БАШКИРСКИЙ ИНТЕРНЕТ: ЛЕКСИКА И ПРАГМАТИКА В КОЛИЧЕСТВЕННОМ АСПЕКТЕ

Орехов Б. В. (nevmenandr@gmail.com)
Российская академия народного хозяйства
и государственной службы при Президенте РФ,
Башкирский государственный университет

Галлямов А. А. (azamat.gallyamov@gmail.com)
Башкирский государственный университет

В работе представляются результаты подсчётов, проведённых в башкироязычном сегменте Интернета. Лексикостатистика показывает ориентированность башкирского интернета не на общение или передачу актуальной информации, а на представление официально-деловых документов и прочие формы представительского присутствия башкирского языка в Сети.

Ключевые слова: башкирский язык, лексикостатистика, прагматика, Интернет

BASHKIR INTERNET: LEXIS AND PRAGMATICS IN A QUANTITATIVE ASPECT

Orekhov B. V. (nevmenandr@gmail.com),
Russian Presidential Academy of National Economy and Public Administration, Bashkir State University

Gallyamov A. A. (azamat.gallyamov@gmail.com),
Bashkir State University

The paper deals with a quantitative aspect of a Bashkir language area of the Internet. We analyze the results of the special crawler's work. Our crawler has indexed Bashkir sites and collected the linguistic valuable data of a word-form frequency. This data differs from the word frequency in Bashkir printed texts or in Russian Internet. Most of the frequent words are marked as official. There is nearly an absence of obscene words. That means that the Bashkir Internet is not designed for any kind of communicative practices but the main goal of it is the message of the existence of the Bashkir language and it's presence on the Web. Internet terms like "site" and others are rare in the Bashkir web-area. Also not so frequent in the Bashnet such popular words in Runet as "job", "mobile phone" and others.

Key words: The Bashkir language, word frequency, pragmatics, Internet

По аналогии с Рунетом постепенно появляются и другие понятия, распространяющиеся на сетевые сообщества, уже не связанные напрямую с доменами первого уровня. Например, вполне устойчивым и употребительным понятием является «Татнет», то есть татарский сегмент всемирной паутины[1]. Существуют национальные сегменты Сети и в зарубежном интернете. Об их объёмах можно судить по размерам Википедии на соответствующем языке. К примеру, на странице статистики интернет-энциклопедии на настоящий момент можно получить такие цифры: каталанский — 370.279 статей, валлийский — 35.614, бенгальский — 23.211, шотландский (гэльский) — 9.792, нижнесаксонский — 4.741. Для сравнения для энциклопедий на некоторых российских национальных языках эти цифры выглядят так: якутский — 8.132, удмуртский — 3.066, кабардино-черкесский — 649. Для башкирского этот показатель составляет 15.935.

Ввиду малого объёма и числа пользователей башкирским интернетом до сих пор не интересовались. Мы попытаемся дать характеристику Башнета на основе того материала, который получен нами в результате нескольких измерительных операций, осуществлённых автоматическим способом.

---

[1] Существует и статья «Татнет» в Википедии, и специально посвящённая предмету вышедшая двумя изданиями книга Айнура Сибгатуллина «Татарский интернет» [Татнет].

Под Башнетом мы будем понимать совокупность доступных в сети Интернет текстовых документов, написанных на башкирском языке.

По нашим данным на конец января 2012 года объём башкирского сегмента составляет 66.199 страниц[2]. Примечательно, что эта величина не находится в линейной зависимости от численности говорящих на башкирском языке. Достаточно сравнить полученные данные с объёмом Рунета. Постоянно растущий Рунет составлял на декабрь 2005 года приблизительно $2538 \cdot 10^6$ документов [Сегалович и др. 2006], на осень 2009 года $3825 \cdot 10^6$ страниц [Контент]. Таким образом, по самой грубой оценке, современный объём Рунета не может быть менее 5 миллиардов документов, то есть разница с объёмом Башнета составляет 5 порядков. По оценке Росстата население России на 1 января 2012 года составляет 143.030.106 человек [Росстат]. При этом число говорящих на башкирском языке в Российской Федерации оценивалось в 2002 году в 1.380.000 (по данным переписи) [Ethnologue]. Таким образом, разница в числе говорящих составляет два порядка.

