[1]Е. А. Гришина, [2]Ю. Г. Зеленков, [3]Б. В. Орехов
[1]Институт русского языка им. В. В. Виноградова РАН, [2]Яндекс,
[3]Национальный исследовательский университет
«Высшая школа экономики»
[1,2,3](Россия, Москва)
[1]rudi2007@yandex.ru, [2]yuryz@yandex-team.ru,
[3]nevmenandr@yandex.ru

НАИВНАЯ ПОЭЗИЯ В АКЦЕНТОЛОГИЧЕСКОМ КОРПУСЕ

В статье рассматривается особенный материал, способный пополнить акцентологический корпус, специфический подкорпус в составе НКРЯ, отражающий место постановки ударений. Наивная поэзия – непрофессиональные стихи, созданные поэтами-любителями, их текстовая продукция не прошла редакторских фильтров и не была опубликована в авторитетных периодических изданиях и издательствах. Так как большинство этих текстов написаны в правильной силлабо-тонике, оказывается возможным автоматически предсказывать ударения и делать разметку для корпуса. Наивная поэзия была извлечена с сайта stihi.ru, самого старого сайта в России, публикующего подобного рода произведения поэтов-любителей. Несмотря на появление альтернативных площадок для публикации, сайт всё ещё остаётся популярным и количество публикаций на нём растёт. Для разметки текстов была применена специальная программа, предсказывающая расстановку ударений на основе машинного обучения. В тексте статьи приводится таблица, показывающая, насколько пополнение корпуса увеличило количество вхождений некоторых конкурирующих форм.

Ключевые слова: наивная поэзия, акцентология, корпусная лингвистика, лингвистические данные

Введенный в свое время фольклористами термин «наивная поэзия» (см. Неклюдов 2001; Минаева, Жигарина 2009) не имеет оценочной окраски. Речь идет о стихах, сочиняемых непрофессиональными литераторами. Определение непрофессиональной литературы наталкивается на известные методологические сложности, но в целом для нас это в первую очередь литература, не прошедшая редакторские фильтры, то есть не опубликованная в признанном сообществом периодическом издании или авторитетном издательстве [Bonch-Osmolovskaya, Orekhov 2014]). У такой текстовой продукции есть несколько содержательных и формальных особенностей, которые отличают ее от современной профессиональной поэзии. В частности, считается, что наивные поэты в большей степени ориентируются на воспроизведение усвоенных из школьной программы классических образцов, чем на создание принципиально новых произведений. Среди следствий этого принципа — доминирование силлабо-тоники в наивной поэзии. Действительно, так как традиционное представление о стихотворной речи предполагает, что поэтическое произведение должно быть написано одним из привычных двухсложных или трехсложных размеров, большинство непрофессиональных литераторов избегают неурегулированных размеров. Предварительные оценки, правда, свидетельствовали, что и современная авангардно-профессиональная поэзия также предпочитает метрическую организованность (хотя и в меньшей степени, чем непрофессиональные поэты) [Сонькин 2009].

Мы устраняемся от обсуждения художественных достоинств этих стихотворных произведений, оно должно быть делом литературных критиков. Для нас, как для составителей корпуса, наиболее важной является именно метрическая урегулированность большинства текстов наивных литераторов, позволяющая в автоматическом режиме предсказывать расстановку иктов в строках и, следовательно, ударений в словах, что представляется ценным при пополнении акцентологического подкорпуса НКРЯ. Другим ценным для составителя корпуса свойством наивной поэзии является ее доступность в последние годы. Непрофессиональные литераторы благодаря развитию Интернета получили широкие возможности для публикации своих произведений на специализированных сайтах, исповедующих принципы User Generated Content, то есть на таких, содержательное наполнение которых происходит за счет обычных пользователей, а

не нанятых администрацией ресурса редакторов или авторов. Одним из таких специализированных сайтов, на которых пользователь имеет простую возможность для публикации своего произведения, является сайт stihi.ru, — безусловно, старейший в Рунете ресурс, предоставляющий такие услуги. За время его работы пользователи stihi.ru успели объединиться в особую социокультурную среду (о некоторых принципах ее функционирования см. [Дианова 2009]). Даже с пришествием в сетевое пространство блогов и социальных сетей популярность stihi.ru не падает. На рис. 1 изображен график количества стихотворений, публикуемых в день на stihi.ru.

