# Corpus regional lexicography: principles, methods, and preliminary results

**Belikov V. I.**
MIPT, ABBYY Lab
vibelikov@gmail.com

**Dubyaga A. O.**
RSUH
dubiaga.al@gmail.com

**Rvanova L. Y.**
MIPT, ABBYY Lab
rvanova.lyu@phystech.edu

**Selegey V. P.**
ABBYY
vladimir_s@abbyy.ru

**Abstract**

The article summarizes the results of the long-term project “Languages of Russian Cities” (LoRC) of the regional vocabulary collecting and researching, which, unfortunately, was not depicted in any academic publications for a number of reasons. About 4 thousand pieces of regional materials were collected, systematized, and became the basis of the typology of regional differences consideration and the concept of a regional norm discussion. Reliability issues and methods of computer-based regional corpus research, including automatic text classification and author profiling, are paid attention to. Along with this article, the "reincarnation" of the LoRC project is also returning to the fund of open lexicographic resources basing on the joint portal for distinctive sociolinguistic research, which includes the General Web-corpus of Russian Language and the interactive dictionary “Languages of Cities and People” (LoC&P)

**Keywords:** WAC; regional diversity; regionalism; regionally biased vocabulary; regional norm; automatic regional classification; author profiling

**DOI:** 10.28995/2075-7182-2021-20-79-93

# Корпусная региональная лексикография: принципы, методы и предварительные результаты

**Беликов В. И.**
МФТИ, ABBYY Lab
vibelikov@gmail.com

**Дубяга А. О.**
РГГУ
dubiaga.al@gmail.com

**Рванова Л. Ю.**
МФТИ, ABBYY Lab
rvanova.lyu@phystech.edu

**Селегей В. П.**
ABBYY
vladimir_s@abbyy.ru

**Аннотация**

В статье подводятся итоги многолетнего проекта «Языки Русских Городов» (ЯРГ) по сбору и исследованию региональной лексики, который, к сожалению, не был «финализирован» по ряду причин в виде академических публикаций. Был собран и систематизирован значительный (ок. 4 тыс. единиц) региональный материал, на базе которого рассматривается типология региональных различий, вводится/обсуждается понятие региональной нормы. Особое внимание уделяется вопросам надежности и методикам компьютерных региональных корпусных исследований, включая автоматическую классификацию текстов и профилирование авторов. Вместе с этой публикацией возвращается в фонд открытых лексикографических ресурсов и «реинкарнация» проекта ЯРГ – теперь на базе объединенного портала для дифференциальных социолингвистических исследований, включающего интернет-корпус ГИКРЯ и интерактивный словарь ЯГеЛь (Языки Городов и Людей).

**Ключевые слова:** WAC; региональная вариативность; регионализм; регионально смещенная лексика; региональная норма; автоматическая региональная классификация; авторское профилирование

## 1 Проект Языки Русских Городов: мотивы и итоги

Можно утверждать, что до старта проекта «Языки русских городов» реального представления о масштабах региональной вариативности в норме языка (см. далее о региональных нормах) не было. Обычная публика охотно принимала анекдоты в популярных изданиях про куру, гречу и поребрик, но за пределами московско-питерской темы существовали только отдельные плоды лексического краеведения, в которых не всегда проводилась грань между сельскими диалектами, топонимикой, автохтонной лексикой и собственно региональной вариативностью «городской» языковой нормы.

Можно сказать, что сама постановка исследовательского вопроса со смещением интереса с устного языка на письменный не была очевидной. В проекте впервые ставилась задача определить масштаб и типологию регионального варьирования РЯ на основании исследования текстов региональных СМИ и социальных сетей и опросов пользователей форумов Lingvo.

Реализация проекта «внутри» Lingvo community оказалось очень правильной идеей: в начале нулевых языковые форумы Lingvo были наиболее популярной дискуссионной площадкой по вопросам не только перевода, но и русского языка, с большим региональным разбросом участников, многие из которых были профессиональными переводчиками и/или филологами.

Технология была простой: участники обсуждений предлагали свои варианты, которые проверялись профессиональными редакторами с помощью базы региональных СМИ Интегрум и работавших на тот момент Яндекс.Блогов.

В результате нескольких лет функционирования форума был собран уникальный материал – несколько тысяч региональных слов (в разной стадии проверки) с высоким индексом цитирования в региональных СМИ и соцсетях, что позволило сделать вывод о глобальном характере проявления языковой вариативности в норме языка. Некоторым «апофеозом» этой деятельности стало издание словаря «Языки Русских Городов», который вошел в юбилейную версию системы Lingvo с номером «ХЗ». В этот словарь вошло около тысячи слов из числа обсуждавшихся на форуме. Форум послужил источником материала для значительного количества научных статей, там же разрабатывалась методика лексикографической работы с материалами интернета в целом [1; 5; 6; 7; 8]. Из работы на форуме вырос фундаментальный словарь неофициальной топонимии России и ближнего зарубежья [2].

К сожалению, проект ЯРГ прекратил активное существование из-за смены лексикографической политики команды Lingvo, которая отказалась от собственных лексикографических проектов и перешла на лицензирование контента. Но причина задержки с публикацией промежуточных итогов более глобальна — она отражает общее падение интереса к кропотливой лексикографической работе в угоду автоматизации, стремление к максимальному покрытию и скорости (дешевизне) получения результата в ущерб качеству. К сожалению, сегодня исчезает понятие «авторитетного» словаря даже для академических толковых словарей (хотя и раньше их авторитетность в случае региональной лексики не означала корректности, полноты и последовательности в описании).

Авторы надеются, что эту тенденцию не поздно еще изменить, причем не возвратом к старому, а за счет применения методов анализа больших корпусных данных, которые являются не только средством получения новых объектов описания, но и верификации этих описаний.

## 2 Типология региональной лексики

Анализ собранного материала позволяет сделать некоторые выводы. Регионализмы можно классифицировать по разным основаниям.

По происхождению:

- Из местных сельских диалектов (что не всегда легко подтверждается в силу малодоступности диалектных словарей; по СРНГ [16] часто не удается выявить диалектный ареал). Есть и экзотика. *Баской/баский* ‘красивый, хороший’ очень слабо представлено в городском узусе на Европейском севере и Сев. Урале, в Сибири — только в сельских диалектах. Но в 1990-х — начале 2000-х в молодежном жаргоне от Норильска до сев.-вост. Казахстана было (сейчас, вероятно, ушедшее) прил. *бащный* — ‘отличный’ — вероятно, от сравнительной степени баще ‘лучше’.

- Из распространенного в ареале нерусского языка, заимствование (*махалля* < узб. *mahalla*) или калька (*самориск* < латышск. *pašrisks*). А также расширение значения: *урюк* ‘дерево и его свежие плоды’ (среднеазиатский *урюк* производит впечатление метонимии, но это результат контактов с местными тюркскими языками).
- Заимствовано из местного неродственного языка, но в городскую речь проникало и из диалектов, и из самого языка — *калега* ‘брюква’ в Удмуртии.
- Свободно порождается системой языка (*башня/свечка* — одноподъездный «высокий» дом); *политсила* ‘организация, активно участвующая в политической жизни страны; партия, политический блок’ (Украина).
- Заимствовано из проф. узуса *точечный дом* (а этап *точечный дом* → *точка* — уже переработка системой языка).
- Неясно откуда (*мультифора* ‘файл для бумаг’, центральная Сибирь) — внутрення форма очевидна: излат. *multus* ‘многочисленный’ и *foro* ‘дырявить’, но пути появления этой номинации не ясны.

