

*С.О. Савчук*  
*S.O. Savchuk*

**ЛЕКСИКО-СЕМАНТИЧЕСКИЕ ОСОБЕННОСТИ  
ТЕКСТОВ РЕГИОНАЛЬНЫХ СМИ ПО ДАННЫМ  
КОРПУСНОГО ОБСЛЕДОВАНИЯ<sup>1</sup>  
CORPUS-BASED STUDY OF LEXICO-SEMANTIC  
PECULIARITIES OF REGIONAL MEDIA TEXTS**

**Аннотация.** В статье описывается новый корпус региональных и зарубежных СМИ, включенный в состав НКРЯ, приводится краткий обзор лексических средств с региональной спецификой, который извлекается с помощью корпусного инструментария, и оцениваются перспективы использования корпуса для анализа региональной вариативности.

**Ключевые слова.** *региональные СМИ, лексико-семантические особенности, типы регионально специфической лексики, корпусное исследование*

**Abstract.** The paper presents a new corpus of regional media within the Russian National Corpus. The corpus includes regional paper texts published in the territory of the Russian Federation and the near abroad and is aimed at studying the specific of regional varieties of the Russian language. Usage of proper names, culture-specific vocabulary, shifts in the meaning and frequency behavior, variation in synonymic correlation of Russian words are analyzed and possibilities of corpus based analysis as a tool of regional variation study are demonstrated.

**Keywords.** *regional media, lexico-semantic features, types of region-specific vocabulary, corpus-based analysis*

---

<sup>1</sup> Исследование выполнено при финансовой поддержке РГНФ, проект № 15-04-12018 и программы ОИФН РАН «Язык и информационные технологии».

## **Введение**

Исследование регионального варьирования русского языка в последние годы стало одним из актуальных направлений в рамках социолингвистического подхода к изучению языка, что отражается в большом количестве публикаций. Поиск региональных особенностей в большинстве исследований ведется прежде всего на материале устной разговорной речи, однако его оказывается недостаточно, если ставится задача описания региональных или национальных вариантов языка. В этом случае необходимо обследовать кодифицированные сферы литературного языка и прежде всего региональные газеты и электронные СМИ. В последнее время отмечается большой потенциал блогосферы для проведения регионально ориентированных исследований<sup>2</sup>. Суть подхода, предложенного в рамках НКРЯ, состоит в том, чтобы создать большой открытый газетный корпус, объединив в нем тексты СМИ различных типов, уровней и региональной принадлежности и снабдив этот массив лингвистической разметкой и инструментами поиска.

В 2014 году пилотный проект Корпуса региональных и зарубежных СМИ открыт для свободного доступа по адресу <http://ruscorpora.ru/search-regional>. В настоящее время в корпусе три модуля: 1) русскоязычная пресса зарубежья, 2) русскоязычные газеты субъектов РФ, 3) региональные выпуски федеральных изданий РФ [Савчук 2014]. Первый модуль представлен подкорпусом русскоязычных газет Гродненского

---

<sup>2</sup> Упомянем словарь «Языки русских городов», при создании которого используется метод краудсорсинга, исследование Яндекса «Местные слова в поисковых запросах», Генеральный интернет-корпус русского языка (ГИКРЯ), который пока находится в стадии разработки и не имеет свободного доступа [Пиперски 2013].

региона за 2012 год (около 2 млн словоупотреблений), подготовленным совместно с белорусскими коллегами в рамках проекта, поддержанного РГНФ и БРФФИ. Второй модуль содержит два подкорпуса российских региональных газет за период 2000-х и 2010-х годов (объемом около 2 млн каждый). Третий модуль включает региональные выпуски газеты федерального уровня «Комсомольская правда» (в регионах России, а также в Беларуси, Молдове, Кыргызстане, странах Балтии, всего около 7 млн словоупотреблений).

В настоящей статье приводится краткий обзор лексических единиц с региональной спецификой, которые извлекаются с помощью корпусного инструментария. Цель проведенного анализа – выявить потенциал газетных текстов как источника регионально-специфических средств и оценить перспективы использования корпуса для исследования региональной вариативности.