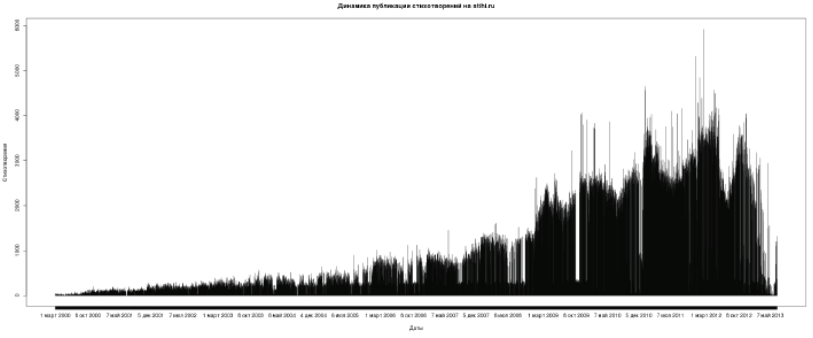

Рис. 1. Динамика публикации стихотворений на stihi.ru

В составе НКРЯ есть поэтический подкорпус, и размечаемые для него тексты автоматически попадают также и в акцентологический подкорпус, так как метрическая разметка предполагает расстановку ударений. Однако наивная поэзия не вписывается в концепцию пополнения поэтического подкорпуса, предполагающую включение культурно значимых и прошедших редакторские фильтры текстов (см. об этом статью К. М. Корчагина в настоящем сборнике). Тем не менее, наивные поэтические тексты представляют интерес для акцентологического корпуса как ценный и объемный материал для изучения актуального состояния русской акцентологии[1].

Локальная копия сайта stihi.ru по состоянию на декабрь 2013 года предоставлена разработчикам корпуса Поиском Mail.Ru, ко-

[1] По сообщению К. К. Богатырёва, А. А. Зализняк еще в доинтернетную эпоху говорил о пользе для акцентологии не принятых редакциями стихов.

торым мы приносим свою благодарность. Из общего объема более 300 млн словоупотреблений нами было отобрано два набора текстов общим объемом примерно 17 млн словоупотреблений. Первый набор текстов (приблизительно 10 млн словоупотреблений) был отобран случайным образом. В отборе второго (около 6,5 млн) учитывался процент совпадений биграммов текста с набором словесных биграммов, составленным по поэтическому корпусу НКРЯ (во второй набор текстов допускались только такие, у которых совпадение было не меньше 60%). Ожидалось, что высокий процент совпадения биграммов с поэтическим корпусом даст и большее число правильных силлабо-тонических текстов. Это соображение основывалось на приводившемся выше принципе, согласно которому наивная поэзия развивается в орбите существующих литературных стереотипов. БОльшая стереотипность в лексике и сочетаниях слов должна означать и бОльшую метрическую стереотипность.

На все отобранные тексты с помощью программы, разработанной Ю. Г. Зеленковым, была автоматически наложена метрическая разметка, то есть каждой строке присвоена помета, сообщающая, каким метром эта строка написана, а также учитывающая стопность и клаузулу.

Алгоритм автоматической расстановки ударений в использованной программе метрической разметки является открытым, т.е. обрабатывает любые русские словоформы бессловарным способом. Алгоритм имеет точность 0.96-0.97 для нарицательной лексики и 0.94-0.95 для имен собственных. В основе алгоритма лежит идея о существовании в русском языке высокой корреляции между буквенным составом концов слов и их грамматическими характеристиками, в частности, позицией ударения.

В качестве иллюстрации ниже, в таблице 1, приводится фрагмент большого (около 150 элементов) списка словоформ, упорядоченных в обратном лексикографическом порядке.