По фиксации в толковых словарях:

- Отсутствуют.
- Присутствуют параллельно с основным с неточным толкованием. Ср. в БТС [10] общее *сурок* (‘небольшое животное сем. беличьих, зимой впадающее в спячку’) и его синонимичные для незоолога именования *байбак* ‘степной грызун из рода сурков, осень и зиму проводящий в спячке’ и ‘грызун рода сурков, обитающий в Забайкалье, на Алтае, в Монголии и Северном Китае’.
- Присутствуют с косвенным указанием на региональность, ср. в [10]: «1. На Дальнем Востоке и в Сибири: небольшая гора с округлой вершиной, курган, холм. 2. На Камчатке и Курильских островах: вулкан».
- Присутствует с ошибочной стилистической пометой (обычно это слова со слабо выраженной региональностью). Ср. в [10]: «**ХОЛОДЕЦ**, -дца*; **м**. Нар.-разг.* 1. =Студень» (*студень* помет не имеет; в современном петербургском узусе, исключая самое старшее поколение, *холодец* частотнее *студня*).
- Присутствует, но региональность не маркируется. Ср. в [10] *банлон* (без помет), *водогрей* (без помет), *хабарик* (*жарг.*). При этом первое с такой фонетикой давно устарело, ср. реакцию петербургских блоггеров при обсуждении вариативности: *Слово «бонлон» я слышала разве что от папы* (1973 г. р.). *Нам банлон не нужен. Мы хотим бадлон!* (1978 г. р.).
- Некогда фиксировалось словарями в разных значениях, в том числе и региональном, а в современных словарях значение заужено: *толока* ‘безвозмездная общественная работа’ (Украина, Белоруссия, Эстония, Латвия) — значение, среди прочих, фиксировалось в словаре Ушакова [18]. В БТС [10] только явно устаревшее ‘поле под паром, используемое для выпаса скота; выпас скота на таком поле с целью удобрения почвы’, которое плохо стыкуется и с региональным, и с распространяемым Яндекс.Толокой: «Заработок в интернете. Простые задания за вознаграждения» (toloka.yandex.ru).

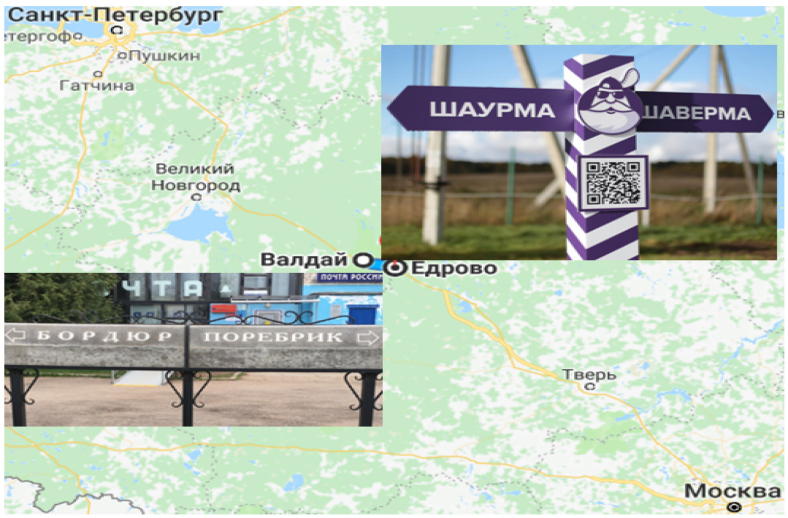
По связи с общенормативным:

- Синоним: *качок* ‘насос’, *лейка* ‘воронка’ (они же и омонимы).
- Как бы синоним: петерб. *латка* = «почти ‘утятница’», но *латка* может быть круглой.
- Как бы «метонимия»: *гардина* ‘карниз для занавесок’, *гамаши* ‘рейтузы’, ср.-аз. *урюк* ‘дерево и его свежие плоды’ (такой *урюк* производит впечатление метонимии, но это результат контактов с местными тюркскими языками)
- Омонимия: дальневосточн. *медведка* (в стандарте — живущее в почве насекомое *Gryllotalpa gryllotalpa*) и *чилим* (в стандарте — *Trapa natans*, водяной орех) — крупные креветки.

По осознаваемости носителями:

- Специфика *массово* не осознается.
- Специфика общеизвестна и в ареале, и за его пределами.

В силу слабости межрегиональных контактов традиционно таких слов было мало, вне ареала слово становилось узнаваемым через беллетристику (*изба/хата*) или из фактов личной биографии (*цветет урюк* у Солженицына; то же в «Золотом теленке» — *Цветет урюк под грохот дней / Дрожит зарей кишлак* … вне региона, вероятно, массово воспринималось как «плохие стихи»).
Интернет породил «мемические» знания, часто неточные: указатель «Шаурма» в Едрово — в сторону Тверь→Москва, но в Твери *шава́рма*):




- Специфика общеизвестна в ареале, служит своеобразным шибболетом. Популяризации такого знания в пределах ареала способствовали распространенные в 2000-х годах бродячие списки *Ты из* [название города], *если* [следует длинный список местных реалий и элементов языковой специфики], а также публикации типа «Только у нас известны слова …»; фактически ареал может быть заметно шире «заявленного» (*грядушка* [спинка] *кровати* — якобы чисто воронежское, но шире используется в Волгоградской обл.) или ареал может быть не единственным (*отсветка* ‘ксерокопия’ — якобы только Амурская обл., но также и Литва). **Вырожденный случай** — поморские *баско* и *порато* в Архангельске (повседневно слова не используются, имеют лишь символическое значение):


Архангельск: парикмахерская и магазин головных уборов

По границам ареала:

- Точечное: *явочка* ‘талон к врачу’ (Ярославль, но не Ярославская область)
- Ареал большой; граница может быть четкой и размытой (образуется периферия ареала): (*по*)*ставить укол/прививку*: западная граница четкая, восточной как бы и нет, но на Дальнем Востоке и в Казахстане *делать укол* сильно частотнее, чем *ставить*.

- Ареал сравнительно небольшой (одна-две-три области или сопредельные части областей), граница четкая: *на зеленую / на зеленой* — о выезде «на природу» с рекреационными целями (Ярославская, Костромская, северо-восток Ивановской обл.). Госграница — не помеха *фыгать* ‘курить анашу’ (только Оренбургская и Актюбинская обл.), граница ареала четкая.
- Территориально далекие независимые ареалы (во всяком случае связь не прослеживается): *отсветка* ‘ксерокопия’: Литва и Амурская обл., *лента* ‘сдвоенный академический час в вузе, «пара»’: Днепропетровск (ныне *Днепр*) и частично соседи, а также Красноярск;
- Точные границы ареала трудно определить. *баллон* ‘трехлитровая банка’ — «южное»: от Украины до Средней Азии, но детали ареала (ареалов) неясны. Примерно такова же ареальность *газводы*, но здесь явно в соответствии с надписями на автоматах *Газвода* vs. *Газированная вода*, — где производились и куда поставлялись те и другие сейчас вряд ли можно установить.
- По географии нас интересует постсоветское пространство. Но специфику языка дальнего зарубежья тоже необходимо учитывать, если она явлена в беллетристике или публицистике, широко читаемой в России (в частности, в российских изданиях). В беллетристике Журнального зала неоднократно встречаются загадочные для многих слова *хенди/хэнди* (Германия), *пелефон* (Израиль) ‘мобильник’, которые не всегда ясны из контекста[1]. То же касается и необычного управления, ср. *на интернете* вм. *в интернете* (характерно для США и, кажется, Франции).

