МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

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

Л. В. Рычкова

Русский язык, региональный вариант русской речи, социокультурная специфика 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

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

Переход от исследования собственно лингвистических реалий к использованию языковых данных для выявления комплекса социокультурных факторов, способствующих варьированию языка и обусловливающих принципиальную изменчивость и вариативность функционирования его единиц, стал возможным благодаря формированию новой, экспериментально-доказательной, парадигмы лингвистики, становлению которой, без сомнения, способствует лингвистика корпусная, в рамках которой не только создаются такие репрезентативные массивы языковых данных, как лингвистические корпусы текстов, но и разрабатываются новые методики и направления интердисциплинарных исследований, основанные на обобщении корпусных данных.

Создание лингвистических корпусов — это сложная и трудоемкая процедура, поэтому важную роль играет грамотное определение этапов формирования корпуса, обеспечение аутентичности отобранного в качестве базы корпуса языкового материала, разработка системы метаразметки, в том числе — тематической. Только учет всех этих факторов позволяет обеспечить информационно-прагматическую релевантность корпуса, достоверность корпусных данных. Степень применимости корпуса зависит также от использованных в нем видов лингвистической разметки. Автоматический режим разметки, как правило, дает много “информационного шума”, обусловленного принципиальной неоднозначно-

[1] Подготовлено в рамках проекта, реализуемого при поддержке Белорусского республиканского фонда фундаментальных исследований (договор № Г13Р-050 от 16.04.2013 г.).

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

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

В частности, важной особенностью создаваемого корпуса, которая проявилась в процессе сбора и систематизации электронного контента, является его "смешанный" характер: в корпусе сохраняются и специальным образом размечаются белорусскоязычные фрагменты и тексты, отражающие естественное в условиях близкородственного белорусско-русского двуязычия сознательное переключение кодов, достаточно часто осуществляемое в текстах СМИ Гродненщины (отметим, что такое сознательное переключение кодов не имеет ничего общего с пресловутой "трасянкой", представляющей собой результат неосознанного психологического двуязычия смешанного типа). Все существующие до сих пор двуязычные корпусы создавались либо как параллельные, либо как сопоставимые. Таким образом, создаваемый корпус представляет собой первый опыт "смешанного" корпуса, отражающего соответствующий языковой материал, что создает объективную базу для проведения межъязыковых исследований.

Второй важной особенностью создаваемого корпуса является его "региональный" характер, так как он строится на материале текстов СМИ Гродненщины — полиэтничного региона Беларуси, непосредственно граничащего с Польшей и Литвой. Как отмечают российские участники проекта — непосредственные разработчики Национального корпуса русского языка (НКРЯ) [Национальный…], — корпус русскоязычных СМИ Гродненщины положит начало формированию "в составе НКРЯ нового модуля, представляющего региональные варианты литературного стандарта" [Кустова, Савчук 2013: 352].

Поскольку "новая языковая ситуация в государствах постсоветского пространства приводит к усилению процессов дивергенции региональных вариантов русского языка и кодифицированного языка метрополии, к постепенному формированию нациолектов" [Кустова, Савчук 2013: 345], то создание такого модуля является объективным отражением "генерального" характера НКРЯ, стремящегося отразить все функциональные варианты русского языка.

Выбор в качестве базы корпуса текстов СМИ не случаен. Именно СМИ наиболее "чувствительны" ко всем факторам социо- и лингвоэкологии и наиболее ярко отражают различные виды актуального дискурса, и поэтому позволяют получать данные, релевантные для выявления различного рода социокультурной специфики национально или регионально обусловленных идиомов. Сфера СМИ занимает сегодня главенствующее положение в системе этнической и межэтнической коммуникации, оказывая влияние на формирование стереотипов публичного речевого поведения и социокультурных конструктов. В текстах СМИ функ-

ционирует не только общеупотребительная лексика, но и лексика специальная, в частности, именно тексты СМИ отражают формирование языков для социокультурных целей. Динамичность языка СМИ позволяет выявлять семантический потенциал лексических единиц и специфику его реализации в различных типах узуса, обусловленную аксиологическими, гендерными и иными социокультурными факторами.

