Контент Рунета

Этот информационный бюллетень рассказывает о контенте Рунета.

Основные данные отчета получены от поиска Яндекса. Также использовались данные поиска Яндекса по картинкам и по видео. Данные исследования охватывают только открытые веб-страницы — для того чтобы попасть на них, не требуется ввод логина и пароля.

В данном исследовании под «Рунетом» подразумеваются сайты, написанные на русском, украинском, белорусском или казахском языках, а также сайты на любых языках, размещенные в национальных доменах .am, .az, .by, .ge, .kg, .kz, .md, .ru, .su, .tj, .ua или uz. Рассматривались текстовые копии всех открытых веб-страниц Рунета, которые хранятся в индексе поисковой системы. Рунет меняется очень быстро, и в силу разных технических ограничений база Яндекса не может быть абсолютно точной его копией.

Содержание

Яндекс

1. Введение

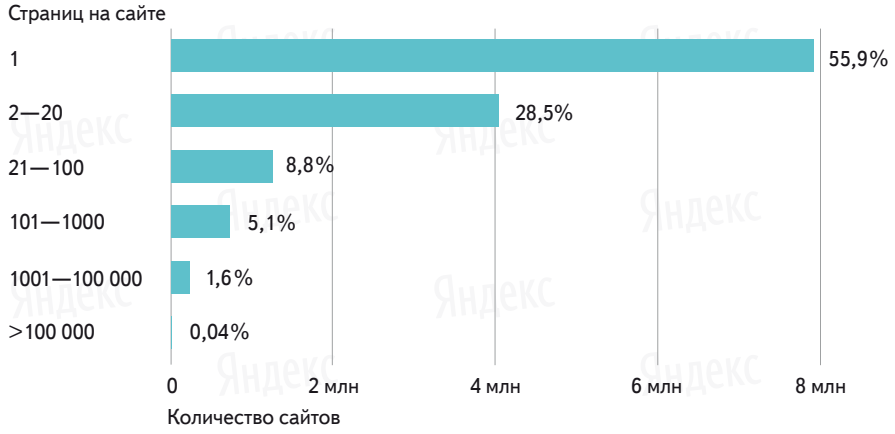
По данным поиска Яндекса на осень 2009, в Рунете — около 15 миллионов сайтов.[1] Это 6,5% от всего интернета.[2] Российские пользователи интернета[3] составляют 2,2% от всех пользователей интернета — то есть на одного пользователя в Рунете приходится больше сайтов, чем в среднем в мире.

Только в текстовом формате (без учета картинок, аудио- и видеофайлов) в Рунете размещено более 140 тысяч Гб данных.[4] Информация в сети распределена неравномерно. 88% всего текста находится менее чем на одном проценте сайтов. Треть всех картинок размещена на четырех крупнейших фотохостингах.

Средний сайт Рунета состоит из 255 страниц, содержит 159 тысяч слов и 204 картинки. Большинство сайтов гораздо меньше среднего — половина сайтов в Рунете состоит всего из одной страницы.

В среднем на одном сайте сейчас столько же страниц, сколько и десять лет назад — в 1999 году средний сайт состоял из 251 страницы. Одна страница занимала тогда около 9 Кб, а по данным на 2009 год — 39 Кб. Всего сайтов в Рунете в 1999 году было в 300 раз меньше, чем сейчас.

Рис. 1. Распределение сайтов Рунета по количеству страниц



По данным поиска Яндекса, лето 2009

По оценке поиска Яндекса, четверть сайтов Рунета — это поисковый спам, то есть страницы, которые почти не содержат полезной информации, и созданы, чтобы привлекать посетителей на другие сайты или влиять на их ранжирование в поисковых системах.

2. Виды информации

Основные виды данных в интернете — это текст и картинки. В интернете они также выполняют служебные функции — с помощью текстов и картинок создается оформление сайтов.

Кроме базовых видов данных в интернете используют флеш, видео и звук. Самый распространенный из них тип передачи информации — флеш — специфичен для интернета. Флеш-объектами могут быть изображения, видеоролики, элементы интерфейса и т.д. Хотя бы один флеш-объект есть почти на 15% сайтов Рунета.

Звуковые файлы и видеоролики встречаются существенно реже. Видеоролики есть где-то на 3% сайтов, а звук в MP3 — менее чем на 0,5%.