По типу «пользователя»:

- Общее разной стилистики. Есть заведомо **нормативное**, использующееся в местном законодательстве: *углярка* ‘угольный сарай’ (Кемерово), судебной практике: *отбивать* (кассовый) *чек* (сев. Урал)[2]; *хабарик* ‘окурок’ (СПб, разговорное), *чибон* ‘окурок’ (Пермь, сленг = «общий жаргон», стилистически гораздо ниже петербургского хабарика).
- Детское. Тут сложно говорить о региональности, поскольку детская лексика плохо фиксируется словарями, то есть «общая кодифицированная норма» в этой сфере неполна (нет, например, явно общих *бибики* и *бибикать*, при том что слово не сугубо детское: есть цитата из СМИ, как Путин, ехавший на Афон, *бибикал ослу*, взрослые собственный автомобиль нередко зовут *бибикой*).
- Жаргон. В первую очередь речь о молодежном жаргоне, хотя понятие «молодежный» довольно туманно. Если потребность в конкретной номинации с возрастом не затухает, то «бывшая молодежь» продолжает его использовать, переходя в средний возраст. Жаргонное, естественно, может быть необщеизвестным в ареале: *зимбура* (Мурманск) «знают автомобилисты и бомжи».
- Профессиональное: *туалет прямого падения* — используется в системе архангельского ЖКХ, поэтому знают его все местные жители; *опанелка* (дверная обналичка, пиломатериал) — непрофессионалы знают, только если сталкивались (как и «московскую» *обналичку*).
- Сугубо административное, вряд ли встречающееся в повседневном узусе: *освободить от транспорта* ‘закрыть проезд на определенных улицах’ (Мурманск). Типы поселений при

[1] И так мы ходили вчера и позавчера, пели “Христос воскресе из мертвых” под писк хэнди, окутанные густым запахом духов, который не может выветрить из нашего крестного хода никакой ветер (Михаил Шишкин. Взятие Измаила // Знамя, № 12, 1999)

[2] По данным базы СМИ «Интегрум» в документах Федерального арбитражного суда Московского округа **Oper** от (*кассовый*) *чек* представлен глаголами *пробивать* (94%) и *выбивать* (6%), ясно, что норма здесь *пробивать*, но допустимо и *выбивать*; tertium non datur. Похожая ситуация и в документах ФАС Северо-западного округа, соответственно, 90% и 10%. Совершенно иное положение в ФАС Уральского округа, где в этом контексте преобладает *отбивать* — 54% (*пробивать* — 35%, *выбивать* — 11%). Преобладание возникает в основном за счет документов северной части Уральского округа — в нижестоящем 17 арбитражном апелляционном суде (Удмуртия, Пермский край, Свердловская область) *отбивать* дает 67% таких контекстов, а на юге, в 18 арбитр. апелл. суде (Башкирия, Челябинская, Оренбургская, Курганская обл.), только 40%. Разумеется, эта практика не кодифицирована ни в толковых словарях, ни в законодательных документах. Основывается она на региональном узусе, ядро ареала которого на севере, а юг Урала — периферия. ***Отбивание** кассовых чеков* противоречит московской и петербургской норме, но **для Урала это норма**, de facto закрепляемая судебными документами.

железнодорожных разъездах официально именуются по-разному. Судя по территориальному распределению, определение типа поселения зависело от региональной администрации (наименование типа поселения приводятся по ОКТМО). есть, *разъезд* и *железнодорожный разъезд* — это разные типы поселений. В одних областях представлены только *разъезды* (в Вологодской обл. их 12, в Пензенской — 8, в Ростовской — 10, в Ульяновской — 19), в других — только *железнодорожные разъезды* (в Волгоградской — 10, в Кировской — 10, в Саратовской —18). В Мордовии есть только более редкий вариант *поселок разъезд*, их 16. В Свердловской области только два поселения-разъезда, но именуются они сложно: *поселок при железнодорожном разъезде 99 км* и *поселок при железнодорожном разъезде 136 км*.

Распространенность по времени:

- Пришло-ушло (обычно жаргонное, но к уходу может подталкивать не только смена моды на экспрессивные слова, но и утрата актуальности: *сотыга* ‘100 рублей’, сибирское).
- Историзм: *мосовская машина* (от советских автомобильных номеров с буквами *МОС* (с 1959 г.) для «больших начальников») — есть литературные примеры, где контекст ничего не разъясняет.
- Утрачивающаяся номинация. Известное с 1970-х гг. петерб. *лабаз* ‘магазин, обычно винный’, судя по возрасту использующих слово, устаревает; при этом в более младших возрастах значение расширяется: ‘любой магазин’.

По степени опознаваемости неносителем:

- от «**неправильно**» (*сайка чёрного* ‘буханка’, Волгоград) и «понятно, но **неграмотно**» (*ставить укол*, *отбивать чек*) через «ясно, что такое» (*красноголовик* ‘подосиновик’) и «в контексте **несложно догадаться**» (*пастик* ‘стержень шариковой ручки’, *химица*, *математица* и т. п. ‘учительница химии’ и т. п.) до **ошибочного понимания** (*медведки к пиву*, *бегал на морозе в одном гольфе* (=водолазке) и **непонимания** (*поместить рисунок в мультифору*, *мосовская машина*).
- Стандартные для Удмуртии примеры типа *Мой отец — удмурт, а мама — русская, из потомственных дворян, и я очень жалею, что не знаю удмуртского языка ‹…› Приедешь, бывало, на ферму, колхозники толкуют о своих проблемах, а ты хоть **не** толкай соседа: «О чем речь?»* (Председатель Союза журналистов Удмуртии Людмила Прокошева, в интервью) в печатном виде, вероятно, будут поняты правильно, но получат реакцию вроде «при редактировании текста забыли убрать ***не***».

## 3 Общеязыковая и региональная норма

Теоретические взгляды на то, что следует считать литературной нормой, изменчивы, меняется языковая ситуация, положение русского языка в мире, эволюционирует и сам язык. Поэтому представляется естественным регулярно возвращаться к тому, что́ именно в языке подлежит нормированию, какие объективные и субъективные препятствия встают на этом пути, насколько строгими могут и должны быть нормы в отношении разных аспектов такого сложного и во многом все еще непознанного феномена, каким является язык.

Мы исходим из концепции В. А. Ицковича, понимавшего под нормой «комплекс закрепленных речевой практикой языковых средств и закономерностей их реализации, объективно существующие в данное время в данном языковом коллективе» [12: 8]. По Ицковичу, норма представлена двумя ипостасями: «Имплицитно норма выступает в виде образца или, точнее, текстов, считаемых образцовыми ‹…› Эксплицитно, в явном виде, сформулированной, норма предстает перед носителями языка в кодификации, отражающей представление авторов грамматических пособий и словарей о языковой норме. Кодификация — это фиксация объективно существующей языковой нормы, сформулированная в виде правил (предписаний) в авторитетном лингвистическом издании (типа грамматики, учебника, словаря) и адресованная всем членам языкового коллектива» [12: 11–12].