### **Типы регионально-специфической лексики**

В текстах корпуса были обнаружены элементы различных типов, которые можно считать регионально-специфическими<sup>3</sup>.

(1) Первую группу составляет «безэквивалентная лексика» (или «реалии», «экзотизмы», «варваризмы», «локализмы» и др.) – лексемы, заимствованные из других языков, функционирующих в полиэтнических регионах в контакте с русским языком. Один из эффективных корпусных инструментов выявления единиц этого типа опирается на лингвистическую аннотацию корпуса, в частности, на разграничение «словарных» и «несловарных» слов.

---

<sup>3</sup> При этом задача выявления различий в функционировании регионально-специфической лексики в российских и зарубежных изданиях в настоящей статье не ставится.

Последние не включены в грамматический словарь корпуса, поскольку не учтены в большинстве словарей, и с высокой долей вероятности могут содержать и регионально специфическую лексику. В газетных текстах корпуса широко представлены следующие группы реалий.

1) **Имена собственные** – антропонимы, топонимы, эргонимы. Доля имен собственных в газетных текстах очень высока, и это нашло отражение в качественном составе несловарных слов. В частности, в списках несловарных существительных, составленных для прибалтийских выпусков КП, доля имен собственных превышает 63%, для киргизских выпусков – 64%.

2) **Общественно-политические реалии** – лексика, называющая политические учреждения и общественные явления. К этой группе относятся, например, названия единиц административного устройства: населенных пунктов: *айыл* (Кыргызстан); органов и носителей власти: *примэрия* ‘орган исполнительной власти’, *муниципалитет*, *примар*, *преатура*, *баишкан* ‘высшее должностное лицо в Гагаузии’ (Молдова); *Жогорку Кенеш* ‘Верховный совет’, *торага* ‘председатель’, *айыл окмоту*, *акимият* ‘органы исполнительной власти’ (Кыргызстан); политических партий: *атажуртовцы*, *арнамысовцы*, *ата-мекеновцы* (Кыргызстан) и др.

3) **Этнографические реалии** – лексика, обозначающая понятия, связанные с бытом, материальной и духовной культурой. Сюда относятся, например, названия блюд национальной кухни: *тартачки*, *пызы*, *цепелины* (Беларусь), *вертута*, *пацинда*, *кэрнэцеи* (Молдова); одежды: *кашуля* ‘традиционная рубашка’ (Беларусь), *кушма* ‘шапка из овечьей шерсти’ (Молдова); денежных единиц: *сом*, *тенге* (*тенга*), *тыйын* (Кыргызстан); праздников, обычаев: *кирмаш* (Беларусь), *ала качуу* (Кыргызстан); строений и сооружений: *каменица* ‘каменный жилой дом’, *тартак* ‘лесопилка’ (Беларусь) жанров фольклора: *кошок*, *арман*, *шырылдан* (Кыргызстан), культовых сооружений, предметов,

праздников, служителей: *Нооруз, байрам*, обряд «*нике*» (Кыргызстан); *арцибискуп, бискуп, ружанцы* ‘четки’ (Беларусь) и мн. др.

(2) Вторую группу регионально-специфической лексики составляют новые лексемы, образованные по продуктивным моделям от исконных или давно освоенных слов. Они, как правило, не зафиксированы в словарях и имеют территориально ограниченное распространение. Единицы этого типа, подобно реалиям, выявляются в процессе сплошного обследования несловарных слов. Примером может служить прилагательное *лозовой* ‘изготовленный из лозы’, ‘относящийся к изделиям из лозы’. Прилагательное широко распространено на территории Беларуси и Украины. *Свислочская фабрика лозовых изделий; Черниговская фабрика лозовой мебели производит мебель из лозы ротанга и абак с каркасом сосна; Чесночница лозовая - прекрасный подарок для хозяйки. Мат лозовый; Лозовые циновки; Что самое главное нужно лозовому мастеру, чтобы удалось сотворить задуманное?* [по материалам сайтов]. В российских регионах, в которых занимаются лозоплетением, прилагательное употребляется значительно реже, вместо него используются синонимические конструкции – *мебель из лозы, плетеная мебель*.

