Институт русского языка РАН им. акад. В. В. Виноградова

Современный русский язык в интернете

Москва
Языки славянской культуры
2014

УДК 811.161.1'276
ББК 81.2Рус

Современный русский язык в интернете / ред. Я. Э. Ахапкина, Е. В. Рахилина. М.: Языки славянской культуры, 2014.

Печатается по решению ученого совета Института русского языка РАН им. акад. В. В. Виноградова.

Рецензенты: к. ф. н. И. Б. Левонтина, д. ф. н. Е. В. Маркасова, д. ф. н. Е. В. Ягунова.

Исследуя современную языковую ситуацию, необходимо учитывать речевую практику носителей языка, связанную с коммуникацией в интернет-пространстве. В этой книге собраны статьи специалистов, анализирующих интернет-тексты как новый языковой ресурс. Рассматриваются особенности структуры текста, лексика и синтаксис поискового запроса, коммуникативная направленность интернет-высказывания и влияние этого нового речевого пласта на современный русский язык. Одновременно анализируются изменения, которые касаются самого русского языка: возникающие грамматические формы, сдвиги в глагольном управлении, лексические неологизмы. Для них интернет выступает «камерой», быстро и точно отражающей мельчайшие новации. Книга адресована всем интересующимся русским языком, в том числе и профессионалам: преподавателям и лингвистам — грамматистам и лексикографам. Она будет полезна и разработчикам контента интернет-сайтов или редакторам порталов для профилактики нарушений цельности, связности, завершенности текста, и пользователям для решения задач по оптимизации поисковых запросов.

Издание осуществлено при финансовой поддержке Фонда содействия развитию интернета «Фонд поддержки интернет».

ISBN XXXXXXXXXXXXX

Б. В. Орехов, К. Ю. Решетников

К оценке Википедии как лингвистического источника: сравнительное исследование[1]

В статье затрагивается один из главных практических вопросов корпусной лингвистики — наличие / отсутствие лингвистически представительных оцифрованных текстов на том или ином языке. Очевидным «кандидатом» на роль релевантного лингвистического источника оказывается Википедия[2], в связи с чем проводится частотный анализ лексики русского раздела этого ресурса в сравнении с википедиями, написанными на некоторых других языках РФ.

Ключевые слова: компьютерная обработка текста, частотный анализ лексики, Википедия

Специалисты, занимающиеся компьютерной обработкой текстов на естественном языке, сталкиваются с необходимостью подбора материалов для создания своих корпусов. У этой задачи много специфических сложностей, связанных с дороговизной оцифровки текстов, ограничениями, которые накладывает система авторского права и др. Когда речь идет о больших письменных традициях, представители которых уже успели перевести в электронную форму часть своего наследия (как это, например, уже произошло с русской культурой, хорошо представленной в интернете), эти трудности не ощущаются в полной мере, потому что в открытом доступе в сети исследователь может найти большой массив текстов, пригодных для его задач. Википедия, на первый взгляд, представляет собой удобный ресурс для извлечения текстов, которые должны послужить материалом для лингвистического исследования. С одной стороны, это интернет-энциклопедия, основным принципом которой создатели провозгласили свободу распространения информации, поэтому тексты ее статей не охраняются авторским правом и с самого начала перешли во всеобщее пользование. С другой стороны, энциклопедия по своей сути содержит разностороннюю информацию и в идеале

[1] Проект выполнен при поддержке Центра исследований интернета и общества РЭШ и Лаборатории исследования социальных отношений и многообразия общества РЭШ.

[2] «Википедия» пишется в статье с прописной буквы в том случае, если речь идет об интернет-энциклопедии и соответствующем портале в целом. Строчная буква используется для обозначения языковых разделов

должна представлять тематически сбалансированный набор текстов. Сюда же можно добавить предусмотренную создателями техническую легкость, с которой все тексты Википедии можно извлечь и проанализировать. Особенно актуально наличие такого источника для тех языковых и письменных традиций, в случае которых оцифровка текстов пока является скорее проектом, нежели реализованной программой. К таким слабо оцифрованным языкам можно отнести все языки народов России, из-за чего компьютерная лингвистика почти не имеет возможности работать с этими языками.