Следует отметить, что норму кодифицируют не только «авторитетные лингвистические издания». Есть **терминологические словари**, которые содержат вполне адекватное описание профессиональной нормы для некоторых научных дисциплин и хозяйственных отраслей; но не для всех: вполне очевидно, что не всякий лингвист сочтет любой словарь лингвистических терминов (общий или специализированный, например, словарь социолингвистических терминов) полностью адекватным описанием лингвистической терминологии.

Есть официально утвержденные **стандарты**. Стандартизация продукции, а значит, и унификация терминологии необходима практически во всех производственных отраслях. Специализированный орган, занимающийся стандартизацией в рамках всего государства, существовал в СССР с 1925 г., именование его менялось. В РФ с 2004 года стандартизация находится в ведении Федерального агентства по техническому регулированию и метрологии (Росстандарт). Стандартизируется п р о д у к ц и я, но описание этой продукции кодифицирует с л о в о у п о т р е б л е н и е. Стандарты бывают отраслевыми, государственными и межгосударственными (в СНГ), а также международными (СССР, а затем Россия — член ISO/ИСО). Далеко не вся кодифицированная таким способом лексика имеет узкоспециальный характер. Есть «простые» слова, по разным (обычно загадочным) причинам игнорируемые толковыми словарями, но давно четко зафиксированные в ГОСТах. Ограничимся одним примером: ни в одном толковом словаре нет свиной *рульки*, широко продающейся в сыром и копченом виде. Давным-давно гостирован разруб свиной туши, где фигурирует и *рулька*, на производство и упаковку готовых мясных изделий также существуют стандарты, эти обязательные к исполнению общегосударственные документы можно считать кодифицирующими именование понятия «рулька».

Кроме ГОСТов, Росстандарт ответствен за разработку разнообразных **общероссийских классификаторов**. Есть «Общероссийский классификатор валют», где кодифицированы слабо отраженные толковыми словарями наименования денежных единиц постсоветского пространства[3]. Существует «Общероссийский классификатор территорий муниципальных образований» [15], где перечислены все поселения с указанием для каждого его типа, изредка неопределенно: *населенный пункт*, но обычно конкретизировано: *город*, *поселок*, *село*, *деревня*, *хутор*, *слобода*, *станция*, *погост*[4] и мн. др. Только что перечисленное — «обычные» русские слова, но в толковых словарях они могут объясняться неточно или не полностью. *Слобода* в БТС толкуется лишь как историзм («В России 11–17 вв. ‹…›»), у Шведовой в [20] есть также значение ‘посёлок около города, пригорода’, но помечено оно *устар*., *станция* — только пункт остановки транспорта, *погост* в обоих словарях только ‘кладбище’. В этом классификаторе закреплены и этнически маркированные типы поселений (188 *улусов* в Бурятии, 52 *аала* в Хакасии, 20 *арбанов* в Туве), словарями они толкуются либо неточно, либо игнорируются.

Есть **общие справочные издания**, в частности, энциклопедии. Адекватность сообщаемой там информации зависит от профессионализма авторов и единообразия интерпретации конкретной единицы на русскоязычном пространстве. Их основным назначением является сообщение энциклопедической информации, но вряд ли принятое в БСЭ (или в современной РСЭ) словоупотребление следует считать ненормативным узусом. А в подобных изданиях хватает общеизвестных лексических единиц, не попавших пока в толковые словари.

Есть **законодательство** и другого рода административные документы, при этом «на местах» и законотворцы, и те, кто ведет официальный документооборот, следуют местному узусу, создавая региональную норму (разнообразные примеры см. в [4]). Некоторые типы региональной документации требуют утверждения на общегосударственном уровне и, как следствие, местная норма de facto утверждается как общероссийская.

Слово *своротъ* ‘поворот (дороги, пути)’ в МАСе[13] получило помету *прост*., то есть в официальном речевом обиходе фигурировать не должно. Тем не менее оно регулярно используется в Иркутской области и Красноярском крае. Подготовленные там документы послужили основой

[3] Есть и занятное. Ко времени введения в действие современного классификатора (принят Постановлением Госстандарта России от 25 декабря 2000, действует 1.07.2001) в Таджикистане была введена денежная единица *сомони*, а в предыдущем классификаторе валют (действовал с 1.07.1995 по 1.07.2001) отразился переход в марте 1997 от именования *таджикский рубль* к таджикизированному написанию *таджикский рубл*. Это особенно примечательно, поскольку на советских рублях были и надписи на таджикском языке: *як сӯм* ‘один рубль’, *панҷ сӯм* ‘пять рублей’ и т. п.; независимый Таджикистан перешел от теоретического *сӯм* к реально функционировавшему и в советские времена *рубл*.

[4] Например, *погост Старая Никола* в составе Вахромеевского муниципального образования Камешковского района Владимирской обл. (код ОКТМО 17625408181).

«Паспорта инвестиционного проекта „Комплексное развитие Нижнего Приангарья“», утвержденного распоряжением Правительства РФ № 1708-р от 30.11.06, где среди «мероприятий, реализуемых в рамках проекта», упоминается *реконструкция автомобильной дороги Канск — Абан — Богучаны на участке Черемухово до* ***поворота*** *на Покатеево (км 124 — км 133) в Абанском районе Красноярского края*.

В нормированном языке юга Западной Сибири в том же значении используется слово *свороток*, последалевской общей толковой лексикографией вообще не фиксируемое. В Республике Алтай, Алтайском крае, Кемеровской области оно широко используется в кадастровой документации, в планах развития дорожного хозяйства и в других случаях, где необходимо упомянуть поворот на второстепенную дорогу. Археолог из Барнаула так указывает местоположение описываемого кургана: *приблизительно в 1,3 км к ЮВ от устья р. Куюм, напротив лесопилки, у свороткa с Чемальского тракта к последней* (Степанова Н. Ф. Погребения в каменных ящиках и их датировка // Погребальный обряд древних племен Алтая. Барнаул, 1996, стр. 54).

Главным различием нормы и кодификации Ицкович считает вполне естественное запаздывание последней. Представляется, что куда важнее субъективность кодификаторов. Как и «рядовые» носители литературного языка, при определении нормативности «кодификаторы ориентируются в первую очередь на собственный узус, во вторую — на узус с в о е г о круга, но лишь настолько, насколько этот узус пассивно знаком самим лексикографам» [3: 361; там же см. разнородные примеры, подтверждающие этот тезис]. Незнакомые реалии толкуются по не всегда аккуратным источникам. Рыба, именуемая на международном языке биологов *Stenodus leucichthys*, на северных реках называется *нельмой*, а в бассейне Каспийского моря (где она практически исчезла) — *белорыбицей*. В словаре Шведовой [20] *нельма* толкуется как «крупная северная рыба сем. лососевых», а *белорыбица* — как «северная промысловая рыба сем. сиговых с серебристой блестящей чешуей»[5]; в петербургских словарях — не совсем так, но столь же ошибочно.

