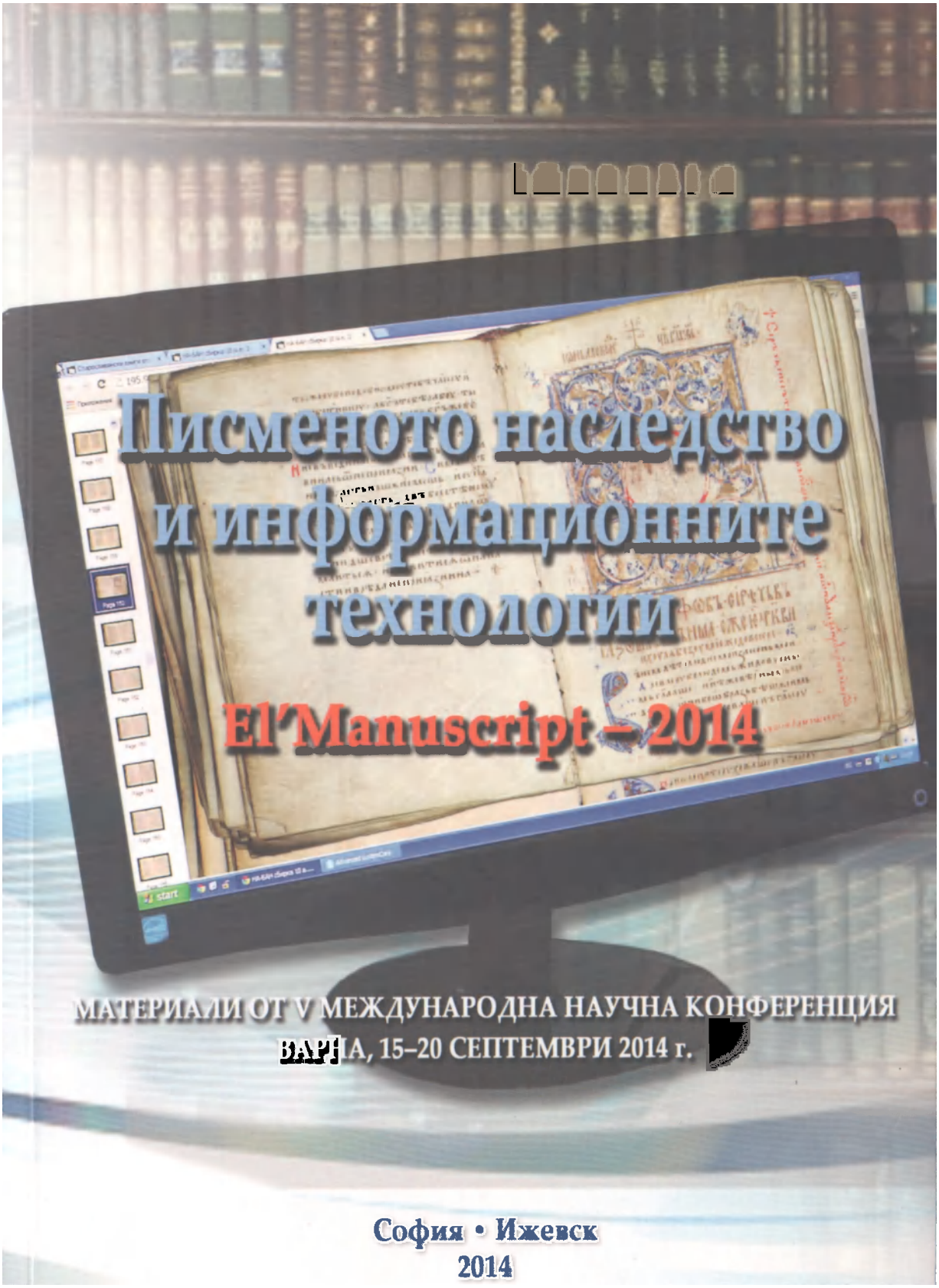

Писменото наследство и информационните технологии

El'Manuscript – 2014

МАТЕРИАЛИ ОТ V МЕЖДУНАРОДНА НАУЧНА КОНФЕРЕНЦИЯ

ВАРНА, 15–20 СЕПТЕМВРИ 2014 г.