Разумеется, Интернет — это особая область проявления языковой компетенции. Помимо числа говорящих для количественной оценки этой сферы имеет значение и число пользователей Интернета среди носителей языка. Президент Башкортостана Рустэм Хамитов в своём блоге в записи от 30 сентября 2011 года назвал следующие цифры: «В республике только чуть более 30% населения постоянно пользуются интернетом. В России — около 40%»[3]. Итак, доля пользователей среди русскоговорящего и башкироговорящего населения сопоставима, а вот пропорции к числу сайтов на этих языках — нет.

Таким образом, несмотря на рост потенциальных пользователей Интернета среди башкир, Башнет остаётся крайне ограниченной областью Всемирной сети.

В проведённом нами замере участвовало 30 доменных имён. Это не выборка, а более-менее полный перечень сайтов, которые вообще содержат тексты на башкирском языке.

Многие из обследованных сайтов являются дву- (и более) язычными. На них присутствуют как тексты на башкирском, так и на русском, и на английском языке. Таким образом, строго говоря, исследовались не все страницы этих 30 сайтов, а только те, которые содержат башкирские слова. Таким образом, нахождение именно башкирского текста на интернет-странице представляет собой некоторую техническую проблему, от решения которой зависят конечные статистические данные. В качестве рабочей гипотезы было принято, что в строке башкирского текста будут обязательно присутствовать слова, содержащие специфические для башкирской графики отсутствующие в русском языке буквы: *ң*, *ҙ*, *ә*, *һ*, *ө*, *ҡ*, *ғ*, *ү*, *ҫ*. Только такие строки и брались для расчёта.

Показательно, что в Рунете по данным компании «Яндекс» на 2009 год опубликовано около 2,3 триллиона слов [Контент]. Методика подсчёта

---

[2] Весь объём сайтов, которые публикуют тексты на башкирском языке, составляет 87.462 страницы, из них 21.263 не содержат башкирских текстов.

[3] http://blog-rkhamitov.livejournal.com/44543.html

специалистов «Яндекса» такова, что Башнет входит в эту цифру. По нашим данным в Башнете содержится 27.252.251 слово. Таким образом, разница в объёмах Рунета и Башнета в словоупотреблениях будет составлять те же 5 порядков, что и при сравнении страниц.

Логически сайты Башнета следует разделить на две группы. В первую войдут те сайты, которые не содержат специального веб-ориентированного контента. Это будут страницы республиканских СМИ, которые печатаются на бумаге или выходят в радиоэфир и одновременно с этим вывешивают свою продукцию в Интернете. Мы будем называть такие сайты «медийными». Во вторую группу будут входить все остальные сайты, то есть те, наполнение которых изначально рассчитано на Интернет. Такое деление представляется оправданным, так как современные мыслители считают, что «глобальная сеть формирует новые языковые системы — знаки, символы, гипертексты без конца и начала, “языковые игры”, требующие анализа» [Тарасенко 2000]. Медийные же сайты не вносят в эти новации ничего от себя, так как изначально ориентируются на традиционные формы бытования своей продукции, а сетевое присутствие является для них всего лишь необязательным дополнением, никак не влияющим на содержательный аспект представленных материалов.

Медийная группа сайтов — это примерно треть всего перечня (9 из 30 сайтов). Однако её доля в страницах почти достигает половины (41.298 страниц из 87.462), а в словоупотреблениях даже превышает половину (15.155.408 из 27.252.251 слов). Именно в этой группе находится самый объёмный сайт Башнета — страница газеты «Йәшлек». На нём опубликовано 8.470.444 слова, что составляет 55.89% всего объёма словоупотреблений на медийных сайтах и 31% от всего корпуса Башнета.

Наиболее значительный ресурс в во второй группе — свободная мультиязычная интернет-энциклопедия «Википедия». Блогов, в которых публиковались бы тексты только на башкирском языке, нам обнаружить не удалось, но нашлись два личных дневника, в которых такие тексты появлялись время от времени, перемежаясь с русскими.