Таблица 1. Словоформы, упорядоченные в обратном лексикографическом порядке

*вольнослу`шательницами
свиде`тельницами
лжесвиде`тельницами

благоде`тельницами
избави`тельницами
прави`тельницами
…
состави`тельницами
вдохнови`тельницами
усынови`тельницами
покрови`тельницами
заяви`тельницами
победи`тельницами
…
руководи`тельницами
сопроводи`тельницами
роди`тельницами
распоряди`тельницами
жи`тельницами
скази`тельницами
…
проси`тельницами
искуси`тельницами
соврати`тельницами
посети`тельницами
похити`тельницами
укроти`тельницами
…
учи`тельницами
утеши`тельницами
прия`тельницами
настоя`тельницами
ва`фельницами
*отше`льницами

Из таблицы хорошо видно, что все без исключения словоформы (кроме самых крайних, которые показывают границы данного списка и сами входят в другие, аналогичные списки) имеют ударение в одной и той же позиции. В таком случае, наш список может быть сокращен всего до одной записи, т.е. примерно в 150 раз, как показано в таблице 2.

Таблица 2. Сокращенный вариант списка

*вольнослу`шательницами
свиде`тельницами
* отше`льницами

Другими словами, словоформа «свиде`тельницами» представляет целый класс словоформ, у которых правильная позиция ударения определяется на основе совпадения их конечных букв с конечными буквами представителя класса.

Этот подход был использован для машинного обучения, в результате которого, с помощью последовательных итераций, был построен достаточно большой список представителей «акцентологических классов», обеспечивающий вполне приемлемые показатели точности при расстановке ударения для любых русских слов.

Процесс анализа стихотворения включает три этапа. На первом этапе для каждой строки (стиха) выполняется расстановка ударений. Затем метрические ударения (икты) основных метров силлабо-тоники накладываются на словесные ударения и путем сравнения их позиций определяются наиболее вероятные метр, стопность и клаузула строки. При этом может распознаваться сверхсхемное ударение и корректироваться исходная расстановка ударения («и го`рьки мне`, горьки` твои` упрё`ки»).

На втором этапе анализируется структура строфы. Поскольку анализ отдельных строк может быть неоднозначным, число комбинаций, после объединения этих строк в строфы, часто оказывается больше одного. При снятии этой неоднозначности используется энтропия, вычисляемая для каждого варианта структуры строфы. В качестве основного выбирается вариант с минимальной энтропией.

Наконец, на третьем этапе, если структуры некоторых строф «выбиваются» из общей картины, используется метрическая инерция, т.е. делается попытка изменить метры отдельных строк на наиболее частотные в данном стихотворении. Здесь также возможна корректировка словесного ударения.

В заключение выполняется фонетическая транскрипция всех строф и определяются их схемы рифмовки.

Выяснилось, что для большей точности метрического анализа к текстам наивных поэтов требуются применять особые правила. В

частности, в случаях, которые программа могла бы трактовать двояко, при выборе между гетерометрией и нарушением стопности при сохранении метра, алгоритму следует предпочесть второй вариант: наивные авторы чаще не могут выдержать стопность, чем используют иной метр.

Все строки, для которых не был распознан силлабо-тонический размер, были помечены как «НУР» (неурегулированный размер). Стихотворения, в которых строк с НУР было больше 30%, были отброшены и в корпус в итоге не попали.

Благодаря автоматической метрической разметке была проверена гипотеза о корреляции между лексической и сочетаемостной традиционностью, с одной стороны, и метрической урегулированностью, с другой. В таблице 3 показано общее число строк и число строк в неурегулированных размерах в первом (отобранном случайно) и втором (отобранном неслучайно) наборах текстов.

Таблица 3. Наборы данных и неурегулированные размеры

Наборы данных	Строки	НУР
1	2 186 617	211 244
2	1 416 562	118 972

Была проведена статистическая проверка гипотезы с использованием критерия хи-квадрат и V Крамера: X-squared = 1644.823, df = 1, p-value < 2.2e-16, Cramer's V: 0.02136568.