1 Сайт — объединённая под одним доменным именем совокупность страниц. Например, страницы с адресами http://site.example.net/a и http://site.example.net/b относятся к одному сайту. А страницы http://b.site.example.net и http://example.net — к разным. Наличие или отсутствие в адресе страницы приставки www не важно, то есть http://www.example.net — это тот же сайт, что и http://example.net.

2 По оценке Netcraft, в октябре 2009 года в сети насчитывалось 230,4 млн сайтов.

3 По данным ФОМ на лето 2009, количество российских пользователей интернета — 37,5 миллионов. Количество пользователей интернета в мире, по данным Internet World Stats на июнь 2009, — 1,7 миллиарда.

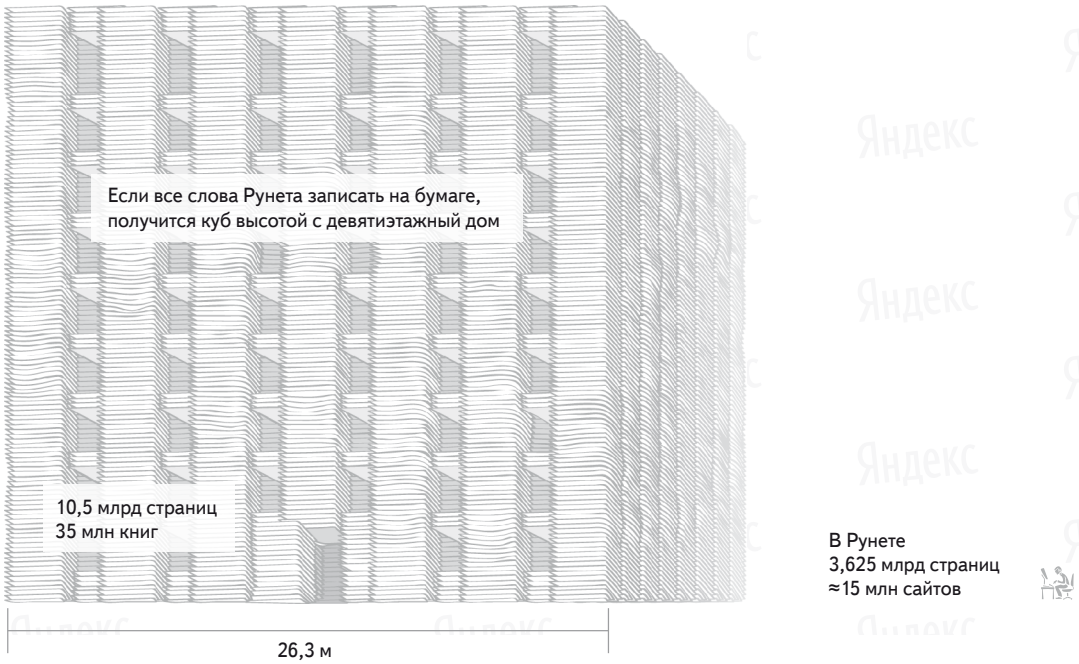
4 Здесь и далее расчеты без учета дублей (идентичных страниц, которые дублируются под разными адресами). С учетом дублей объем данных Рунета достигает почти 200 тысяч Гб.

В 2000 году количество сайтов в Рунете выросло по сравнению с 1999 более чем в три раза, а среднее число страниц на одном сайте в 2000 году уменьшилось до 139. Начиная с 2002 года среднее количество страниц снова стало расти.

2.1 Текст

В открытом доступе — без учета страниц, доступных только после ввода логина и пароля, — в Рунете опубликовано около 2,3 триллиона слов. На каждого российского пользователя приходится более 60 тысяч слов — этого хватило бы на книгу стандартного формата в 280 страниц.

Рис. 2. Весь текст Рунета на бумаге



По данным поиска Яндекса, осень 2009

89% всех сайтов содержат совсем немного текста — в среднем по 1630 слов, как полторы журнальных страницы. На один большой сайт (таких менее 1%) приходится в среднем 18 миллионов слов — объем текста небольшой домашней библиотеки из 250-300 книг.

2.2 Картинки

По данным Яндекса на лето 2009, в Рунете размещено по крайней мере 1,6 миллиарда уникальных изображений [5] — это фотографии и рисунки, элементы оформления страниц, рекламные баннеры и т. п. Картинок, которые можно увидеть в Рунете, в том числе тех, которые отображаются сразу на нескольких сайтах, несколько больше — около 2,1 миллиарда. То есть в среднем где-то две трети картинок можно увидеть только на одном сайте, а остальные — на двух и более.