Тематический и жанровый обзор свидетельствует, что для башкирского сегмента Сети фактически отсутствуют все основные сервисы, обычно занимающие большую часть времени пользователя. Иными словами, нет того, зачем обычно приходят в Интернет: электронная почта, поисковая система, информация о погоде, современные новостные ресурсы, социальные сети.

Как минимум на треть Башнет является всего лишь приложением к СМИ. В гораздо меньшей степени, чем это развито в Рунете и всемирной паутине в целом, Башнет является ресурсом общения. Это означает, что, по большому счёту, размещённые в Интернете башкирские тексты не предполагают существования читателя. Заходящий в сеть носитель башкирского языка воспользуется для своих повседневных нужд сайтом на другом языке, просто потому что таких востребованных сервисов в Башнете нет. Главным образом, башкирские сайты экстравертны — они обращены к внешнему миру, к сообществу людей, которые не являются носителями башкирского языка, и главным их сообщением является сигнал о существовании башкирского языка как такового.

По частотности лексики язык Интернета заведомо отличается от языка печатных изданий (это ещё одна причина разделять сайты на медийные и остальные). На материале русского языка было сделано наблюдение, что «существительные, распространенные в текстах на сайтах и в письменных бумажных текстах, совпадают очень мало. Это неудивительно: топ-20 популярных в интернете существительных наполовину состоит из интернет-терминов» [Контент]. Для Башнета это не так. Самым частотным словом из лексики сетевой коммуникации в башкирском интернете выступает «форум», который находится в средине третьего десятка. Следующий за ним специфический интернет-термин — «Википедия» — обнаруживается в середине четвёртого десятка.

Очевидно, что такое положение вещей обусловлено неразвитостью интернет-коммуникации на башкирском языке и, соответственно, отсутствием сетевого метаязыкового уровня.

Для лучшего уяснения ситуации с лексическим составом верхней части частотного словаря Башнета можно использовать сравнение получившегося в результате обследования башкирских сайтов частотного списка с существующими частотными словарями башкирского языка. Мы использовали составленные З. А. Сиразитдиновым словари на основе научных текстов (Н), художественной прозы (П) и произведений писателя Даута Юлтыя (Ю) [Сиразитдинов 1997, Сиразитдинов 2002, Частотный].

Соединительный союз *һәм* ‘и’ и послелог *менән* ‘с’ вполне ожидаемо занимают верхние строчки в частотнике Башнета. Столь же высокие позиции у этих слов и в указанных частотных словарях башкирского языка.

Но вот третье слово в частотном словаре Башнета в сравнении с другими частотниками довольно неожиданно, и, как мы считаем, отражает крен в сторону официального языка. Это послелог *буйынса* ‘поэтому, вследствие этого’ (78.230 вхождений), который не встречается в числе 50 наиболее употребляемых слов ни в одном из частотных словарей башкирского языка. Употребительность этого послелога может быть связана с такими случаями, как этот: *баш мөхәррирҙең дөйөм мәсьәләләр буйынса урынбаҫары* ‘заместитель редактора **по** общим вопросам’. Такого рода контекстов достаточно много, что лучше всего иллюстрирует мысль о том, что Башнет в его нынешнем виде, по сути, не предназначен для общения и более всего отражает норму официально-деловых текстов.

Четвёртое место в частотнике занимает слово *башҡорт* ‘башкирский’ (66.945 вхождений). В других словарях: Н: 15 место (1.126 вхождений); в П и Ю: не входит в первые 50 слов (в Ю: 68 место). Как и слово *башҡортостан* (17 место и 43.272 вхождений), это, прежде всего, слово официального языка, а кроме этого ещё и главный элемент в ряду знаков саморепрезентации языка.

Неожиданно высокие позиции в частотном списке занимают формы *ағасым* ‘моё дерево’ (6 место и 59.792 вхождений), *ҡошом* ‘моя птица’ (8 место, 59.739 вхождений), *ырыуым* ‘мой род’ (10 место, 47.789 вхождений), *ораным*

‘мой клич’ (16 место, 43.791 вхождений). Такое необычное положение вещей объясняется форматом башкирского форума (http://www.bashforum.net/), регистрация на котором предполагает указание пользователем признаков единства своего племени[4], которые отображаются на каждой странице рядом с добавленной им репликой. Таким образом, формы *ағасым*, *ҡошом*, *ырыуым* и *ораным* встречаются на сайте столько раз, сколько реплик пользователей он показывает.