Трактовка фауны и флоры в толковых словарях ориентирована на научную картину мира, как видим, делается это не всегда аккуратно. А «биологически аккуратные» толкования часто плохо соотносятся с картиной мира образованного русскоязычного «обывателя», язык которого и должен быть отражен в словаре. Один из авторов настоящего текста многократно в разных аудиториях (лингвисты, школьные учителя, студенты-филологи) и разных регионах (от Воронежа до Благовещенска и от Петербурга до Волгограда) предъявлял изображения трех растений с вопросом, которое из них *камыш*. Обычно большинство указывало на *Typha latifolia*, который толковые словари вслед за ботаниками именуют «рогозом», с ним конкурировал *Phragmites australis* (словарный «тростник»), а «камыш» (*Scirpus lacustris*) иногда вообще никто не считал *камышом*. Показательно, что некогда популярный строительный материал *камышит* изготовлялся из того, что в словарях является тростником.

Определенная доля общерусской лексики в повседневном узусе столиц (и лексикографов) не встречается. Как кажется, ни в Ленинграде/Петербурге, ни в Москве — в отличие от большинства городов СССР — не функционировали *уличкомы*[6]; этого слова нет и в словарях. Между тем утвержденное в 1996 г. «Положение об уличных комитетах (уличкомах) г. Воронежа» попало в качестве типового образца в хрестоматию по муниципальному праву [14: 339–342]. В этом случае можно говорить о своеобразной **антирегиональности**: «везде» есть, но где-то не встречается.

* * *

Существование региональных различий в норме вполне очевидно. Очевидны и причины игнорирования этого факта официальной русистикой. В постсоветской истории это всего лишь традиция, а прежде была и идеология.

В первые послереволюционные годы во многом оказалось неизбежным расшатывание нормы: с одной стороны, в общегосударственный коммуникативный процесс вовлекались широкие массы населения, недостаточно владевшие нормативным языком, с другой стороны, менялся

[5] Для жителей Дагестана и Азербайджана это «восточная рыба», в Казахстане и Туркмении — «западная», а «северной» она оказывается лишь при взгляде из Ирана.

[6] В окраинной Москве *уличкомы* вполне могли существовать, в частности, на территориях, вошедших в черту города 17 августа 1960 (города Очаково, Кунцево, Тушино, Бабушкин, Перово, Люблино, села Медведково, Тропарево и мн. др.).

спектр функций литературного языка и охватываемая им проблематика. Происходило это одновременно по всей стране, так что о единообразии результатов не могло быть и речи[7]. С середины 1930-х гг. можно говорить об относительной стабилизации нормы, но в области лексики «новая» норма довольно заметно отличалась от «старой»[8].

Реальное осмысление сложившейся социолингвистической ситуации началось лишь с 1950-х гг. В немногих работах, посвященных региональным лексическим расхождениям в языке города, они интерпретировались как влияние местных диалектов и просторечия. Редким исключением оказалась статья Р. Р. Гельгардта «О литературном языке в географической проекции», справедливо утверждавшего, что «местные различия ‹…› литературного языка могут и не иметь источников в народной диалектной среде. Тогда они являются только вариантами литературной нормы» [11: 98].

Однако такие взгляды были признаны идеологически вредными: «старые» диалекты отмирают, любые их следы в городской речи — пережитки, «новых» различий, возникших в рамках литературного языка, в принципе не может быть, поскольку для них нет социальной базы. Возобладал декларативный тезис о полном единообразии русского литературного языка на всей территории его распространения, «общеобязательности его норм как образцовых для всех, кто им владеет и пользуется, независимо от социальной, профессиональной и территориальной принадлежности» [19: 3]. Писалось это за семь лет до предполагавшегося стирания классовых границ, а бесклассовому обществу положен монолитный язык.

Упоминание социальной принадлежности параллельно с профессиональной и территориальной указывает на довольно примитивное понимание *социального* всего лишь как *классового*, что характерно для далекой от социологии части научного сообщества, которая использовала собственные элементарные познания в этой сфере в административно-идеологической борьбе[9]. Профессионалам же известно, что любой город и другой населенный пункт представляет собой самостоятельный социальный организм, устроенный иногда сложно, иногда очень сложно. Каждый из них занимает собственное место в иерархически организованной системе «однотипных» социальных организмов — поселений. Однотипность тут условная, всякое поселение имеет свое лицо, определяемое многими показателями: историей образования, численностью населения, родом занятий жителей, местом в административной иерархии, физико- и экономикогеографическим положением, развитостью культурной среды, сетью учебных заведений и другими параметрами, вплоть до локальных мифологем. И было бы удивительно, если бы все это не находило отражения в лексиконе, в частности, в его нормативной части.

## 4 Технологии поиска и верификации регионализмов

Для задачи выявления региональной вариативности наиболее принципиальным вопросом является определение надежных регионально маркированных источников данных, причем в том количестве, которое позволяет выносить статистически значимые суждения. Стандартный «пайплайн», приводящий к появлению новых словарных входов регионального словаря, состоит из следующих этапов:

1. Поиск кандидатов на статус региональных нормированных вариантов значений, входящих в некоторый универсальный национальный Инвентарь Значений, сущности, которая, увы, в реальности никак не представлена. Далее мы будем называть **регионализмами** именно такие единицы, отличая их от топонимов и другой **регионально смещенной лексики**.
2. Определение достоверной региональной картины употребления регионализмов.
3. Лексикографическое описание.

Остановимся более подробно на каждом из этих этапов.

[7] По замечанию Р. О. Шор, в дореволюционном языке имелись значительные лексические лакуны, например, среди «терминов кухни и домашнего хозяйства». «Очевидно, что при отсутствии соответствующих слов в „литературном языке“ „образованные классы“ общества принуждены заимствовать их из народных говоров данной местности» [21: 137].

[8] Не случайно в словаре под ред. Д. Н. Ушакова появились пометы *новое* («слово или значение возникло в русском языке в эпохи мировой войны и революции») и *дореволюционное* («слово обозначает предмет или понятие, вытесненные послереволюционным бытом») [18: XXVII—XXVIII].

[9] Вообще-то такой взгляд заслуживает старой советской этикетки «вульгарный социологизм». В действительности социально в человеке все, что не обусловлено исключительно биологией.

### 4.1 Поиск кандидатов

Есть целый ряд проблем, типичных для современных социолингвистических исследований, препятствующих эффективному поиску таких объектов автоматически:

- малое число надежных полномасштабных источников региональных данных;
- проприетарность таких источников или существенные ограничения в их академическом использовании;
- ложная/неточная региональная атрибуция текстов и/или авторов, связанная, в частности, с принципиальным различием геометок, ассоциированных с местом текущего пребывания автора (геолокация) и местом его рождения. Заметим, что на идиолект могут действовать оба фактора, при этом первый является случайно смещающим реальную региональную картину.

На первом этапе проекта ЯРГ основным источником потенциальных регионализмов были предложения, сделанные участниками форумов Lingvo, в основном, профессиональными переводчиками, представляющими самые разные регионы России и Ближнего Зарубежья и привыкшими не только внимательно относиться к нюансам в употреблении слов, но и обсуждать эти нюансы с “peer-to-peer” коллегами. Это обусловило высокий КПД обсуждений, позволивший быстро набрать наиболее очевидные частотные регионализмы. Число таких регионализмов, несколько тысяч, оказалось сюрпризом как для донаторов (для которых сама идея наличия таких слов в их собственных идиолектах вовсе не была очевидной), так и для идеологов проекта.