Каждый третий сайт не содержит ни одной картинки, а еще половина использует для оформления не более десятка изображений.

5 Картинок, проиндексированных поиском Яндекса, у которых есть уникальный адрес.

Рис. 3. Распределение сайтов Рунета по количеству картинок

По данным Яндекс.Картинок, лето 2009

В общем количестве картинок не учтены фотографии, размещенные на крупных фотохостингах.[6] На четырех крупнейших фотохостингах Рунета — Photofile.ru, Radikal.ru, Фото Mail.ru и Яндекс.Фотки — находится, по их собственным оценкам, в общей сложности почти 800 миллионов картинок, загруженных пользователями. То есть на одного пользователя Рунета приходится в среднем 21 фотография на хостингах и еще 57 картинок с остальных сайтов.

6 Полностью все фотографии, размещенные на фотохостингах, роботом Яндекса не индексируются. Их очень много, и часть фотографий размещена на закрытых страницах — например, с доступом только для друзей или за паролем.

2.3 Видео

Видео в Рунете популярнее звука — в том числе благодаря видеохостингам, позволяющим легко добавлять новые видео и вставлять уже загруженные ролики на другие сайты.

На крупнейших видеохостингах Рунета без учета файлобменных и социальных сетей, а также YouTube.com[7] размещено, по данным поиска Яндекса на лето 2009, 7,2 миллиона видеороликов. Ролик, размещенный на видеохостинге, можно легко вставить на любую страницу, где его можно будет просматривать. Таких видеовставок в Рунете — по крайней мере 19,1 миллиона (в том числе с YouTube), их можно увидеть по крайней мере на 2,4% сайтов Рунета.

7 Точное число русскоязычных роликов и роликов, размещенных пользователями Рунета на сайте YouTube.com, неизвестно.

Другой способ распространения видеороликов — с помощью прямой ссылки на видеофайл — популярен гораздо меньше. Его используют около 0,7% сайтов Рунета.

Рис. 4. Распределение видеороликов[8] по длительности

По данным Яндекс.Видео, лето 2009

8 Распределение посчитано по базе видеороликов, известных сервису Яндекс.Видео.

2.4 Звук

Самый популярный формат звуковых файлов в сети — MP3. Сайтов, где есть ссылка на MP3-файлы, в десять раз больше, чем тех, где встречаются ссылки на файлы в форматах WAV, WMA и RAM. В целом звук нельзя назвать распространенным в открытом (доступном без регистрации и ввода пароля) интернете типом информации. Сайтов, где в открытом доступе выложены MP3-треки, — менее 0,5% от общего количества.

Рис. 5. Распределение MP3-треков по длительности[9]

По данным поиска Яндекса, лето 2009

9 Не все эти MP3-файлы расположены в Рунете.

Значительная часть треков, по всей видимости, — музыкальные композиции. Кроме того, заметную долю звука в интернете занимают аудио-подкасты — записанные пользователями выступления на разные темы. Сайт rpod.ru, посвященный подкастам, содержит более 450 тысяч аудиотреков и входит в число крупнейших хранилищ MP3-файлов.

3. Язык Рунета

Основной язык для 91% сайтов Рунета — русский.[10] 2% сайтов написаны на украинском, 1% — на белорусском и менее одной десятой процента — на казахском. Самый распространенный иностранный язык — английский. Он основной для 3% сайтов.

[10] Сайт считается русскоязычным, если на русском языке написано более половины его страниц.

3.1 Частоупотребляемые слова[11]

Рис. 6. Самые частоупотребляемые существительные и прилагательные русского языка в Рунете и в Новом частотном словаре русской лексики

По данным поиска Яндекса и Нового частотного словаря русской лексики, лето 2009

[11] Данные о частоте слов в современном русском языке получены из Нового частотного словаря русской лексики (создан Институтом русского языка им. В. В. Виноградова РАН на основе Национального корпуса русского языка, www.ruscorpora.ru, http://dict.ruslang.ru/freq.php).