Перенесённая в Сеть традиционная башкирская форма определения себя в сопоставлении с окружающими, особенно знаменательна на фоне высказанной выше гипотезы о самом существовании Башнета как способе определения языка в современных условиях информационного общества.

Глагол *үҙгәртергә* ‘изменить’ (9 место и 49.017 вхождений в частотном списке Башнета) тоже не характерное часто употребляемое слово в башкирском языке за пределами Интернета. Его высокие позиции в перечне лексики объясняются дизайнерским решением «Википедии», в которой слово «изменить» сопровождает каждый раздел статьи (которых на одной странице может быть много). А так как «Википедия» составляет 31,28 % всего объёма немедийного Башнета, частотности слова *үҙгәртергә* удивляться не следует. То же относится и к слову *мәҡәлә* ‘статья’ (12 место 45.984 вхождений), отсутствующему в ряду самой частотной лексики в Н, П и Ю. При этом можно было бы ожидать высокие позиции этого слова в Н, однако, в реальности мы сталкиваемся с иным положением вещей.

Специального очерка заслуживает слово *йүнәлештәр* ‘направления’ (14 место, 44.049 вхождений). Его мы тоже вряд ли смогли бы ожидать на верхних позициях частотного списка, составленного на основе сбалансированного корпуса текстов данного языка (в Н, П и Ю не входит в число наиболее частотных). Однако чрезмерно канцеляризированный язык Башнета демонстрирует аномальную популярность этой словоформы. Она появляется в характерных контекстах, в которых речь идёт о развитии новых направлений, стремлении к новым направлениям исследований, о работе по следующим направлениям, об основных направлениях работы. Самая частая коллокация с участием этой лексемы в корпусе Башнета: *төп йүнәлеш* ‘основное направление’. Обилие контекстов свидетельствует, что *йүнәлеш* относится к ряду особенно «модных» слов с предельно абстрактным значением, об обилии которых в современной русской языковой ситуации писал М. А. Кронгауз [Кронгауз 2007] (ср. также *нанотехнологии*, *инновации*, *модернизация* и т. д.). Вот несколько репрезентативных иллюстраций: *ә 2009 йылдан факультетта бөтә йүнәлештәр буйынса ла студенттарҙың олимпиадаһы уҙғарыла* ‘а с 2009 года на факультете проводятся олимпиады по всем направлениям’; *Баймаҡ районында тап ошо йүнәлештәрҙе үҫтереү өсөн бөтә мөмкинлектәр бар* ‘в Баймакском районе для развития вот этих вот направлений есть все возможности’.

---

[4] См. статью «Структура родоплеменной организации башкир» в Краткой энциклопедии «Башкортостан» (1996).

Ещё одну неожиданность составляет 15 позиция и 43.955 вхождений словоформы *бейҙәр* ‘бии, князья’, отсутствующей среди частотных в Н, П и Ю. Высокую частотность этому слову обеспечивает «Википедия».

Что характерно, мы не находим слов *буйынса*, *ағасым*, *ҡошом*, *ырыуым*, *үҙгәртергә*, *мәҡәлә*, *йүнәлештәр*, *бейҙәр* в верхней части частотного списка, составленного нами на материале сайтов башкирских СМИ. В то же время наиболее закономерно употребительные служебные слова *һәм*, *был*, *менән*, *өсөн* присутствуют и там, и там.

Отдельную проблему составляет форма *йылға* (19 место, 40.237 вхождений), которая одновременно может быть и сущ. в им. п. ‘река’, и сущ. в дат. п. ‘год’. Таким образом, частотность этой формы складывается из всех энциклопедических статей Википедии, в которых так или иначе упоминаются реки, и всех контекстов, где этим словом обозначен некоторый промежуток времени, что особенно востребовано на интернет-страницах.