Этот результат можно считать наиболее значительным, поскольку он показывает **реальность существования региональной вариативности** не только в узусе, но и **в норме** в статистически значимых объемах: вывод, с которым обязаны теперь считаться любые лексикографы, занимающиеся толковыми словарями русского языка.

С другой стороны, переход от наиболее частотных и очевидных регионализмов к менее частотным требует уже иных методов: увеличение числа участников обсуждений неизбежно приводит к падению среднего качества предложений, росту «фейковых» обсуждений и т.п. Это подвело естественную черту под первым этапом проекта ЯРГ.

Второй этап проекта начался с появлением корпуса ГИКРЯ, который позволяет не просто проверять региональное смещение запроса, но и проводить сплошную обработку регионально маркированных текстов в поисках кандидатов на регионализмы. Сразу скажем, что этот этап проекта еще не полностью реализован, и речь пойдет о тех подходах, которые активно исследуются, и о некоторых предварительных выводах из этих исследований.

Здесь следует немного отвлечься от основной темы статьи и коснуться некоторых побочных, но важных тем, связанных с оценкой качества регионально маркированных подкорпусов ГИКРЯ, особенностям распределения регионально окрашенной лексики и возможности ее использования для задач извлечения кандидатов и автоматической региональной классификации текстов и регионального профилирования авторов (как для расширения корпуса, так и для верификации априорной разметки).

Основные выводы следующие:

1. Регионализмы имеют очень низкую плотность распределения в текстах. Кроме того, в отличие от других социолингвистических категорий (например, гендера и возраста), региональные признаки крайне неравномерно представлены в доступных для сплошного компьютерного анализа данных, что препятствует получению статистически значимых результатов для многих «маленьких», хотя и, возможно, интересных в языковом отношении регионов.
2. В результате исследований по автоматической региональной классификации текстов на основании использования региональных словарей был получен вполне естественный негативный результат: высокая точность идентификации при крайне низкой полноте [17]. Конкретные цифры есть в статье, но сейчас неважны, поскольку были получены на довольно грязных данных.
3. Автоматическая региональная классификация текстов стандартными методами классического и глубокого обучения дает результаты, которые значимо выше случайных (в особенности в искусственных условиях подбора максимально ортогональных и при этом хорошо представленных укрупненных регионов, например, «Краснодарский край и Кавказ», «Урал

и Сибирь», «Восточная Украина и Киев»), но не позволяют рассматривать такие методы как надежное средство верификации априорной или получения новой региональной разметки.

4. При исследовании значимости признаков выявлено, что основной вклад в качество региональной идентификации произвольного текста вносят **не регионализмы, но другие виды регионально смещенной лексики**, прежде всего, топонимы и прочие регионально значимые именованные сущности и нерегиональная по сути лексика, связанная с важными локальными событиями. Это, разумеется, вполне предсказуемый результат, который, тем не менее, стоило проверить.

Таким образом, для решения двойной задачи расширения регионально размеченных данных и автоматического поиска регионализмов необходимо:

- Перейти от задачи классификации текстов к задаче авторского профилирования (тем самым решая проблему низкой плотности и неравномерности распределения регионально смещенной лексики любого типа).
- Задачу поиска кандидатов в регионализмы решать статистическими методами, элиминируя из ранжированных списков прочие типы регионально смещенной лексики.
- Универсальным средством повышения качества является очистка данных. Эта задача в значительной степени решена в новой версии ГИКРЯ (см. [9].)

### 4.2 Определение «карты» употребления регионализмов

Итогом первого этапа является список кандидатов в регионализмы, полученный как в результате предложений участников проекта (в версии ЯРГ), так и автоматическим анализом регионально смещенной лексики по данным ГИКРЯ.

Этап проверки реальной картины регионального распределения не получается пока делать полностью автоматически. Эту проблему нам еще предстоит решить, прежде всего увеличением как общего объема корпуса, так и применением автоматических методов профилирования для увеличения регионально маркированной части, которая и сейчас весьма значительна, но крайне неравномерно представляет регионы, которые, естественно, существенно различаются числом авторов в соцсетях.

Для верификации регионального распределения регионализмов в проекте ЯРГ использовалась, помимо соцсетей, и база данных региональных СМИ «Интегрум». К сожалению, этот замечательный ресурс не имеет API-доступа к текстам для пользователей, что не позволяет использовать его не только для верификации, но и для поиска кандидатов. Возможно, этот вопрос удастся когда-нибудь решить.

Результатом верификации являются данные, подобные приведенному ниже распределению вариантов *магаз/магазин/лабаз*:

В Ленинграде 1980-х лабаз — обычно винный магазин, у современных авторов младших возрастов — любой магазин. Вот цитаты из старшего поколения:

В. Гаврильчик (1929–2017): *Одиннадцать пропикало, / Народ бежит в лабаз (1978); Нас мотало в метро. И в лабазах давили. / В зной и стужу стояли мы у пивного ларь*ка (1979).

*За водкой и более деликатными алкоголями бились (в буквальном смысле слова) в специальных отделах, полуподвалах, лабазах, и все равно только БЛАТ давал вожделенную влагу в нужном для праздника количест*ве (Сергей Юрский. Вспышки, 2001, В Москве с 1978 года, с 43 лет.)

**Региональный и возрастной анализ в ЖЖ ГИКРЯ (без учета семантики):**

Нужен эталон для сравнения. 1. нейтральное *в магазин* и 2. «молодёжное» *в магаз*.

**Избранные регионы** (число словоупотреблений)

| **Регион** | *в магаз* | *в магазин* | *в лабаз* |
|---|---|---|---|
| Мск | 2621 | 38001 | 58 |
| СПб | 749 | 10893 | 52 |
| Моск. обл. | 100 | 1835 | 4 |
| Ленобласть | 37 | 611 | 6 |
| Мск/СПб | 3,50 | 3,49 | 1,12 |
| **Всего** | 10130 | 178215 | 350 |

**Доля избранных регионов, %.**

| **Регион** | *в магаз* | *в магазин* | *в лабаз* |
|---|---|---|---|
| Мск | 25,9 | 21,3 | 16,6 |
| СПб | 7,4 | 6,1 | 14,9 |
| Моск. обл. | 1,0 | 1,0 | 1,1 |
| Ленобласть | 0,4 | 0,3 | 1,7 |

Слово явно петербургское.

**Разбивка по возрасту** (число словоупотреблений)

| **Год рождения** | *в магаз* | *в магазин* | *в лабаз* | *в магазин / в лабаз* |
|---|---|---|---|---|
| 1950–1969 | 72 | 3508 | 18 | 194,9 |
| 1970–1979 | 609 | 13582 | 58 | 234,2 |
| 1980–1999 | 2448 | 30279 | 30 | 1009,3 |
| 1950–1999 | 3129 | 47369 | 106 | 446,9 |

**Доля отдельных когорт, %**

| **Год рождения** | *в магаз* | *в магазин* | *в лабаз* |
|---|---|---|---|
| 1950–1969 | 2,3 | 7,4 | 17,0 |
| 1970–1979 | 19,5 | 28,7 | 54,7 |
| 1980–1999 | 78,2 | 63,9 | 28,3 |

Выражение *в магазин* дает обычное возрастное распределение для нейтральной семантически умеренно маркированной лексики (для немаркированной доля младшей возрастной когорты не сильно превышает 50%, но молодежь походы в магазин упоминает чаще).
Межпоколенная разница в тяге к хождению *в магазин* и *в лабаз* очевидна.