Существительные, распространенные в текстах на сайтах и в письменных бумажных[12] текстах, совпадают очень мало. Это неудивительно: топ-20 популярных в интернете существительных наполовину состоит из интернет-терминов, которые относятся не только к языку, на котором говорят и пишут пользователи, но и к интерфейсам *(форум, регистрация, комментарий, поиск* и т.п.). Такие слова, как *новость, тема, карта* и *игра,* на самом деле тоже отражают специфику интернета. На многих сайтах есть раздел *Новости, тема* — это тема на форуме, *карта* — оглавление сайта, *игра* — это компьютерные офлайн- и онлайн-игры.

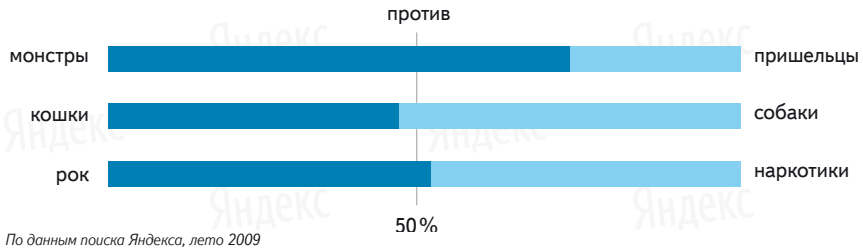
[12] В исследовании все тексты, составляющие Национальный корпус русского языка, называются «бумажными». В реальности помимо текстов, которые были опубликованы на бумаге, в Корпус также входят устные и электронные тексты — их доля от общего объема текстов составляет менее 10%.

Кроме того, в число распространенных в интернете слов попали коммерческие — *цена, компания и товар*. Они встречаются на многих сайтах, принадлежащих коммерческим компаниям, которые предлагают разного рода товары и услуги. В бумажных текстах эти слова распространены существенно меньше — например, слово *товар* встречается там в десять раз реже, чем в интернете, а слово *цена* — почти в шесть раз реже.

Частотные прилагательные в языке Рунета и Новом частотном словаре русской лексики похожи гораздо больше, чем существительные. Топы глаголов также в значительной мере схожи. Только четыре слова попали

в первую двадцатку сетевого рейтинга глаголов и не вошли в общий языковой рейтинг — *находить, скачивать, покупать* и *зарегистрировать*.

Рис. 7. Соотношение количества сайтов, на которых встречается слово из пары



По данным поиска Яндекса, лето 2009

3.2 Эмоции

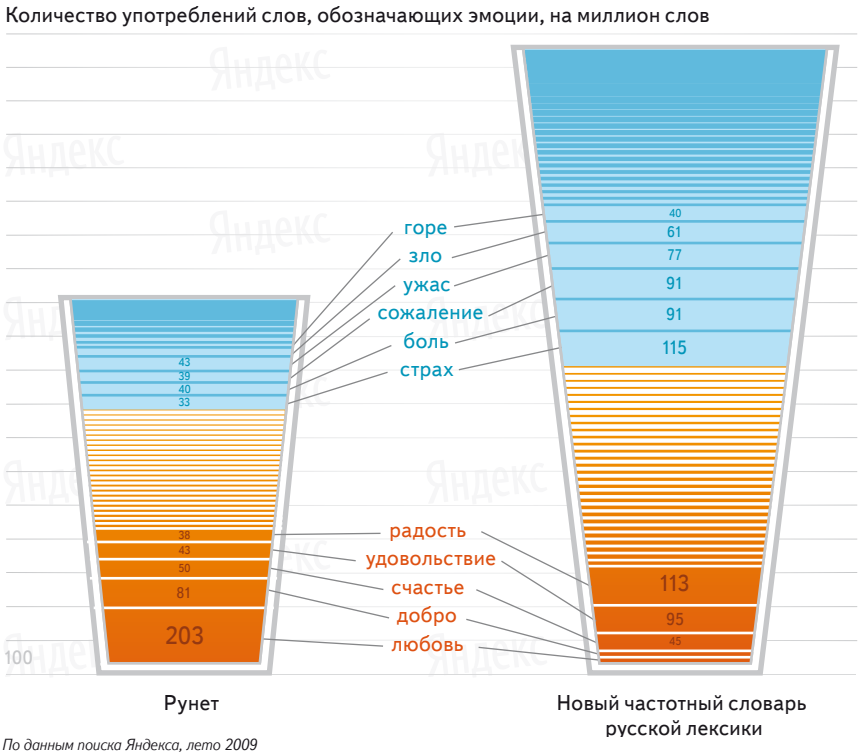
Слова, обозначающие позитивные эмоции и чувства, в интернете встречаются в два раза чаще, чем негативные.