Эти цифры свидетельствуют, что зависимость между наборами данных и «неурегулированными размерами» отсутствует.

Конечный результат составил пополнение акцентологического корпуса на 14,3 млн словоупотреблений (162,8 тыс. стихотворений).

Приведенная ниже таблица 4 позволяет оценить существенность пополнения акцентологического корпуса с точки зрения присутствия в текстовой коллекции некоторых «проблемных» случаев вариативности ударения.

Таблица 4. Вариативные словоформы в пополнении акцентологического корпуса

словоформа	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
балованный	0	0	0	0	0	1
баловать	0	6	7	13	2	16
балую	0	2	6	8	1	3
бедны	2	28	29	59	2	8
бледны	1	84	75	160	1	10
близки	6	133	79	218	25	137
бодры	0	13	7	20	0	2
важны	1	32	35	67	21	99
верны	5	54	59	118	3	57
видны	17	135	167	219	24	138
вкусны	1	14	4	19	0	4
влажны	2	11	4	17	0	11
вольно	6	319	17	342	22	103
вольны	2	44	23	69	3	28
воспринял	0	28	1	29	14	19
восприняла	0	0	0	0	0	1
восприняли	0	16	0	0	14	6
вредны	2	25	10	40	3	2
глупы	3	41	5	49	1	51
гнусны	1	14	0	15	0	0
годны	1	6	5	12	0	6
горды	2	63	47	112	4	17
горстей	0	0	3	3	0	2
горьки	3	44	6	53	0	37

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
грешно	4	2	123	129	1	44
грешны	0	11	10	21	2	20
грозны	4	59	5	68	0	3
грубы	3	81	15	99	1	15
грузны	0	3	4	7	0	2
грустны	2	24	15	41	0	35
грязны	2	8	3	13	0	3
дружны	1	38	16	55	1	13
душны	0	6	0	6	0	1
жадны	0	23	2	25	0	1
жирны	0	5	5	10	0	3
завит	0	2	8	10	1	0
залит	1	38	23	62	0	38
звучны	0	21	4	25	0	2
знатны	2	7	0	9	0	0
йогурт	0	0	0	0	0	6
квартал	2	2	64	68	6	12
кислы	0	3	0	3	0	1
колки	0	15	1	16	0	9
крепки	3	61	44	108	0	29
круглы	1	6	15	22	0	4
ловки	0	8	11	19	0	1
милы	2	206	31	239	2	68
мокры	2	9	10	21	0	8
мрачны	0	72	8	80	0	11
мудры	1	21	5	27	3	6
мутны	0	15	3	18	0	4
мягки	3	22	2	27	0	15
налит	1	22	11	34	0	3

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
нежны	3	138	47	188	0	84
низки	0	38	0	38	0	2
новы	2	142	13	157	0	17
нужны	13	194	901	1108	228	1209
областей	1	0	20	21	8	4
плотны	0	3	1	4	0	2
полно	60	745	289	1094	47	278
полны	13	391	415	819	2	303
предпринял	1	15	0	16	2	5
пресны	0	1	0	1	0	2
прожит	1	44	2	47	0	48
пройден	1	21	4	26	3	36
простыней	0	0	3	3	0	24
прочны	0	11	4	15	2	10
пышны	1	62	3	66	0	3
пьяны	5	76	31	112	4	19
редки	2	66	2	70	1	31
резвы	2	15	3	20	0	3
ровны	0	9	3	12	0	2
серьгам	0	0	3	3	0	0
серьгами	0	2	5	7	1	1
серьгах	1	7	9	17	0	2
сильны	10	76	73	159	10	64
скромны	1	16	10	27	0	9
скудны	2	18	1	21	0	5
скучны	1	38	16	55	1	9
слабо	10	207	24	241	25	77
слабы	4	90	16	110	2	33

