Башкирский интернет: лексика и прагматика в количественном аспекте

Борис Орехов, Азамат Галлямов

nevmenandr@gmail.com

Лаборатория компьютерной филологии
БашГУ, г. Уфа

Башкирский интернет = Башнет

- ***.ru → Рунет
- ... → Татнет
- ... → Башнет

Что такое Башнет?

- Башнет:
  - сайты о Башкирии на русском языке
  - сайты на башкирском языке

- в нашем понимании только второе (никто не включает в Рунет сайты о России на английском языке)

Что сделано?

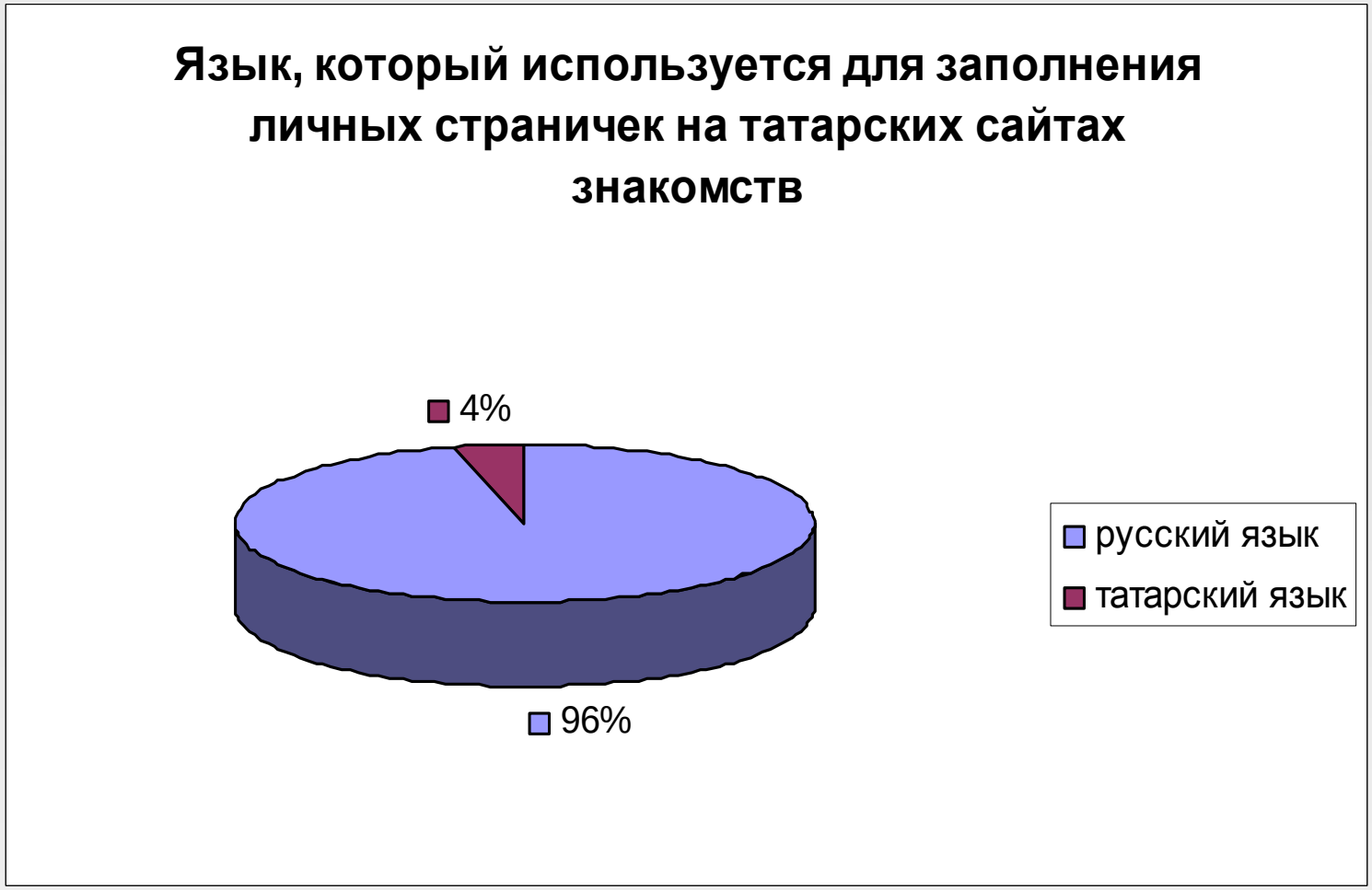
- До сих пор национальные интернеты почти не изучались:

  - Малое число пользователей
  - Отсутствие коммерческого интереса
  - Инертность национальной научной элиты

Исключение: Татнет (статья в Википедии)

Некоторые данные по Татнету

- Данные З. Махмутова (Казань)



Исследование Башнета

- Сплошной обход сайтов
- 30 доменных имён
- Парсинг строк, содержаших башкирские слова
- *ҥ*, *ҙ*, *ә*, *һ*, *ө*, *ү* , *ғ*, *ү*, *ҫ* (специфические элементы башкирской графики)

Объём Башнета

- На конец января 2012 г.:
- 66 199 страниц
- Рунет:
- Февраль 2004: 428571·10^3
- Декабрь 2005: 2538·10^6
- Осень 2009: 3825·10^6
- На текущий момент (грубая оценка): 5 млрд.