Соотношение синонимических наименований может быть также «регионально зависимым». Неофициальное наименование для работников жилищно-коммунального хозяйства – *коммунальщики* и реже, в отдельных регионах, – *жилищники*. Ср., например, названия управляющих организаций в сфере ЖКХ: «*Жилищник*» в Смоленске, Владимире, Омске, Кемерове и «*Коммунальщик*» в Ижевске, Красноярске, а в Барнауле так называется хоккейный клуб. Общее соотношение лексем *жилищник* и *коммунальщик* в региональном корпусе – 60 к 642, а в большом газетном корпусе, сформированном в основном из центральных газет, – 97 к 2521, то есть налицо подавляющее преобладание второго варианта. Однако если рассмотреть распределение лексем по региональным газетам, обнаружатся

регионы, в которых соотношение двух номинаций отличается от общероссийского. Например, в казанских газетах разрыв между *жилищниками* и *коммунальщиками* сокращается и составляет 10 к 31 в выпусках КП и 3 к 3 в местных газетах; а в Омске, по данным местных выпусков КП, это соотношение равно 11 к 3 в пользу *жилищников*. В подкорпусе белорусских газет *жилищник* не зафиксирован ни разу, *коммунальщик* – 6 раз.

(3) Третью группу образуют семантические регионализмы – слова русского языка, употребляющиеся на какой-либо территории в значениях, не зафиксированных в других регионах. Семантические регионализмы невозможно обнаружить с помощью метода несловарных слов, поскольку в плане означаемого они не выходят за пределы словарного фонда литературного языка. Однако о сдвиге в значении может свидетельствовать непривычная сочетаемость, которая обнаруживается при анализе частотных списков коллокаций. Например, в списке биграмм по региональному корпусу в *застенках* (1), *фашистских застенках* (2), *шляхетских застенков* (1) отмечаем последнее сочетание, которое приводит нас к следующему контексту: *Вплоть до Второй мировой войны поселения вокруг Гродно составляли очень интересный конгломерат деревень, шляхетских застенков, хуторов и владений-маёнтков, где жили крупные землевладельцы.* [«Вечерний Гродно», 2012.09.05]. Здесь слово *застенок* употребляется не в одном из основных значений, отмечаемых словарями ('задняя комната или место между двух стен'; 'стар. малый придел, отделенный от главного храма стеною'; 'место пыток, при допросах' и др.), а в региональном значении, отмеченном в словаре Даля для «западного наречия»: «на старинном западном наречии, полоса, участок земли и леса <...>. *Поньине* (от *sciana*, стена, межа) односелье, хутор, заимка, однодворок, одинокий поселок на пустоши, вне общей крестьянской межи».

Таким образом, анализ частотных списков словоформ и коллокаций, созданных на материале текстовых корпусов, можно рассматривать как метод поиска регионализмов семантического типа. Дальнейшее развитие корпуса предусматривает создание списков n-грамм по корпусу и организацию поиска по ним.

(4) Четвертую группу можно назвать узуальными регионализмами, ее формируют лексемы, как правило, многозначные, которые для разных значений имеют регионально распределенные частотные характеристики: значение, высокочастотное в одном регионе, может оказаться низкочастотным в другом регионе. Методом обнаружения таких единиц также является анализ частотных списков словоформ, лексем и коллокаций.

Слова *агрогогородок* (в Белоруссии это официальное название одного из типов сельских населенных пунктов) и *агроусадьба* (объект экологического туризма) имеют выраженную региональную специфику, которая проявляется в различии частотных характеристик. Так, в белорусских газетах в составе регионального корпуса слово *агрогогородок* встретилось 296 раз (на 2,7 млн словоупотреблений), а в российских – лишь 4 раза (на 10,1 млн), причем два раза – в материалах о Беларуси. В корпусе современных российских СМИ (173 млн словоупотреблений) зафиксировано всего 19 вхождений этого слова. Слово *агроусадьба*, при 77 вхождениях в белорусских газетах регионального корпуса, в корпусах российских газет вообще не зафиксировано и 2 раза встретилось в основном корпусе (в публикации журнала «Туринфо» за 2002 г.).