Цель нашей работы — с помощью простейших лексико-статистических методов проверить пригодность текстов википедий на некоторых языках народов России (включая русский) для использования их в качестве лингвистического источника.

В нашем исследовании мы будем оперировать нестрогим понятием «естественности» частотного распределения словоформ в словаре, составленном на основе какого-то текста или коллекции текстов. Суть этого понятия в следующем. И из общих соображений, и из практики частотной лексикографии следует, что в текстах на языках флективного и агглютинативного строя наиболее частотными оказываются служебные части речи: союзы, предлоги, частицы, местоимения. Если корпус, на котором составлен частотный словарь, сбалансирован, т. е. включает в более-менее равном соотношении тексты разной тематики и коммуникативной направленности (и таким образом отражает соотношение, в котором естественный язык обычно фигурирует в жизни носителя), то полнозначная лексика появляется в таком словаре только в третьем-четвертом десятке (конкретные позиции будут зависеть от языка). Вот, например, первая десятка лемм из частотного словаря, составленного О. Н. Ляшевской и С. А. Шаровым на материале текстового фонда Национального корпуса русского языка [Ляшевская, Шаров 2009]:

Таблица 1. Первые десять лемм
из частотного словаря русского языка

№	Лемма	Часть речи	Частота (ipm)
1	и	conj	35801.8
2	в	pr	31374.2
3	не	part	18028.0
4	на	pr	15867.3
5	я	spro	12684.4
6	быть	v	12160.7
7	он	spro	11791.1
8	с	pr	11311.9
9	что	conj	8354.0
10	а	conj	8198.0

Важно отметить, что в этом списке словоформы уже приведены к единой лемме, и поэтому у изменяемых слов (к которым чаще всего и относятся полно-

значные, в отличие от служебных) шансы попасть в верхнюю часть списка гораздо выше. Однако даже с этим условием, как можно увидеть, наиболее частотными словами оказываются представители служебных частей речи.

Исходя из этого, наиболее «естественным» мы будем считать набор таких текстов, в которых частотное распределение слов наиболее близко к ожидаемому, т. е. демонстрирует сильные позиции для неизменяемых служебных слов и слабые позиции для форм полнозначных лексем. Соответственно, в случае, когда мы наблюдаем обратную ситуацию, это будет служить для нас основанием видеть в тексте дисбаланс. При этом, разумеется, дело не только и не столько в распределении словоформ. Мы исходим из гипотезы, что если диспропорция возникает в частотном словаре, составленном на основе какой-то коллекции текстов, то и данные по единицам других уровней языка могут быть (и даже скорее всего являются) не вполне корректными для произвольного текста на данном языке.

С этих позиций мы попробуем рассмотреть в сравнительном аспекте сначала википедию на русском языке, а затем несколько пар википедий на других языках народов России.

Верхняя часть частотного списка словоформ русской википедии (на момент исследования ее объем составляет 1 059 783 статьи) выглядит следующим образом:

Таблица 2. Первый десяток из частотного списка словоформ русской википедии

№	Словоформа	Встречаемость
1	в	10859129
2	и	5761105
3	на	3214393
4	с	2439469
5	года	1637221
6	по	1555831
7	году	1249646
8	из	1055953
9	был	940992
10	к	900353

Как легко удостовериться, в основном этот список соответствует аналогичным позициям частотного словаря, составленного на материалах сбалансированного корпуса, так как состоит в основном из предлогов, союзов и глагола «быть», используемого в качестве связки.

Однако 5 и 7 позиции интересны тем, что на них попадают формы полнозначного слова *год*. Это единственный случай нарушения естественного распределения словоформ в соответствии с критерием «полнозначные слова — неполнозначные слова» в верхней части частотного списка русской википедии. Он вызван, как нетрудно догадаться, жанровой спецификой энциклопе-

дического текста и структурой словника, в котором большое место занимают отдельные статьи, посвященные отдельным годам в мировой истории.

Второй десяток словоформ в частотном списке также составляют главным образом предлоги, местоимения и союзы, что вполне ожидаемо для частотного распределения в сбалансированном текстовом корпусе. Полнозначные слова в этом диапазоне не фиксируются.