В «бумажном» русском языке слова, обозначающие какие-либо чувства, используются в 1,8 раза чаще, однако негативных эмоций там больше, чем позитивных.

Настроение пользователей интернета можно оценить не только по словам, но и по смайликам. Веселые смайлики популярнее грустных — сайтов, где есть улыбающиеся смайлики, в 2,5 раза больше, чем сайтов, где хотя бы раз появлялись грустные.[13]

13 Смайлики — это сочетания :), :-), :(, :-(, а также повторение трех и более круглых скобок одного типа подряд. Смайлики в виде картинок не учитывались.

Рис. 8. Относительная частота существительных, обозначающих позитивные и негативные эмоции

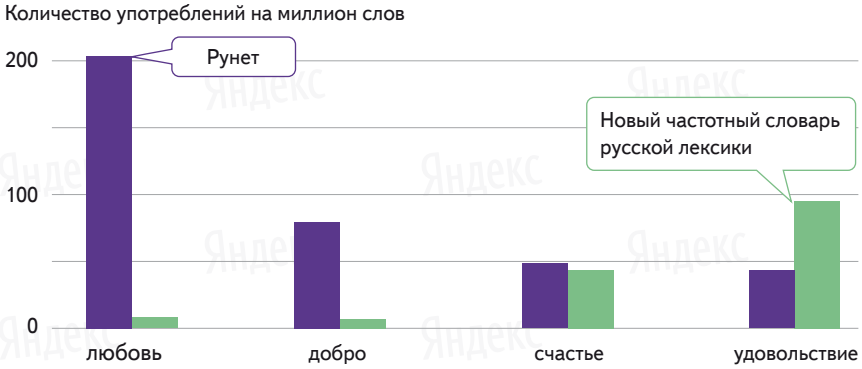


По данным поиска Яндекса, лето 2009

Стакан Рунета полон чуть более, чем наполовину.

Самые частоупотребляемые в интернете слова, обозначающие позитивные эмоции, — это *добро* и *любовь*. В «бумажном» языке эти существительные не вошли даже в топ-50 популярных слов-эмоций.

