūkstančius relevantiškų tekstų ir vaizdų. Nors klasikinio folkloro rinkimo metodai ne visada tinkami, nes folkloro pateikėją ir tyrėją skiria erdvė ir laikas, virtualaus „lauko“ specifika ir technologiniai terpės ypatumai atveria iš esmės naujas tyrimo galimybes ir reikalauja specialios lauko tyrimų metodikos.

Gausybė žanrų gyvuoja pašaliniams faktiškai neprieinamuose kanaluose (pvz., cirkuliuoja elektroniniu paštu), todėl daugelis interneto folkloro tyrimų grindžiami tekstais, tyrėjo kaupiamais pasitelkiant asmeninius kontaktus. Kiti tekstai cirkuliuoja atviroje interneto erdvėje. Ieškant šių formų ir jas tiriant, svarbios paieškos sistemos. Tačiau, dirbant su ieškyklėmis ir automatiškai generuojamais reitingais, iškyla nemažai problemų. Pirmiausia, svetainės ar teksto reitingas negali būti interpretuojamas neatsižvelgiant į konkretų jo formavimo mechanizmą. Vėliau kyla klausimas dėl ieškyklių pateikiamos automatizuotos statistikos duomenų patikimumo. Didžiulis interneto segmentas iš principo nepasiekiamas indeksuoti. Be to, indeksacija atsilieka nuo naujai pasirodančių adresų, o kai nėra nuolatinės stambių ieškyklių teikiamos indeksavimo paslaugos, gana sunku gauti senesnius duomenis. Taip pat būtina įvertinti galimai nepilną pašalintų ir paslėptų įrašų indeksaciją ir esmingą automatiškai sugeneruotų įrašų, skirtų veikti reitingus, kiekį. Pagaliau jokios paieškų sistemos duomenų bazė neapima visos interneto erdvės. Vadinasi, siekiant kuo tikslesnių duomenų, būtina naudotis keliomis ieškyklėmis, relevantiškomis tiriamame kalbiniame segmente, arba naudotis statistiniais atrankos korekcijos metodais.

Atskiras tyrimo uždavinys yra užklausų formulavimas. Paprastai užklausos, įvardijančios žanrą, tyrėją atveda į svetaines – atitinkamų tekstų talpyklas arba į svetaines, kuriose įkelti kalbamosios problematikos mokslo darbai ir tik retais atvejais aptinkami natūraliai gyvuojantys tekstai. Dėl to pačia tinkamiausia laikoma paieška pagal raktažodžius. Kitas būdas – konkretaus teksto paieška. Maksimaliam variantų kiekiui išryškinti ir kiekybinei folklorinio teksto sklaidos analizei taip pat būtina atskirų esminių teksto fragmentų paieška.

Galimybė fiksuoti socialinį demografinį autoriaus ar teksto „republikatoriaus“ (rus. *республикатора*) profilį kelia virtinę klausimų. Dabar daugelis interneto svetainių gerokai iš-

viešintos, o interneto komunikantų teikiamų asmens duomenų patikimumas išties abejotinas. Vis dėlto tyčinis informacijos klastojimas pasitaiko palyginti retai. Asmens duomenų paieška ir patikra – sunkus ir etiniu požiūriu anaiptol ne neginčijamas uždavinys, tačiau šių duomenų fiksavimas leidžia susidaryti nuomonę apie tą terpę, kurioje gimsta ir sklinda tiriamoji medžiaga.

Kai tiesioginio pokalbio galimybė ribota, o reiškinio masiškumas reikalauja atitinkamų tyrimo metodų, teksto gyvavimo priežastis ir būdus prieinamiausiai apibūdina pats komunikacijos, prasidėjusios ir fiksuotos tekstu, procesas. Detalus teksto gyvavimo internete suvokimas ir statistinė folkloro sklaidos analizė padeda atsakyti į principinius klausimus, kaip sutvarkytas komunikacijos procesas internete, kaip tekstai yra kuriami, keičiami, kaip jie plinta.

Iš rusų kalbos vertė *Povilas Krikščiūnas*

Gauta 2013-01-28