Особый интерес представляют формы, встречающиеся чаще в печатных текстах, чем в Башнете. К таким можно отнести, например, неизменяемую форму *ине*, употребляемую в прошедшем времени со значением усиленной неопределённости: *ҡайтты* ‘он (точно) вернулся’; *ҡайтҡан* ‘он (вроде бы, я не видел) вернулся’; *ҡайтҡан ине* ‘он, кажется, вроде, вернулся’. Для научных текстов такие конструкции нехарактерны, а в П (11 место) и в Ю (16 место), напротив, демонстрируют свою востребованность. По сравнению с этими показателями Башнет демонстрирует свою умеренную заинтересованность в форме *ине* (78 место, 18.393 употреблений), что объясняется малым количеством коммуникативно-ориентированных текстов.

Интересным показателем жанрового состава попавших в базу документов являются обсценные слова. Их наличие свидетельствовало бы о ситуации живого общения. Здесь мы наблюдаем довольно ущербную парадигму, в которой актуализированы лишь слова со значением ‘meretrix’ (*уйнаш*, *уйнашсы*: 21 и 105 употреблений соответственно), а остальные лексемы, связанные с именованием телесного низа и под., в Башнете не встречаются.

Среди слов, имеющих высокую частотность и в Рунете, и в Башнете, можно упомянуть *год / йыл* (3 место в списке существительных Рунета[5] / 50 место для начальной формы в Башнете и сравнительно высокие позиции у омонимичных форм типа *йылға* в общем списке, см. выше[6]), *новость / яңылыҡтар* (4 место в Рунете / 176 место в Башнете, 12.530 вхождений), *форум / форум* (7 место в Рунете / 25 место в Башнете), *поиск / эҙләү* (8 место в Рунете / 26 место в Башнете, 33.217 вхождений), *день / көн* (9 место в Рунете

---

[5] Данные взяты из [Контент].

[6] Цифры для сравнения весьма приблизительные, так как рейтинги для Рунета и Башнета составлены по разным принципам: в Рунете считаются леммы, а в Башнете словоформы; в Рунете каждая часть речи приводится в составе собственного рейтинга, а в Башнете учитывается место слово в общем (смешанном с точки зрения частеречной принадлежности) рейтинге. Тем не менее, есть слова, для которых разница в употребимости достаточно показательна, несмотря на принципы подсчёта.

/ 339 место, 6.137 вхождений), *пользователь / ҡулланыусы* (14 место в Рунете / 562 место в Башнете, 3.333 вхождений), *время / ваҡыт* (19 место в Рунете / 199 место в Башнете, 10.790 вхождений), *человек / кеше* (20 место в Рунете / 93 место в Башнете, 16.745 вхождений). В этом списке видны ключевые слова, отражающие, как можно представить, универсальные для сайтостроительства понятия, независимо от языка оккупирующие верхнюю часть списка. Не менее показателен список слов, которые характерны для Рунета, но редко встречаются в Башнете. Он демонстрирует, для чего, в отличие от Рунета, Башнет пока не предназначен: *телефон / телефон* (5 место в Рунете / 2.721 место в Башнете, 474 вхождения) *регистрация / регистрациялау* (16 место в Рунете / 107.110 место в Башнете, 4 вхождения), *комментарий / комментарий* (18 место в Рунете, 11.965 место в Башнете, 75 вхождений) *товар / тауар* (13 место в Рунете / 6.689 место в Башнете, 153 вхождения). Таким образом, Башнет не может быть рабочим инструментом для того, в чём давно преуспел Рунет: в нём нельзя искать телефоны, товары, оставлять комментарии, в нём мало сайтов с возможностью регистрации (а значит, привлекающих постоянную аудиторию пользователей).

В частеречной группе прилагательных особенно репрезентативна частотность адъектива *мобильный* (*мобиль* в башкирском). В Рунете это слово находится на 7 месте в перечне прилагательных, в Башнете его частотность равна 1 употреблению (в нашем списке оно попало на 233.607 позицию).

Итак, как можно заключить из частотного анализа лексики Башнета, его язык демонстрирует закономерное своеобразие по сравнению с языком печатных текстов. Своеобразие это продиктовано не столько специфическими реалиями сетевой коммуникации, сколько дизайнерскими решениями вебстраниц, соединёнными с сильным акцентом на официально-деловой стороне публикуемых документов.