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
сложны	0	3	7	10	2	17
слышны	12	181	63	256	3	168
смелы	0	45	11	56	0	9
смуглы	0	8	7	13	0	0
сочны	1	8	2	11	0	5
спелы	1	8	1	10	0	0
средам	0	4	1	5	0	7
страшны	9	154	87	250	2	233
строги	2	127	5	234	0	19
стройны	0	29	25	54	0	13
тверды	7	57	23	87	0	9
творог	2	8	16	26	6	7
теней	9	78	607	694	4	250
тесны	1	18	6	25	0	17
тихи	5	87	31	123	0	16
толсты	1	8	6	15	0	2
тонки	1	51	18	70	0	10
точны	0	14	6	20	2	7
трезвы	0	4	0	4	0	5
трудны	1	15	4	20	1	6
тусклы	0	21	4	25	0	2
тучны	0	22	3	25	0	0
узки	0	21	12	33	2	3
узко	1	41	0	42	5	5
хитро	3	60	59	122	8	59
хитры	0	9	11	20	2	6
храбры	4	22	5	31	0	1
черствы	1	1	4	6	0	3
честны	3	12	8	23	0	15

словоформа	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
чисты	6	91	57	154	1	86
чужды	5	216	29	250	2	104
шумны	0	35	2	37	0	1
шустры	0	0	1	1	0	1
щедры	0	13	10	23	4	9
щелей	1	14	28	43	2	18
ярки	1	90	7	98	0	16
ясны	4	158	61	223	1	45

Как видим, пополнение акцентологического корпуса в «проблемных» точках — по сравнению с предшествующим состоянием данных, относящихся к XXI веку, — достаточно существенно (ячейки, которые демонстрируют особенно заметную разницу, выделены полужирным). Таким образом, очевидно, что пополнение акцентологического корпуса наивными поэтическими текстами дает специалистам по современной русской акцентологии принципиально новый материал, который позволяет изучать состояние и тенденции развития современной русской акцентологической системы на статистически значимом материале, который, кроме того, в значительной степени отражает живое русское произношение, а не рекомендации нормативных справочников, — что позволяет, очевидно, как строить предположения о тенденциях развития системы, так и более полно описывать реализацию тенденций, заложенных в языке предшествующими этапами его развития[1].

[1] В дальнейшем возможности программы Ю.Г. Зеленкова предполагается несколько расширить за счет автоматической разметки в поэтических текстах т.н. «зоны рифмовки», т.е. тех слов и словосочетаний, которые включают в себя последнюю клаузулу строки. Это позволит исследовать проблемы, связанные, в частности, с выбором между *е* и *ё* в ряде лексем и словоформ (например, не вполне очевидно, какое в настоящий момент предпочитается произнесение — *крестный/крёстный*, *Крез/Крёз* и под.). И такие данные можно получить либо из прозаического модуля акцентологического корпуса (где анализируемые лексемы встречаются довольно редко), либо из его поэтического модуля. И здесь данные о зоне рифмовки в наивных поэтических текстах могут оказаться совершенно бесценными.

Литература

Гришина Е. А. Микроизменения в акцентологической системе русских прилагательных по материалам Национального корпуса русского языка // Вопросы культуры речи. Вып. 11. М., 2012

Дианова Т. Б. «Эффект Сидоренко»: образ виртуальной личности в сети и манипуляция массовым сознанием // Интернет и фольклор. М.: Государственный республиканский центр русского фольклора, 2009. С. 32–43.

До и после литературы: тексты наивной словесности. / Сост. А. П. Минаева. Отв. ред. Е. Е. Жигарина. М., 2009.

«Наивная литература»: исследования и тексты / Сост. С. Ю. Неклюдов. М., 2001. [URL: http://www.ruthenia.ru/folklore/luriem43.pdf]

Савчук С. О. Активные процессы в системе русского словоизменения: опыт корпусного исследования акцентологических норм // Труды II Международной конференции «Русский язык и литература в международном образовательном пространстве: современное состояние и перспективы». Т. 2. Гранада, 2010. С. 1549–1554.

Сонькин В. Традиционный стих в контексте современной русской поэзии // Славянский стих. VIII: Стих, язык, смысл. М: Языки славянских культур, 2009. С. 390–393.