Национальные интернеты в других странах

- Объём Википедии ~ объём национального сегмента:
  - Каталанский: 370 279
  - Валлийский: 35 614
  - Бенгальский: 23 211
  - Шотландский: 9 792
  - Нижнесаксонский: 4 741
  - Клингонский: 71

Башнет: 15 964 Татнет: 15 259

Пользователи

«В республике только чуть более 30% населения постоянно пользуются интернетом. В России – около 40%» (Президент Р. Хамитов)

«Последними переписями населения (1979, 1989 и 1994 гг.) в городах Башкортостана зафиксирован устойчивый рост численности башкир (соответственно, 12.1 %, 14.5 %, и 14,9 % городского населения). По переписи 1989 г., 42,3 % всего башкирского населения республики проживает в городах, причем, по прогнозам демографов, в следующем столетии больше половины башкир будут городскими жителями». По данным переписи 2002 года доля городских башкир составляла 42,4 %, доля сельского населения – 57,6 %

Статистика по лексике

- Башнет: 27 252 251 слово
- Рунет (на 2009 г.): 2,3 трлн. слов

- Разница на 5 порядков (как и с количеством документов)

Медийные vs. немедийные сайты

- Медийные сайты (интернет-представление вторично):
  - 9 из 30 доменов
  - 41 298 страниц
  - 15 155 408 слов

Сайт газеты «Йәшлек»: 8 470 444 слова, 55.89% объёма словоупотреблений на медийных сайтах и 31% от всего корпуса Башнета, больше башкирской Википедии

Немедийные сайты

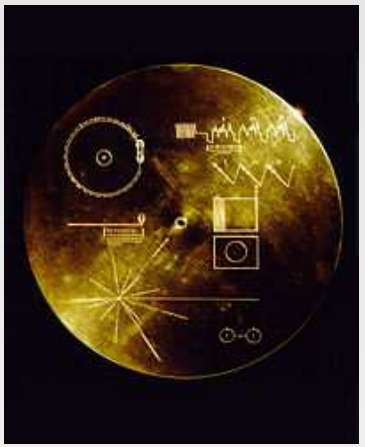
- Википедия
- Блоги (моноязычных башкирских нет)
- Отсутствуют востребованные сервисы (электронная почта, новостные ресурсы, социальные сети)
- Двуязычный форум (bashforum.net)

Количество слов на странице

На сайте газеты «Йәшлек» в среднем на страницу приходится 595,3 слова, а на сайте радиостанции «Ашҡаҙар», которая публикует, в основном, не тексты, а аудиозаписи передач, среднее количество слов на странице – 42,6.

Промежуточные выводы

- Башнет на половину – приложение к СМИ
- Имиджевый проект
- Не предполагает наличия читателя
- Сигнал во внешний мир (золотая пластинка ”Вояджера”)



Подробнее о лексике

- Сравнительные данные:

Сиразитдинов З. А. Частотный словарь башкирского языка. Т.1 (наука). Уфа, 1997; Сиразитдинов З.А. Частотный словарь башкирского языка. Т.2 <в выходных данных ошибочно: «Т.1»> (проза). Уфа:, 2002;. Частотный словарь языка произведений Даута Юлтыя / Составитель З. А. Сиразитдинов. Уфа, 1995.

Составлены на коллекции в 200.000 словоупотреблений

Контент Рунета (данные Яндекса)

Частотные слова

- Союз *һәм* ‘и’ и послелог *менән* ‘с’
- Послелог *буйынса* ‘поэтому, вследствие этого’ (78.230 вхождений, нет в других словарях в топ-50)

баш мөхәррирҙең дөйөм мәсьәләләр буйынса урынбаҫары ‘заместитель редактора **по** общим вопросам’

Частотные слова

башҡорт ‘башкирский’ (66.945 вхождений). В других словарях: ”научный стиль”: 15 место (1 126 вхождений); в публицистическом и словаре Юлтыя: не входит в первые 50 слов (в Ю: 68 место). Как и слово *башҡортостан* (17 место и 43 272 вхождений)

аҕасым ‘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)

Частотные слова

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» (31,28% всего объёма немедийного Башнета)

То же: *мәҡәлә* ‘статья’ (12 место 45.984 вхождений)

Частотные слова

үнәлештәр ‘направления’ (14 место, 44.049 вхождений, в других словарях отсутствует в топ-50)

төп йүнәлеш ‘основное направление’

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’.

Башнет и печатные тексты

мы не находим слов *буйынса*, *аҕасым*, *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.

Башнет и печатные тексты

чаще в печатных текстах, чем в Башнете: неизменяемая форму *ине* (употребляется 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) демонстрируют свою востребованность. Башнет демонстрирует свою умеренную заинтересованность в форме *ине* (78 место, 18.393 употреблений), что объясняется малым количеством коммуникативно-ориентированных текстов.

Башнет и Рунет

год / *йыл* (3 место в списке существительных Рунета / 50 место для начальной формы в Башнете и сравнительно высокие позиции у омонимичных форм типа *йылҕа* в общем списке

новость / *яңылыҡтар* (4 место в Рунете / 176 место в Башнете, 12.530 вхождений)

форум / *форум* (7 место в Рунете / 25 место в Башнете)

Башнет и Рунет

поиск / эҙләү (8 место в Рунете / 26 место в Башнете, 33.217 вхождений)

день / көн (9 место в Рунете / 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)

Башнет и Рунет

- То, для чего 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).

Башнет и фольклорный фонд

- ЭНИ ”Фольклорный архив БашГУ) lcph.bashedu.ru/index.php?go=editions

глагол *ти*, маркирующий передачу чужой речи, что вполне объяснимо для фольклорных произведений.

3 место: усилительная частица *ғына*, свойственная эмоционально насыщенным текстам.

Слова *бер* и *менән*, находящиеся и в верху списка частотности Башнета, находятся на 2 и 4 месте в частотном указателе ЭНИ.

Что дальше?

- Нужна морфология, чтобы не зависеть от графики
- Нужны подобные обследования других национальных интернетов