Литература

1. *Контент* Рунета [Электронный ресурс] // URL: http://company.yandex.ru/researches/reports/ya_content_09.xml
2. *Кронгауз М. А.* Русский язык на грани нервного срыва. — М.: Языки славянских культур, 2007.
3. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка на материалах Национального корпуса русского языка. — М.: Издательский центр «Азбуковник», 2009.
4. *Росстат*. Демография [Электронный ресурс] // URL: http://www.gks.ru/wps/wcm/connect/rosstat/rosstatsite/main/population/demography/
5. *Сегалович И. В.* Методы сравнительного анализа современных поисковых систем и определения объема Рунета [Электронный ресурс] / И. В. Сегалович, Ю. Г. Зеленков, Д. О. Нагорнов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Восьмая

Всероссийская научная конференция. Суздаль, 17–19 октября 2006 года. — Суздаль, 2006. — URL: http://download.yandex.ru/company/paper_76_v1.pdf
6. *Сиразитдинов З. А.* Частотный словарь башкирского языка. Т.1 (наука). — Уфа: Гилем, 1997. — 330 с.
7. *Сиразитдинов З. А.* Частотный словарь башкирского языка. Т.2 <в выходных данных ошибочно: «Т.1»> (проза). — Уфа: Гилем, 2002. — 413 с.
8. *Тарасенко В. В.* Антропология Интернет: самоорганизация человека кликающего // Общественные науки и современность. 2000. № 5
9. *Татнет* // Википедия. [2011–2012]. Дата обновления: 15.11.2011. URL: http://ru.wikipedia.org/?oldid=39263025 (дата обращения: 1.2.2012).
10. *Частотный* словарь языка произведений Даута Юлтыя / Составитель З. А. Сиразитдинов / УНЦ РАН. АН РБ. — Уфа, 1995. — 292 с.
11. *Ethnologue* report for language code: bak [Электронный ресурс] // URL: http://www.ethnologue.com/show_language.asp?code=bak

References

12. *RuNet* content [Kontent Runeta], available at: http://company.yandex.ru/researches/reports/ya_content_09.xml
13. *Krongauz M. A.* Russian language at the edge of a nervous breakdown. M.: Languages of Slavonic Cultures [Russkiy yazyk na grani nervnogo sryva], 2007.
14. *Lyashevskaya O. N., Sharov S. A.* Frequency dictionary of modern Russian language based on the materials of the National Russian Language Corpus [Chastotny slovar sovremennogo russkogo yazyka na materialakh Natsionalnogo korpusa russkogo yazyka]. M., Azbukovnik publishing center, 2009.
15. *Russian* State Statistics Agency (Rosstat). Demography, available at: http://www.gks.ru/wps/wcm/connect/rosstat/rosstatsite/main/population/demography/
16. *Segalovich I. V.* (2006), Methods of comparative analysis of modern search engines and of determining the volume of RuNet [Metody sravnitelnogo analiza sovremennykh poiskovykh system I opredeleniya obyema Runeta] In Electronic libraries: prospective methods and technologies, electronic collections: The Eighth All-Russian Scientific Conference [Eletronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii: Vosmaya vserossiyskaya nauchnaya konferentsiya], available at: http://download.yandex.ru/company/paper_76_v1.pdf
17. *Sirazitdinov Z. A.* (2007), Frequency dictionary of the Bashkir language [Chastotnyj slovar' Bashkirskogo jazyka]. V.1 (science). Ufa, Gilem.
18. *Sirazitdinov Z. A.* (2002), Frequency dictionary of the Bashkir language [Chastotnyj slovar' Bashkirskogo jazyka]. V.2 <the imprint mistakenly puts it as V.1> (prose), Ufa, Gilem.
19. *Tarasenko V. V.* (2000), Anthropology of the Internet: self-organization of the homo clickens [Antropologiya Internet: samoorganizatsiya cheloveka klikayushchego] In Obshchestvennye nauki I sovremennost [Social Sciences and the Modern Times], # 15.

20. *Tatnet* // Wikipedia (2011–2012), last updated: 15.11.2011, available at: http://ru.wikipedia.org/?oldid=39263025
21. *Frequency* dictionary of the language of works by Daut Yultiy [Chastotny slovar proizvedeniy Dauta Yultiya] (1995)
22. *Ethnologue* report for language code: bak, available at: http://www.ethnologue.com/show_language.asp?code=bak