Таблица 3. Второй десяток из частотного списка словоформ русской википедии

№	Словоформа	Встречаемость
11	не	843697
12	от	804136
13	а	754425
14	для	718569
15	что	676643
16	его	665984
17	до	637904
18	как	633286
19	он	611867
20	за	590437

Насколько мы можем судить, тексты русской википедии с поправкой на соответствующий жанр, если опираться на наши диагностические критерии, вполне могут быть использованы для компьютерных лингвистических исследований.

Однако для русского языка проблема отбора материала как раз не стоит остро. В свободном доступе в интернете представлены миллиарды страниц с текстами на русском языке, и при сравнительно небольших затратах лингвист может получить сверхбольшой корпус в десятки миллионов (а с применением некоторых инженерных решений и в десятки миллиардов) словоупотреблений. При этом тексты будут относиться к разным жанрам, отражать разную тематику.

В орбите русского языка и русской культуры, а конкретно — на территории Российской Федерации расположены ареалы обитания десятков народов, говорящих на собственных языках, которые располагают гораздо меньшим количеством оцифрованных текстов. Однако Википедия как престижный ресурс, претендующий на то, чтобы свободно распространять информацию на всех языках, имеет свои разделы и на многих языках народов России. Мы остановимся на нескольких примерах таких языковых разделов. В нашу выборку попали некоторые тюркские и финно-угорские языки и написанные на них википедии.

Из тюркских языков мы рассмотрим татарский и башкирский, а из финно-угорских — марийские, мордовские и два языка коми — зы-

рянский и пермяцкий. Такая структура выборки обусловлена тем, что нам представляется интересным проанализировать национальные википедии, составляющие друг с другом некие пары, внутри каждой из которых идет речь, с одной стороны, о максимально близком языковом родстве, с другой — о специфическом сходстве соответствующих энциклопедических ресурсов. Башкирская википедия очевидным образом близка к татарской, причем дело здесь не только в единстве башкирского и татарского языков как членов поволжско-кыпчакской общности (носящей, согласно разным трактовкам, либо генетический, либо ареальный характер), но и в том, что авторы этих википедий пользуются при создании контента схожими специфическими приемами. При детальном изучении татарского и башкирского разделов Википедии можно также отследить соревновательный момент. Аналогичные и даже еще более тесные википедийные пары составляют лугово-восточный марийский и горно-марийский, эрзя-мордовский и мокша-мордовский, а также коми-зырянский и коми-пермяцкий. В каждом из этих случаев мы имеем дело с двумя литературными языками, основанными на разных диалектах одного диалектного континуума, и с двумя родственными народами, которые, несмотря на тесную историческую связь друг с другом, имеют разные традиции и разное национальное самосознание. Сравнение соответствующих википедий дает не менее любопытные результаты, чем сравнение башкирского и татарского разделов.

Порядок, в котором будут рассмотрены верхние части частотных списков, соответствует иерархии википедий по количеству статей.

Самой большой по названному параметру национально-региональный раздел Википедии — это раздел на татарском языке, относящемся к тюркской семье. В настоящее время татарская википедия содержит 50 893 статьи. Вот верхняя часть частотного списка, составленного на основе этого ресурса:

Таблица 4. Первый десяток из частотного списка словоформ татарской википедии

№	Словоформа	Перевод/значение	Встречаемость
1	елга	“река”	132567
2	бассейны	“бассейн”	75706
3	су	“вода”	54689
4	буенча	“по”	48838
5	Русия	“Россия”	48722
6	урнашкан	“расположенный”	38043
7	км	“километр”	36962
8	һәм	“и”	27231
9	кече	“малый”	27203
10	дәүләт	“государство”	26888

Если не считать действительно высокочастотного соединительного союза и предлога со значением “по”, все слова, попавшие в этот перечень, относятся к категории полнозначных. Особенно интригует высокая встречаемость слова *елга*; в русском частотном словаре *река* находится на 916 месте.