Рис. 9. Относительная частота самых распространенных в Рунете слов-эмоций

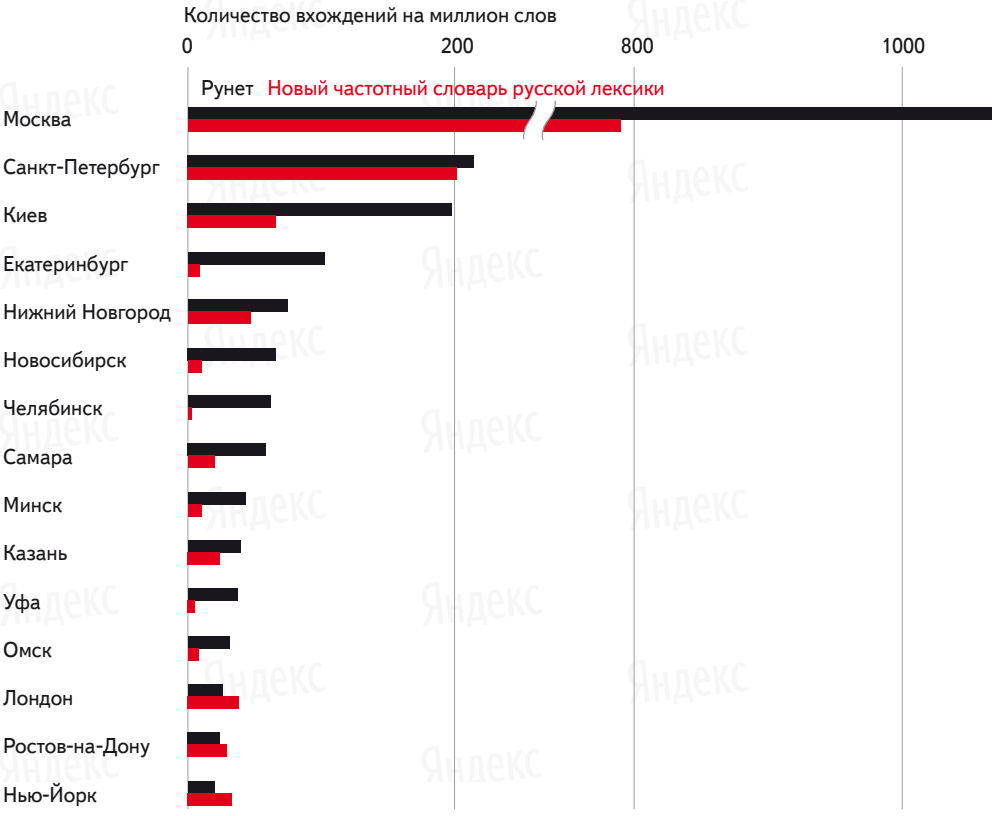


По данным поиска Яндекса, лето 2009

3.3 Географические наименования

По сравнению с бумажными текстами в интернете больше пишут про регионы России и меньше — про города дальнего зарубежья.

Рис. 10. Относительная частота названий городов в Рунете и офлайне



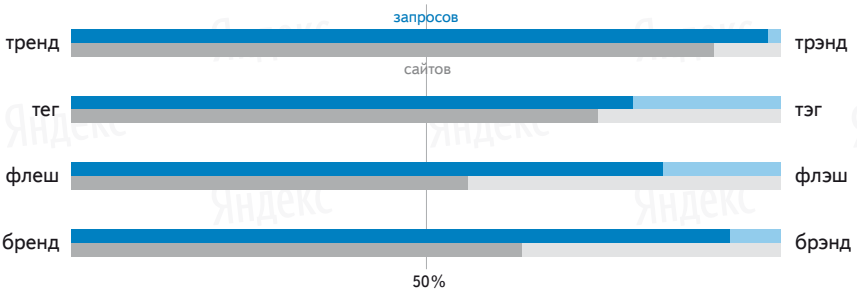
По данным поиска Яндекса, лето 2009

Названия российских городов-миллионников встречаются в сети в полтора раза чаще, чем в бумажных текстах. Отчасти это происходит из-за обилия профилей пользователей на различных форумах и блогохостингах. Среди прочих регистрационных данных пользователи часто указывают город, в котором живут.

3.4 Новые слова

Новые слова поначалу имеют несколько вариантов написания, однако рано или поздно остается один наиболее употребляемый, который становится нормой. Так уже произошло, например, с транслитерацией слова trend. Сейчас сайтов, где это слово написано как *тренд*, почти в десять раз больше сайтов, предпочитающих написание *трэнд*. С тем же, как писать слово flash, в Рунете пока нет определенности. То, к чему склоняется язык, хорошо видно по статистике поисковых запросов.

Рис. 11. Соотношение количества сайтов, на которых встречается слово из пары, соотношение поисковых запросов



По данным поиска Яндекса, сервиса wordstat.yandex.ru, лето 2009

3.5 Ошибки

Орфографических ошибок и опечаток в текстах, размещенных в интернете, не так много. Даже для тех слов, в которых часто делают ошибки, — например, *педиатр* (популярная ошибка — *педиатор*), *агентство* (распространенный неправильный вариант — *агенство),* *трансцендентально* (*трансцедентально*) — средняя доля ошибок не превышает 5—6%.

Доля сайтов, содержащих ошибки в каком-либо слове, часто оказывается больше доли ошибочных написаний этого слова. Например, на семнадцать употреблений слова *агентство* неправильно написано только одно, но ошибка в этом слове встречается на каждом третьем сайте, рискнувшем его использовать.

В масштабах Рунета даже сравнительно небольшая доля ошибок означает огромные числа. 5,78% неправильных написаний слова *агентство* в Рунете — это 21 миллион *агенств*.

Рис. 12. Относительная частота ошибки для слова агентство, доля сайтов с ошибкой

По данным поиска Яндекса, лето 2009

В некоторых случаях грамматически неправильные формы встречаются чаще, чем правильные. Например, сайтов, которые образуют множественное число от слова *брелок* по правилам — *«брелоки»*, меньше, чем сайтов с *«брелками»*. Та же ситуация и с глаголом *победить*. Строго по правилам, у этого глагола нет формы первого лица будущего времени. Однако сайтов, использовавших форму *«победю»*, в три раза больше, чем тех, где выбрали грамматически правильную форму *«одержу победу»*. Формы *«побежду»* и *«побежу»* употребляются очень редко.