Корпус региональных и зарубежных СМИ успел зарекомендовать себя как эффективная экспериментальная база. На материале корпуса белорусских газет в ходе реализации проекта проведены исследования региональных отличий узуальных норм речи, употребления специальной лексики, фразеологии и прецедентных имен как отражения межкультурных и межъязыковых контактов, дана оценка корпуса как источника актуального

страноведческого материала [Рычкова 2014]. Возможности статистических методик были опробованы для исследования лексико-семантических особенностей районных и городских газет Гродненщины [Шайкевич, Савчук 2014].

### Литература

1. *Пиперски А. Ч.* (2013), Генеральный интернет-корпус русского языка и понятие репрезентативности в корпусной лингвистике // *Современные проблемы науки и образования*. № 5; URL: [www.science-education.ru/111-9895](http://www.science-education.ru/111-9895) (дата обращения: 23.03.2015)

2. *Рычкова Л. В.* (2014), Изучение социально-коммуникативных особенностей русского языка на основе использования лингвистических корпусов региональных средств массовой информации (на примере корпуса Гродненщины). *Universitas Catholica Rosenbergensis Studia Russico-Slovaca. – Verbum. – Ruzomberok*, 2014. – С. 48–58.

3. Словарь «Языки городов», форум «Городские диалекты» [электронный ресурс] available at: <http://community.lingvo.ru/goroda/>

4. *Савчук С. О.* (2015), Корпус региональных газет России и зарубежья. Труды института русского языка. М., 2015 (в печати)

5. *Шайкевич А. Я., Савчук С. О.* (2014), Анализ лексико-семантических особенностей региональной прессы (на примере газет Гродненского региона Беларуси). *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»* (Бекасово, 4 — 8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, с. 585-598.

### References

1. *Piperski A. Ch.* (2013), General'nyj internet-korpus russkogo yazyka i ponyatie reprezentativnosti v korpusnoj lingvistike [General



Internet-corpus of the Russian language and the concept of representativeness in corpus linguistics]. *Sovremennye problemy nauki i obrazovaniya* [Modern problems of science and education], 5. URL: [www.science-education.ru/111-9895](http://www.science-education.ru/111-9895) (date accessed: 23.03.2015)

2. *Rychkova L. V.* (2014), *Izuchenie social'no-kommunikativnyx osobennostej russkogo yazyka na osnove ispol'zovaniya lingvisticheskix korpusov regional'nyx sredstv massovoj informacii (na primere korpusa Grodnenshhiny)*. [The study of socio-communicative peculiarities of the Russian language through the use of regional mass media linguistic corpora] *Universitas Catholica Rosenbergensis Studia Russico-Slovaca, Verbum, Ruzomberok*, pp. 48–58.

3. *Slovar' «Yazyki gorodov»*, forum «Gorodskie dialekty» [Dictionary «Languages of cities», forum «Urban dialects» [electronic resource], available at: <http://community.lingvo.ru/goroda/>

4. *Savchuk S.O.* (2015), *Korpus regional'nyx gazet Rossii i zarubezh'ya*. [The corpus of regional Newspapers in Russia and abroad.] *Trudy instituta russkogo yazyka* [Proceedings of the Institute of the Russian language]. Moscow (in print).

5. *Shajkevich A. Ya., Savchuk S. O.* (2014), *Analiz leksiko-semanticheskix osobennostej regional'noj pressy (na primere gazet Grodnenskogo regiona Belarusi)*. [Distributional-statistical Analysis of Regional Press (Newspapers of Grodno Region of Belarus)] *Komp'yuternaya lingvistika i intellektual'nye texnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferencii «Dialog»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2014»]. Moscow, pp. 585-598.

**Svetlana Savchuk**

Vinogradov's Institute of the Russian Language RAS (Russia).

***E-mail: savsvetlana@mail.ru***