Замечательным образом похож на этот список и частотный словарь башкирской википедии (30 724 статьи). Башкирский язык, как уже отмечалось, по отношению к татарскому является близкородственным (а республики, в которых эти языки являются титульными, расположены по соседству одна с другой на территории России).

Таблица 5. Первый десяток
из частотного списка словоформ башкирской википедии

№	Словоформа	Перевод/значение	Встречаемость
1	йылға	“река”	122849
2	бассейны	“бассейн”	85709
3	һыу	“вода”	64261
4	км	“километр”	38644
5	Рәсәй	“Россия”	33245
6	йылғаһы	“река”	30299
7	тиклем	“до”	28871
8	буйынса	“по”	25968
9	урынлашҡан	“расположенный”	23200
10	Дәүләт	“государство”	20786

Здесь только послелоги *тиклем* и *буйынса*, располагающиеся на 7 и 8 местах, могут претендовать на частотные позиции в естественном распределении словоформ. Зато полнозначные слова со значением “река”, “бассейн”, “вода”, “Россия”, “расположенный”, “государство” присутствуют среди самых частотных и в татарской, и в башкирской википедиях.

Такой дисбаланс в сторону определенных словоформ «водной» тематики объясняется способом пополнения татарской и башкирской википедий. Большая часть статей для разделов на этих языках не написана людьми, а сформирована автоматически из текстового шаблона на соответствующем языке, в который при программной обработке вставлены количественные данные. Абсолютное большинство такого рода статей посвящено рекам России, а данные о них взяты, по всей видимости, из Государственного водного реестра РФ.

Разумеется, совсем иным образом выглядит верхняя часть частотного словаря башкирского языка, составленного на корпусе научных текстов [Сиразитдинов 1997: 227].

Таблица 6. Первые десять лемм из частотного словаря башкирского языка

[номер лексемы в списке]	Һүҙ [слово]	F [частотность]
[1]	һәм	4936
[2]	бул	4787
[3]	менән	3003
[4]	бер	2425
[5]	ул	2336
[6]	был	2235
[7]	ит	1666
[8]	улар	1523
[9]	кил	1408
[10]	ти	1404

Можно обнаружить, что пересечений с башкирской википедией в первой десятке самых частотных слов не наблюдается.

Описанный способ автоматического наполнения статей Википедии на жаргоне активистов интернет-энциклопедии называется «ботозаливкой», т. е. совершаемой (ро)ботом «заливкой» новых текстов на сайт ресурса. На сайте Википедии есть специальная техническая страница, которая фиксирует статистику автоматического создания статей для 120 наиболее развитых википедий. Эта страница сообщает, что если для русской википедии доля статей-заготовок, созданных «ботами», сравнительно невелика и равняется 15 %, то для татарской википедии этот показатель составляет 73 %, а для башкирской 89 %. Статистика «ботозаливок», отмеченная в башкирском разделе, может показаться рекордной, однако это не так: в данном отношении башкирская википедия уступает еще как минимум восьми аналогичным ресурсам, гораздо активнее использующим роботизированное создание статей. Википедии на финно-угорских языках России в этой статистике не приводятся.

В самой обширной из рассматриваемых финно-угорских википедий, горно-марийской (5 110 статей), мы видим смешанный случай в плане «естественности» и «неестественности» наполнения верхней части частотного списка:

Таблица 7. Первый десяток из частотного списка словоформ википедии на горно-марийском языке

№	Словоформа	Перевод/значение	Встречаемость
1	Ин	“года” (род. падеж)	3694
2	дӓ	“и; а”	3351
3	эдем	“человек”	2606
4	ӹлен	“жил”	2173

№	Словоформа	Перевод/значение	Встречаемость
5	тӹштӹ	“там”	2053
6	гӹц	“из; от; через; по”	1441
7	доно	“с, при помощи”	1402
8	пырышы	“вошедший”	1347
9	йыхыш	“в род”	1263
10	ӹдӹрӓмӓш	“женский”	1152