3.6 Изменение норм русского языка

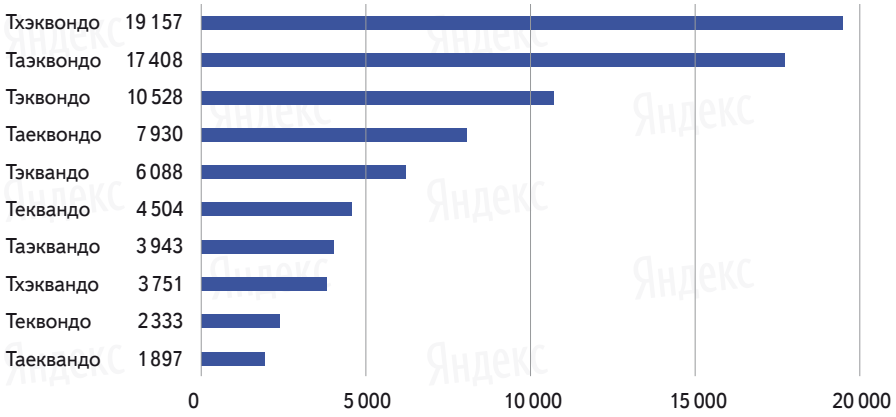
В том, что касается определения рода кофе, Рунет пока придерживается мужского рода. Сайтов, где есть *«хороший кофе»*, в 12 раз больше, чем сайтов, пишущих *«хорошее кофе»*. *«Черный кофе»* встречается 16 раз чаще чем *«черное кофе»*, а *«растворимый кофе»* — в 19 раз чаще, чем «*растворимое*».

Сайтов, где встречается слово *брачующиеся,* в пять раз больше, чем тех, где употребляется равноправная форма *брачащиеся*, и почти в 19 раз больше, чем тех, где есть устаревшее *брачущиеся*.

Карате в Рунете употребляется почти в два раза чаще, чем второй вариант — *каратэ*.

Если для карате распространено только два названия, то для другого восточного единоборства — тхэквондо или таэквондо — встречается по крайней мере восемь вариантов.

Рис. 13. Соотношение количества сайтов, на которых встречаются разные варианты написания



По данным поиска Яндекса, осень 2009

Приложение 1. Основные цифры и факты

По данным поиска Яндекса на осень 2009, в Рунете — около 15 миллионов сайтов. Это около 6,5% от всего интернета. Российские пользователи интернета составляют 2,2% от всех пользователей интернета — то есть на одного пользователя в Рунете приходится больше сайтов, чем в среднем в мире.

Только в текстовом формате (без учета картинок, аудио- и видеофайлов) в Рунете размещено более 140 тысяч Гб данных. Информация в сети распределена неравномерно. 88% всего текста находится менее чем на одном проценте сайтов.

Средний сайт Рунета состоит из 255 страниц, содержит 159 тысяч слов и 204 картинки. Большинство сайтов гораздо меньше среднего — половина сайтов в Рунете состоит всего из одной страницы.

По данным Яндекса на лето 2009, в Рунете размещено по крайней мере 1,6 миллиарда уникальных изображений — это фотографии и рисунки, элементы оформления страниц, рекламные баннеры и т. п. Картинок, которые можно увидеть в Рунете, несколько больше — около 2,1 миллиарда.

Топ-20 популярных в интернете существительных наполовину состоит из интернет-терминов, которые относятся не только к языку, на котором говорят и пишут пользователи, но и к интерфейсам.

Слова, обозначающие позитивные эмоции и чувства, в интернете встречаются в два раза чаще, чем негативные. Самые частоупотребляемые в интернете слова, обозначающие позитивные эмоции, — это *добро* и *любовь*.

Веселые смайлики популярнее грустных — сайтов, где есть улыбающиеся смайлики, в 2,5 раза больше, чем сайтов, где хотя бы раз появлялись грустные.

По сравнению с бумажными текстами в интернете больше пишут про регионы России, и меньше — про города дальнего зарубежья.

Орфографических ошибок и опечаток в текстах, размещенных в интернете, не так много. Даже для тех слов, в которых часто делают ошибки (например, *педиатр, агентство, трансцендентально*) средняя доля ошибок не превышает 5-6%.