### 4.3 Лексикографическое описание.

Региональный словарь является толковым. Поэтому вопросы собственно лексикографического описания регионализмов не имеют какой-то очевидной специфики, если не считать наличия в словарной статье собственно региональных признаков.

Лексикографическое описание является трудоемким процессом, особенно в ситуации, когда идиолект лекскографа не включает описываемого значения. Ограниченность «редакторского» ресурса привела к тому, что не все очевидные регионализмы, собранные в ходе проекта ЯРГ, получили полноценное лексикографическое описание. Так, в словарь ЯРГ под Lingvo была включена лишь примерно четверть того, что было предложено к обсуждению, и около половины того, что получило статус проекта словарной статьи.

Образцы словарных статей (в условном формате) можно увидеть ниже:

**смитник**
мусорный бак, мусорное ведро, помойка
**Ex***: Эти, юные тогда одесситы, были затем высоко (аж до Харькова, Киева и Москвы) подняты гребнем 1920-х годов, раздавлены и вышвырнуты на смитник страны в конце 1930-х…* (Порто-Франко, Одесса; 23.03.2007
**Syn**: альтфатер, жбан, мульда, пухто
**Reg**: Одесса, возможно, вся Украина

**альтфатер**
мусорный контейнер с крышкой
**Ex***: Женщина (откуда-то с провинции, вроде как с Москвы) сильно хочет выкинуть кулек с мусором в альтфатер…* (блог, Одесса); *Пацаны вышли за хлебом, таз атаковал альтфатер* (форум, Черновцы).

**Syn**: жбан, мульда, пухто, смитник
**Reg**: Одесса, Черновцы

**пухто**
*тж*. пухта
мусорный контейнер
**Ex***: Там, во дворах, двадцать лет стояли баки и пухто, но ЖКС № 2 решил их убрать, хотя деньги за вывоз отходов с жильцов собирает* (Невское время, Санкт-Петербург; 15.12.2000); *Оставленную пустую пухту заполнят за несколько часов, и через день за ней снова приедет КамАЗ* (Сельская новь, Волосово, Ленинградская область; 11.06.2005).
**Syn**: альтфатер, жбан, мульда, смитник
**Reg**: Петербург, возможно вся Ленинградская область

**мульда**
мусорный бак
**Ex***: Кстати, скоро в Ростове появится порядка трех тысяч новых контейнеров, но вблизи строек установят не их, а более вместительные мульды* (Наше время, Ростов-на-Дону; 05.08.2003); *Ни кабинок для переодевания, ни душевых кабинок, и даже мульд для мусора* (АиФ Удмуртии; 17.07.2003).
**Syn**: альтфатер, жбан, пухто, смитник
**Reg**: Ижевск, Ростов-на-Дону

Объем статьи не дает возможность осветить устройство регионального классификатора: это вопрос, в котором перемешаны социолингвистические, культурно-исторические и административные соображения. Скажем только, что в ГИКРЯ используется трехуровневый классификатор (грубо говоря, страна — регион — город), в сильной степени по необходимости связанный с системой региональных признаков соцсетей. В ЯГеЛе используется несколько отличная система, это несоответствие еще предстоит преодолеть. Общий объем региональных признаков – около 1000. Все в целом делает задачу автоматической классификации весьма нетривиальной.

## 5 Интерактивный региональный словарь проекта ЯГеЛь

Желание вернуть к жизни форум ЯРГ, перенеся его из ставшего чужим проекта Lingvo в «профильный» ГИКРЯ, возникло давно. Однако этому мешала необходимость решить сначала вопросы строительства новой версии корпуса, что оказалось делом непростым и небыстрым, поскольку проект около 5 лет оставался (и остается) без грантовой поддержки. Как оказалось, эксперты полагают, что «корпус Русского Языка у нас уже есть». Фактически развитие проекта проходило и проходит за счет практики и НИР студентов специальности Компьютерная Лингвистика МФТИ и РГГУ при общей оргподдержке отдела перспективных исследований ABBYY и с 2020 года – Лаборатории ABBYY Lab в Физтех-Школе ПМИ МФТИ.

Проект ЯГеЛь был начат студентами МФТИ в ходе т.н. ИннПрака (рук. Т.О. Шаврина). Была сделана попытка поэкспериментировать с региональной разметкой первой версии ГКИРЯ с прицелом на интересы неискушенного пользователя, «любителя слов».

В дальнейшем идеи этого проекта были несколько переосмыслены и в настоящее время реализуется идея объединенного портала для дифференциальных социолингвистических исследований, включающего:

1. интернет-корпус ГИКРЯ новой расширенной и очищенной от неавторских текстов версии;
2. интерактивный словарь ЯГеЛь (Языки Городов и Людей), в состав которого вошли полностью словарные и методические материалы проекта ЯРГ, импортированные из формата форума Lingvo (уже закрытого) в более современный формат Wiki-dictionary. Таким образом, сотни пользователей этого ресурса вновь получают доступ к нему: https://int.webcorpora.ru/reg2/index.php/Языки_городов_и_людей

В отличие от проекта ЯРГ, который был связан с поиском и описанием только региональной лексики, ЯГеЛЬ включает материалы для словаря паремий и разделы для будущих гендерных и возрастных словарей.

## Благодарности

Мы благодарим за соучастие и поддержку всех участников Lingvo-форума «Языки русских городов» и в особенности редактора первой версии регионального словаря Марию Ахметову!

Мы также глубоко признательны участникам первой студенческой версии ЯГеЛя, сделанной в рамках ИннПрака МФТИ под руководством Татьяны Шавриной.

## References

[1] Akhmetova M.V. (2014), Lexical regionalisms and localisms in Runet: problems of collecting materials [Leksitscheskie regionalism i lokalismy v russkoyazychnom Internete: problem sbora materiala] // Russian language and new technologies [Russkiy yazyk i novye tehnologii], Moscow, pp. 156-171.

[2] Akhmetova M.V. (2015), From A-Aty till Yarsk: unofficial townnames dictionary [Ot A-Aty do Yarska: slovar neofitsyalnykh nazvaniy naselennyh punktow]. – M:FORUM, 2015. – 496 p.

[3] Belikov V.I. (2009), Stereotypes in literary norms understanding [Stereotipy v ponimanii literaturnoi normy] // Language, communication and culture stereotypes [Stereotipy v yazyke, kommunikatsyi i kulture], Moscow, pp. 357-377.

[4] Belikov V.I. (2009), Lexical usus of official documents and codified dictionary norm [Leksicheskyi uzus ofitsyalnych documentov I kodifitsirovannaja slovarnaja norma] // Social language options – VI [Sotsyalnye variant yazyka - VI], Nizhniy Novgorod, pp. 65-68.

[5] Belikov V.I. (2010), Metodic news in the social lexicography of the XXI century [Metodicheskie novosti v sotsialnoj leksikografii XXI veka] // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian, Helsinki University Press, pp.32 -49.

[6] Belikov V.I. (2012), On the method of identifying isoglossae of urban regionalisms [K metodike vyavlenia isogloss gorodskih regionalismov] // Modern problems of cultural and linguistic regionalism [Sovrememmye problem kulturno-yazykovoi regionalistiki], Perm, PONITSAA, pp. 8-14.