С одной стороны, слова *дӓ, тӹштӹ, гӹц, доно* вполне могли бы оказаться на первых позициях в частотном словаре горно-марийского языка, составленном на материале сбалансированного корпуса. С другой стороны, в этом перечне присутствуют и слова, обладающие специфически высокой именно для горно-марийской википедии частотностью. К таким можно отнести форму род. п. слова со значением “год” (*ин*), слова со значением “человек” (*эдем*), “жить” (*ӹлен*), “женский” (*ӹдӹрӓмӓш*), причастие *пырышы* (“вошедший”). Присутствие в этом списке существительного *йыхыш*, означающего “в род” (т. е. слова со значением “род”, представленного в одной из падежных форм) имеет особое объяснение, которое мы дадим ниже. В верхней части извлеченных из Википедии и ранжированных по частотности лексических перечней на финно-угорских языках вообще часто встречаются слова со значением “род”, “категория”, “семья”.

В лугово-восточном марийском (3 814 статьи) мы видим приблизительно ту же картину.

Таблица 8. Первый десяток из частотного списка словоформ википедии на лугово-восточном марийском языке

№	Словоформа	Перевод/значение	Встречаемость
1	да	“и; а”	2808
2	марий	“мариец”	2270
3	дене	“с; от, из-за; по”	2243
4	ийыште	“в году”	1566
5	гыч	“из; от; через; по”	1559
6	тыгак	“так”	1360
7	ий	“год”	1205
8	лийын	“будучи; из-за”	1157
9	ончо	“смотри”	1092

Трудно ожидать от википедии на русском языке, что на второй позиции по частотности в ней окажется слово «русский» (и выше мы убедились, что в русской википедии ничего подобного не происходит). Зато форма от слова со значением “год” (*ийыште*) говорит о некотором лексическом единстве жанра русской и горно-марийской википедии. Однако сомнительно, чтобы *ончо* “смотри” (аналог русского «см.») попало бы на такую высокую позицию в какой-нибудь более развитой википедии. То,

что это слово, пусть и частотное, но не самое частотное, находится так высоко, можно считать одним из признаков неразвитости википедии: если бы контент был более равномерным и богатым, это слово ушло бы далеко вниз.

В коми-зырянском (3 971 статья), опять же, встречаются слова со значением “род, порода”, “принадлежащий к роду”, а также обязательное “год”.

Таблица 9. Первый десяток из частотного списка словоформ википедии на коми-зырянском языке

№	Словоформа	Перевод/значение	Встречаемость
1	да	“и; но; так как”	2781
2	коми	“коми”	1328
3	кыв	“язык, речь”	1325
4	км	“километр”	1058
5	тайӧ	“этот, это”	917
6	во	“год”	916
7	И	“и”	628
8	воын	“в году”	606
9	увтыр	“род, порода”	551
10	котырса	“принадлежащий к роду, к семье”	543

Частотный список словоформ коми-пермяцкой википедии (3 427 статей) также во многом схож с уже приведенными. Как и в других википедиях на языках народов России, в верхней части перечня доминируют служебные слова.

Таблица 10. Первый десяток из частотного списка словоформ википедии на коми-пермяцком языке

№	Словоформа	Перевод/значение	Встречаемость
1	да	“и”	2470
2	коми	“коми”	1380
3	котырись	“из рода, из семьи”	1243
4	вид	“вид”	1226
5	пантасьӧ	“встречается”	1130
6	район	“район”	1070
7	и	“и”	818
8	увтыр	“род, порода”	779
9	морт	“человек”	709
10	кыв	“язык, речь”	660

В эрзя-мордовский википедии (1 582 статьи) сходная ситуация. Здесь доминируют слова, входящие в состав ссылок на другие статьи или группы статей (категории) википедии («см.» и «также» из выражения «см. также»).

Таблица 11. Первый десяток из частотного списка словоформ википедии на эрзя-мордовском языке

№	Словоформа	Перевод/значение	Встречаемость
1	Истяжо	“также”	1470
2	категория	“категория”	1453
3	Вн	“смотри”, “см.” (ваномс “смотреть”)	1452
4	Чи	“день”	1121
5	Ды	“и; но”	898
6	Ие	“год”	770
7	Иенть	“года” (род. падеж)	742
8	покшчить	“праздники”	729
9	Те	“этот”	545
10	Ульнесь	“был”	502

В мокша-мордовской википедии (1 154 статьи) выборка выглядит наиболее экстравагантно. Хотя наверху списка находятся вполне законные для этой позиции служебные слова, но уже с третьей строчки начинаются странные для частотного перечня слова ботанической тематики.