София • Ижевск
2014

Специализированный лингвистический корпус социокультурной специфики регионального варианта русской речи[1]

Л. В. Рычкова

...й язык, региональный вариант русской речи, социокультурная специ... речи, лингвистический корпус, корпус СМИ

A Specialized Linguistic Corpus Created for Research on the Sociocultural Characteristics of a Regional Variant of Russian Speech

Ludmila Rychkova

...aper describes peculiarities of the development of a specialized linguistic ... comprising Russian-language texts from the mass media printed in the ... Region of Belarus, the goal of which is research on the sociocultural char... ...tics of one of the regional varieties of Russian speech.

...ход от исследования собственно лингвистических реалий к использова... ...ковых данных для выявления комплекса социокультурных факторов, ...ующих варьированию языка и обусловливающих принципиальную из... ...сть и вариативность функционирования его единиц, стал возможным ... формированию новой, экспериментально-доказательной, парадигмы ...тики, становлению которой, без сомнения, способствует лингвистика ..., в рамках которой не только создаются такие репрезентативные мас... ...ковых данных, как лингвистические корпусы текстов, но и разрабаты... ...ные методики и направления интердисциплинарных исследований, ос... ...на обобщении корпусных данных.

...ание лингвистических корпусов — это сложная и трудоемкая процедура, ...ажную роль играет грамотное определение этапов формирования кор... ...спечение аутентичности отобранного в качестве базы корпуса языково... ...риала, разработка системы метаразметки, в том числе — тематической. ...учет всех этих факторов позволяет обеспечить информационно... ...ческую релевантность корпуса, достоверность корпусных данных. Сте... ...мимости корпуса зависит также от использованных в нем видов лин... ...кой разметки. Автоматический режим разметки, как правило, дает мно... ...рмационного шума", обусловленного принципиальной неоднозначно...

...лено в рамках проекта, реализуемого при поддержке Белорусского республи... ...фонда фундаментальных исследований (договор № Г13Р-050 от 16.04.2013 г.).

...чкова, 2014

стью языковых объектов. В связи с этим первично размеченный корпус до ввода в эксплуатацию пройти этапы выверки и апробации.

В рамках международного белорусско-российского проекта создается гвистический корпус нового типа, представляющий собой оригинальный тронный языковой ресурс, специфика которого обусловлена, прежде всего, бенностями избранного в качестве основы для создания данного корпуса вого материала.

В частности, важной особенностью создаваемого корпуса, которая про в процессе сбора и систематизации электронного контента, является сто шанный" характер: в корпусе сохраняются и специальным образом разме белорусскоязычные фрагменты и тексты, отражающие естественное в усл близкородственного белорусско-русского двуязычия сознательное перекл кодов, достаточно часто осуществляемое в текстах СМИ Гроднснщины тим, что такое сознательное переключение кодов не имеет ничего общего словутой "трасянкой", представляющей собой результат неосознанного пси гического двуязычия смешанного типа). Все существующие до сих пор дву ные корпусы создавались либо как параллельные, либо как сопоставимые образом, создаваемый корпус представляет собой первый опыт "смеш корпуса, отражающего соответствующий языковой материал, что создаст тивную базу для проведения межъязыковых исследований.

Второй важной особенностью создаваемого корпуса является его "рег ный" характер, так как он строится на материале текстов СМИ Гродненщ полиэтничного региона Беларуси, непосредственно граничащего с Поль Литвой. Как отмечают российские участники проекта — непосредственны работчики Национального корпуса русского языка (НКРЯ) [Национальный корпус русскоязычных СМИ Гродненщины положит начало формиров составе НКРЯ нового модуля, представляющего региональные варианты турного стандарта" [Кустова, Савчук 2013: 352].

Поскольку "новая языковая ситуация в государствах постсоветского странства приводит к усилению процессов дивергенции региональных ва русского языка и кодифицированного языка метрополии, к постепенно мированию нациолектов" [Кустова, Савчук 2013: 345], то создание такого является объективным отражением "генерального" характера НКРЯ, ст гося отразить все функциональные варианты русского языка.

Выбор в качестве базы корпуса текстов СМИ не случаен. Именно С более "чувствительны" ко всем факторам социо- и лингвоэкологии и ярко отражают различные виды актуального дискурса, и поэтому позвол лучать данные, релевантные для выявления различного рода социоку специфики национально или регионально обусловленных идиомов. СМ занимает сегодня главенствующее положение в системе этнической и ческой коммуникации, оказывая влияние на формирование стереотипов ного речевого поведения и социокультурных конструктов. В текстах СМ

[illegible]рует не только общеупотребительная лексика, но и лексика специальная, в [illegible]ти, именно тексты СМИ отражают формирование языков для социокуль[illegible]рных целей. Динамичность языка СМИ позволяет выявлять семантический [illegible]ал лексических единиц и специфику его реализации в различных типах [illegible]обусловленную аксиологическими, гендерными и иными социокультур[illegible]факторами.