"Иллюстративный" характер корпуса текстов СМИ Гродненщины, обусловленный вынужденным, по причине высокой степени трудоемкости создания корпуса, ограничением объема исходного языкового материала, актуализировал проблему отбора конкретных изданий, тексты которых составили базу для формирования "сырого" (неразмеченного) электронного языкового ресурса на первом этапе создания корпуса.

В частности, была осуществлена первичная характеристика СМИ с учетом следующих критериев: название газеты, сайт, место издания, официальность/неофициальность, год основания, периодичность, язык статей, адресат/читатель, тематика. Как важные факторы в отборе СМИ для формирования основы лингвистического корпуса рассматривались следующие: наличие интернет-версий СМИ, их доступность и возможность конвертации в формат корпуса. Таким образом, был выполнен мониторинг интернет-версий русскоязычных СМИ Гродненщины, в результате чего был определен характер существующих электронных архивов различных СМИ, их объем, пригодность для формальной и содержательной обработки для целей создания лингвистического корпуса.

Поскольку описываемый в данной работе корпус создается как экспериментально-доказательная база для исследования социокультурной специфики русской речи Гродненщины, то особое внимание было уделено разработке системы метаразметки, позволяющей должным образом структурировать исходный языковой материал в соответствии с релевантными цели исследования частными прагматическими задачами.

Остановимся подробнее на двух параметрах метаразметки иллюстративного корпуса текстов СМИ Гродненщины — "Сфера функционирования" и "Тематика".

Параметр "Сфера функционирования" в массиве текстов корпуса представлен следующими значениями: "бытовая", "официально-деловая", "производственно-техническая", "публицистика", "реклама", "учебно-научная", "художественная", "церковно-богословская", "электронная коммуникация".

На первый взгляд, набор значений не слишком велик, но в любом случае, он отражает фактическое стилевое разнообразие текстов современных СМИ, выходящих за пределы собственно публицистики.

Еще одним важным параметром системы метаразметки является тематика текстов СМИ. В газетном корпусе НКРЯ тематическая разметка не предусмотрена. Тем не менее, в основном корпусе НКРЯ все нехудожественные тексты размечаются в соответствии с тематикой.

Определение тематики текста при осуществлении разметки вызывает определенные сложности, так как большие по объему тексты чаще всего, как правило, информационно неоднородны, то есть в них представлено несколько тем. Поскольку при осуществлении тематической разметки текстов нехудожественной литературы основного корпуса НКРЯ в случае неоднозначности дается сумма индексов, соответствующих всем предметным областям, нашедшим отражение в тексте, то аналогичным образом осуществлялась тематическая разметка текстов и при создании иллюстративного корпуса русскоязычных СМИ Гродненщины. Для определения перечня адекватных фактическому языковому материалу тематических индексов в автоматическом режиме были получены списки тематических маркеров архивов газет — источников исходных для корпуса текстов. Затем в ручном режиме была составлена таблица соответствий полученных таким образом тематических маркеров значениям параметра "тематика", предложенным разработчиками НКРЯ для текстов нехудожественной литературы основного корпуса. Анализ полученных таким образом данных показал, что, с известной долей допущения, этих индексов достаточно для тематической метаразметки текстов корпуса СМИ Гродненщины. В итоге было добавлено лишь одно значение параметра — "происшествие".

Принципиальное соответствие значений параметров метаразметки иллюстративного корпуса текстов СМИ Гродненщины характеристикам основного корпуса письменных текстов НКРЯ — важный фактор, обеспечивающий "включение" этого корпуса в новый модуль НКРЯ — корпус региональной и зарубежной прессы.

Литература

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

Национальный — *Национальный корпус русского языка* [Электронный ресурс]. Режим доступа: http://www.ruscorpora.ru, свободный (дата обращения: 17.02.2014).