Таблица 12. Первый десяток из частотного списка словоформ википедии на мокша-мордовском языке

№	Словоформа	Перевод/значение	Встречаемость
1	Ди	“и”	514
2	И	“и”	474
3	тьналста	“из семьи”	385
4	касыкссь	“растение”	376
5	Панчф	“цветок”	361
6	орхидея	“орхидея”	358
7	мокшень	“мокшанский”	328
8	васьфневихть	“встречается”	250
9	кизоня	“в году”	248
10	Сонь	“его, ее”	241

Разгадка проста: аналогично тому, как в вышеприведенных тюркских википедиях «ботозаливки» делаются с помощью данных Государственного водного реестра, в мордовских разделах автоматически создаются статьи о цветах. Едва ли не большая часть эрзянской и мокшанской википедий состоит из статей про растения, прежде всего про многолетние травянистые растения из семейства орхидных. Статьи ботанической тематики составляют преобладающий процент раздела на мордовских языках, и именно с этим связана высокая частотность слов со значением “род, семья, порода”, которые следует отнести к терминологии классификации видов. Мы предлагаем называть этот феномен «синдромом орхидеи».

Таким образом, мы обнаруживаем в википедиях на региональных языках России большое количество коротких статей-заготовок, созданных роботами и не наполненных текстами, которые были бы написаны людьми. В таких заготовках большую роль играют слова-ссылки (вроде «см. также...») и тематическая лексика из той области, к которой относятся статьи. Чтобы учесть этот аспект, мы посчитали среднюю длину статьи в словах для каждой из рассматриваемых википедий. Получились следующие данные:

Таблица 13. Средняя длина статьи для разных википедий

Википедия	**Средняя длина статьи в словах**
русская	184.47
татарская	58.47
башкирская	56.84
горно-марийская	36.74
лугово-марийская	47.43
коми-зырянская	29.58
коми-пермяцкая	35.45
эрзя-мордовская	34.46
мокша-мордовская	34.90

Из приведенных сведений видно, что развитая википедия с большой средней длиной статьи дает более качественные тексты, которые можно было бы использовать для лингвистических исследований.

Это один из возможных маркеров релевантности википедии, хотя релевантность в целом должна оцениваться по целому ряду признаков. Как следует из сказанного выше, одним из важных критериев является степень тематической сбалансированности. Если в том или ином национальном разделе Википедии более или менее пропорционально представлены разные тематические категории (подобно тому, как это имеет место в русской википедии), то релевантность такого ресурса можно оценить как высокую. Отсутствие больших статистических перекосов в употреблении лексики и относительное лексико-статистическое сходство с неэнциклопедическим корпусом демонстрируется адекватностью соответствующего частотного списка.

В свою очередь, вышеназванные качества того или иного национального раздела Википедии, обуславливающие степень его лингвистической релевантности, зависят от социального бэкграунда и условий создания этого раздела. Ключевую роль играют такие факторы, как большое количество активных пользователей и присутствие независимых активистов-аналитиков, которые могут выполнять редакторские функции, противодействуя как автоматическому порождению статей, так и необоснованному доминированию какой-либо одной тематики.

Национально-региональные википедии — по крайней мере, те из них, которые были рассмотрены здесь, — пока что создаются, по всей видимости, в основном небольшими сообществами энтузиастов, и потому в отношении этих википедий отсутствует практика массового редактирования и многостороннего взаимоконтроля. Однако вполне возможно, что со временем ситуация изменится, и усилия авторов, культивирующих национальные языки России в интернете, приведут к созданию более адекватных энциклопедических разделов, которые смогут служить достаточно релевантным лингвистическим источником.

Литература

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материале Национального корпуса русского языка). М., 2009.

Сиразитдинов З. А. Частотный словарь башкирского языка. Т. 1 (наука). Уфа, 1997.