Bonch-Osmolovskaya A., Orekhov B. Distant reading of naïve poetry: corpora comparison as research methodology // Digital humanities Lausanne — Switzerland '14. URL: http://dharchive.org/paper/DH2014/Paper-777.xml

Larsson J. О некоторых изменениях тенденций акцентного развития прилагательных в современном русском языке // Russian Linguistics. 2006. Vol. 30. P. 235–262

[1]E. A. Grishina, [2]Yu. G. Zelenkov, [3]B. V. Orekhov
[1]Vinogradov Russian Language Institute
of the Russian Academy of Sciences,
[2]Yandex, [3]National Research University "Higher School of Economics"
[1,2,3](Russia, Moscow)
[1]rudi2007@yandex.ru, [2]yuryz@yandex-team.ru,
[3]nevmenandr@yandex.ru

NAÏVE POETRY IN ACCENTOLOGIC CORPUS

The article deals with specific material that is able to enlarge the Accentologic Corpus, the subcorpus in the RNC reflecting accent patterns in Russian words. Naïve poetry is the term for unprofessional poems written by amateur poets. Their textual products have not passed any editorial filters and have not been published in reputable periodicals and publishing houses. Since majority of these texts are written in the correct syllabic-tonic, it is possible to predict stress automatically and make the markup for the Corpus. Examples of naive poetry have been downloaded from the site stihi.ru, the oldest one in Russia that publishes such works by amateur poets. Despite the existence of alternative platforms for publication, the site is still popular and the number of publications is on the rise. A special program for the markup of texts was used. This program based on machine learning predicts the place of the accents. There is a table in the article that shows how the enlargement has increased the number of occurrences of some competing forms in Corpus.

Key words: naïve poetry, accentology, corpus linguistics, linguistic data.

References

Grishina E. A. [Micromutations in accentologic system of Russian adjectives based on Russian National Corpus]. *Voprosy kul'tury rechi* [Questions of speech culture]. Issue 11. Moscow, 2012. (In Russ.)

Dianova T. B. [“Sidorenko Effect”: an image of virtual personality on the web and mass conscience manipulation]. *Internet i fol'klor* [Internet and Folklore]. Moscow, Gosudarstvennyi respublikanskii tsentr russkogo fol'klora Publ., 2009, pp. 32–43. (In Russ.)

Do i posle literatury: teksty naivnoj slovesnosti [Before and after a

literature: naïve texts]. A. P. Minaeva (comp.), E.E. Zhigarina (ed.). Moscow, 2009.

«Naivnaja literatura»: issledovanija i teksty [Naïve literature: studies and texts]. S. Yu. Nekljudov (comp.). Moscow, 2001. Available at URL: http://www.ruthenia.ru/folklore/luriem43.pdf (accessed on 14.06.2015)

Savchuk S. O. [Active processes in the system of Russian inflections: a corpus-based study of accentologic norms] *Trudy II Mezhdunarodnoi konferentsii "Russkii yazyk i literatura v mezhdunarodnom obrazovatel'nom prostranstve: sovremennoe sostoyanie i perspektivy"* [Proceedings of the IInd International conference "Russian language and literature in the international educational space: current status and prospects"]. B. 2. Granada, 2010, pp. 1549-1554. (In Russ.)

Son'kin V. [Traditional verse in a context of contemporary Russian poetry]. *Slavjanskij stih. VIII: Stikh, jazyk, smysl* [Slavic verse VIII. Verse, language, meaning]. Moscow, Jazyki slavyanskikh kul'tur Publ., 2009, pp. 390–393. (In Russ.)

Bonch-Osmolovskaya A., Orekhov B. Distant reading of naïve poetry: corpora comparison as research methodology. *Digital humanities* Lausanne — Switzerland '14. Available at URL: http://dharchive.org/paper/DH2014/Paper-777.xml (accessed on 14.06.2015)

Larsson J. [About some mutations of tendencies of accentologic evolution of adjectives in contemporary Russian language]. *Russian Linguistics*, 2006, vol. 30, pp. 235–262.