[7] Belikov V.I. (2014), On the methodology of corpora research of vocabulary [K metodike korpusnogo issledovanija leksiki] // Russian and new languages [Russkij yasyk i novyje], Moscow, pp. 99-130.

[8] Belikov V.I. (2016), Беликов В. И. What can a linguist get from digitized texts and its ways [Chto i kak mozhet poluchit lingvist is otsifrovannykh tekstov]// Siberian Philological Journal [Sibiskij filologicheskij zhurnal], No 3, pp. 17 -34.

[9] Belikov V.I., Selegey V.I., Selegey D.V. (2020), Web-corpus as a tool for linguistic research: differentiation, authorization, thematic biases (or corpora we want so much to believe) [Internet-korpus kak instrument lingvisticheskikh issledovanij: differentsialnost, avtorixatsija, tematicheskije smeshenija (ili korpusy, kotorym tak hochetsa verit'] – Computational Linguistics and Intelligent Technologies [Kompjuternaja lingvistika i intellektualnyje tekhnologii].

[10] Comprehensive Explanatory Dictionary of the Russian Language [Bolshoj tolkovyj slovar russkogo yazyka], SPb, 1998.

[11] Gel'gardt R.R. (1959), About the literary language in the geographical area [O literaturnom yazyke v geograficheskoj proektsii] // VJA, No 3.

[12] Itskovich V.A. (1982), Essays on the syntactic norm - M.: Nauka.

[13] Evgenjeva A.P. (1981 – 1984), Russian Language Dictionary, 4 vol. [Slovar russkogo yazyka], vol. 2, Moscow, Rus.yaz.

[14] Belousova E.V. (1999), Municipal Law of the Russian Federation: Reader [Munitsypalnoe pravo Rossijskoj Federatsii: Khrestomatija], M.: Jurist, 544 p.

[15] ARCoMT: All-Russian classifier of municipal territories [OKTMO: Obsherossijskij klassificator territorij municipalnykh obrazovanij], OK 033-2013, M.: Standartinform, 2013.

[16] Dictionary of Russian folk dialects [Slovar russkikh narodnykh govorov], vol. 1 et al, M. – L.: Nauka, 1965 - …

[17] Sorokin A.A. (2015), Automatic regional classification based on the dictionary of regional vocabulary: a trial study [Avtomaticheskaja regionalnaja classificatsija na osnove slovarja regionalnoj leksiki: probnoje issledovanije], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2015" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2015"], Bekasovo.

[18] Ushakov D.N. Explanatory dictionary of the Russian language [Tolkovyj slovar russkogo yazyka], vol. 1, M.:SE, OGIZ, 1935.

[19] Filin F.P. (1973), On the structure of the modern Russian Literary language [O structure sovremennogo russkogo literaturnogo yazyka] // Linguistics issues [Voprosy yazykpznanija], No 3.
[20] Shvedov N.J. Explanatory dictionary of the Russian language with the etymology of the words [Tolkovyj slovar russkogo yazyka s vklucheniem informatsii o proishozhdenii slov] / Institute of the Russian Language of the Russian Academy of Sciences [Institut russkogo yazyka RAN], M.: Azbukovnik, 2007.
[21] Shor R.O. (1926, 2009), Language and Society [Yazyk I obschestvo], vol. 3, M.: Lenand.
[22] Age and Gender Identification in Unbalanced Social Media. Juan Carlos Gomez, Luis-Miguel López-Santamaría, Mario-Alberto Ibarra-Manzano, Dora-Luz Almanza-Ojeda, 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2019

## Список литературы

[1] Ахметова М. В. Лексические регионализмы и локализмы в русскоязычном Интернете: проблемы сбора материала // Русский язык и новые технологии: М.: Нов. лит. обозрение, 2014. С. 156–171.
[2] Ахметова М. В., От А-Аты до Ярска: словарь неофициальных названий населенных пунктов / Отв. ред. В. И. Беликов. М.: ФОРУМ, 2015. 496 с.
[3] Беликов В. И. Стереотипы в понимании литературной нормы // Стереотипы в языке, коммуникации и культуре. М.: РГГУ, 2009-а. Стр. 357 -377.
[4] Беликов В. И. Лексический узус официальных документов и кодифицированная словарная норма // Социальные варианты языка — VI. Нижний Новгород: НГЛУ, 2009-б. Стр. 65 -68.
[5] Беликов В. И. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian / Editors: A. Mustajoki, E. Protassova, N. Vakhtin. Helsinki: — Helsinki University Press, 2010. Pp.32 -49.
[6] Беликов В. И. К методике выявления изоглосс городских регионализмов // Современные проблемы культурно-языковой регионалистики. Пермь: ПОНИЦАА 2012. С. 8 -14.
[7] Беликов В. И. К методике корпусного исследования лексики // Русский язык и новые. М.: Нов. лит. обозрение, 2014. С. 99 -130.
[8] Беликов В. И. Что и как может получить лингвист из оцифрованных текстов // Сибирский филологический журнал. 2016, № 3. С. 17 -34.
[9] Беликов В.И., Селегей В. П., Селегей Д. В. Интернет-корпус как инструмент лингвистических исследований: дифференциальность, авторизация, тематические смещения. В сб. «Компьютерная лингвистика и интеллектуальные технологии» 2020.
[10] Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. — СПб., 1998.
[11] Гельгардт Р. Р. О литературном языке в географической проекции // ВЯ, 1959, № 3.
[12] Ицкович В. А. Очерки синтаксической нормы. М.: Наука, 1982.
[13] Словарь русского языка: В 4 т. / Под ред. А. П. Евгеньевой. — 2-е изд., испр. и доп. — М.: Рус. яз., 1981–1984
[14] Муниципальное право Российской Федерации: Хрестоматия / Сост. Е. В. Белоусова. — М.: Юристъ, 1999. 544 с.
[15] ОКТМО: Общероссийский классификатор территорий муниципальных образований. ОК 033-2013. Тт. 1 -8. М.: Стандартинформ, 2013.
[16] Словарь русских народиых говоров. Вып. 1 (и следующие, издание не завершено) М.-Л.: Наука, 1965-…
[17] Сорокин А. А. Автоматическая региональная классификация на основе словаря региональной лексики: пробное исследование, Компьютерная лингвистика иинтеллектуальные технологии. По материалам ежегодной конференции Диалог, Бекасово, 2015
[18] Толковый словарь русского языка: В 4 т. Т. 1. / Под ред. Д. Н. Ушакова. — М.: СЭ; ОГИЗ, 1935.
[19] Филин Ф. П. О структуре современного русского литературного языка // Вопросы языкознания, 1973, № 2.
[20] Толковый словарь русского языка с включением сведений о происхождении слов / Ин-т рус. яз. РАН. Отв. ред. Н. Ю. Шведова. — М.: Азбуковник, 2007.
[21] Шор Р. О. Язык и общество. М., 1926. Издание 3-е. М.: Ленанд; 2009
[22] Age and Gender Identification in Unbalanced Social Media. Juan Carlos Gomez, Luis-Miguel López-Santamaría, Mario-Alberto Ibarra-Manzano, Dora-Luz Almanza-Ojeda, 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2019