[illegible]постративный" характер корпуса текстов СМИ Гродненщины, обуслов[illegible]вынужденным, по причине высокой степени трудоемкости создания кор[illegible]ограничением объема исходного языкового материала, актуализировал [illegible]му отбора конкретных изданий, тексты которых составили базу для фор[illegible]ния "сырого" (неразмеченного) электронного языкового ресурса на пер[illegible]е создания корпуса.

[illegible]ности, была осуществлена первичная характеристика СМИ с учетом [illegible]ших критериев: название газеты, сайт, место издания, официаль[illegible]официальность, год основания, периодичность, язык статей, адре[illegible]тель, тематика. Как важные факторы в отборе СМИ для формирования [illegible]лингвистического корпуса рассматривались следующие: наличие интер[illegible]ий СМИ, их доступность и возможность конвертации в формат корпуса. [illegible]образом, был выполнен мониторинг интернет-версий русскоязычных [illegible]родненщины, в результате чего был определен характер существующих [illegible]ших архивов различных СМИ, их объем, пригодность для формальной и [illegible]тельной обработки для целей создания лингвистического корпуса.

[illegible]кольку описываемый в данной работе корпус создается как эксперимен[illegible]показательная база для исследования социокультурной специфики рус[illegible]чи Гродненщины, то особое внимание было уделено разработке системы [illegible]метки, позволяющей должным образом структурировать исходный язы[illegible]материал в соответствии с релевантными цели исследования частными [illegible]ческими задачами.

[illegible]новимся подробнее на двух параметрах метаразметки иллюстративного [illegible]текстов СМИ Гродненщины — "Сфера функционирования" и "Темати-

[illegible]метр "Сфера функционирования" в массиве текстов корпуса представлен [illegible]шими значениями: "бытовая", "официально-деловая", "производственно-[illegible]ская", "публицистика", "реклама", "учебно-научная", "художественная", [illegible]но-богословская", "электронная коммуникация".

[illegible]первый взгляд, набор значений не слишком велик, но в любом случае, он [illegible]т фактическое стилевое разнообразие текстов современных СМИ, выхо[illegible]ы пределы собственно публицистики.

[illegible]одним важным параметром системы метаразметки является тематика [illegible]СМИ. В газетном корпусе НКРЯ тематическая разметка не предусмотре[illegible]не менее, в основном корпусе НКРЯ все нехудожественные тексты раз[illegible]ся в соответствии с тематикой.

Определение тематики текста при осуществлении разметки вызывает от ленные сложности, так как большие по объему тексты чаще всего, как пр информационно неоднородны, то есть в них представлено несколько тем. скольку при осуществлении тематической разметки текстов нехудожест литературы основного корпуса НКРЯ в случае неоднозначности дается индексов, соответствующих всем предметным областям, нашедшим отраж тексте, то аналогичным образом осуществлялась тематическая разметка и и при создании иллюстративного корпуса русскоязычных СМИ Гродне Для определения перечня адекватных фактическому языковому материалу тических индексов в автоматическом режиме были получены списки тема ских маркеров архивов газет — источников исходных для корпуса текстов в ручном режиме была составлена таблица соответствий полученных таким зом тематических маркеров значениям параметра "тематика", предлож разработчиками НКРЯ для текстов нехудожественной литературы осн корпуса. Анализ полученных таким образом данных показал, что, с изв долей допущения, этих индексов достаточно для тематической метара текстов корпуса СМИ Гродненщины. В итоге было добавлено лишь одно ние параметра — "происшествие".

Принципиальное соответствие значений параметров метаразметки илл тивного корпуса текстов СМИ Гродненщины характеристикам основного са письменных текстов НКРЯ — важный фактор, обеспечивающий "вклю этого корпуса в новый модуль НКРЯ — корпус региональной и зар прессы.

Литература

Кустова, Савчук 2013 — *Кустова Г. И., Савчук С. О.* Изучение лексико-семан кой и социокультурной специфики русской речи на территории Республики русь (на материале текстов СМИ) // Труды междунар. конф. "Корпусная ли тика–2013", Санкт-Петербург, 25–27 июня 2013 г. СПб.: С.-Петербургск. го Филол. фак-т, 2013. С. 344–352.

Национальный — *Национальный корпус русского языка* [Электронный ресурс жим доступа: http://www.ruscorpora.ru, свободный (дата обращения: 17.02.20