Российская академия наук
Институт русского языка им. В. В. Виноградова

Национальный корпус русского языка: 2006–2008

Новые результаты и перспективы

Санкт-Петербург
«НЕСТОР–ИСТОРИЯ»
2009

Содержание

III. Новые проекты в рамках НКРЯ

IV. Семантика в НКРЯ

V. Техническое обеспечение корпусных задач

VI. Перспективы использования НКРЯ в различных областях

Образование

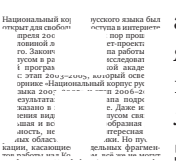
Наука

VII. НКРЯ и другие корпуса

Е. В. Рахилина

Корпус как творческий проект

Введение



Национальный корпус русского языка был открыт для свободного доступа в интернете 29 апреля 2004 года — с тех пор прошло 5 с половиной лет, для интернет-проекта это много. Закончились два этапа работы над корпусом в рамках особой исследовательской программы Российской академии наук: этап 2003–2005, который освещен в сборнике «Национальный корпус русского языка 2003–2005» и этап 2006–2008. О результатах второго этапа подробно рассказано в этом сборнике. Даже из оглавления видно, что с Корпусом связана большая и всё более разнообразная деятельность, несомненно, интересная для разных областей лингвистики. Но публикации, касающиеся отдельных фрагментов работы над Корпусом, всё же не могут дать представления о проекте в целом, его развитии, общих задачах и перспективах, его, если можно так сказать, «философии». Восполнить этот пробел мы и попробуем в настоящей статье.

Прежде всего, напомним, что первый этап работы был нацелен на создание корпуса как такового: нужно было собрать как можно больше текстов, сделать корпус представительным и организовать по имеющимся текстам хотя бы самый простой поиск. Все усилия разработчиков были направлены именно на это. Имелось в виду, что главной задачей является «канонический» сбалансированный стомиллионный корпус современного русского языка, хронологические границы которого задавались периодом с 50-х годов XX века по настоящее время. Дополнительно предполагался корпус XIX и первой половины XX века в качестве, так сказать, диахронической составляющей. Все другие разработки, касающиеся диалектного корпуса, корпуса устных текстов, параллельного корпуса и проч. на первом этапе представлялись как экспериментальные, они создавали задел на будущее. Сами эти корпуса в то время либо отсутствовали, либо были очень малы, но активно обсуждались принципы их формирования, их структура, поисковые возможности и т.п. Кроме того, в рамках НКРЯ развивались еще два самостоятельных больших корпусных проекта: корпус XI–XIV вв. и синтаксически размеченный корпус современного русского языка. Работа над первым частично отражена в статье А. И. Зобнина и А. В. Сахаровой в настоящем сборнике; о втором проекте можно прочитать в [Апресян и др. 2005], а воспользоваться этим подкорпусом и изучить принятую в нем систему разметки можно теперь непосредственно на сайте НКРЯ (*http://ruscorpora.ru/search-syntax.html*).

Задачи первого этапа удалось выполнить почти все; собственно, тогда сил не хватило только на систематический сбор текстов первой половины XX века, поэтому данная часть работы завершается только сейчас. В остальном, к 2005 году Национальный корпус русского языка действительно существовал в довольно солидном объеме: 100 млн словоупотреблений, как и планировалось, для современного русского языка и более 20 млн словоупотреблений — для (в основном художественных) текстов XIX века. На этих текстовых массивах уже тогда работал морфологический анализ и пилотный проект семантической разметки. Кроме того, был создан значительный по объему (более 4 млн словоупотреблений) корпус со снятой вручную грамматической омонимией, который давал возможность высокоточной выдачи результатов по запросам,

Корпуса славянских языков в Интернете: основные параметры

язык	корпус	содержание	объем корпуса (в млн. словоупотреблений)	Типы разметки		
				морфологическая	снятие грамматической омонимии (**а**втоматическое/ **р**учное)	синтаксическая
чешский	ЧНК — подкорпуса письменного языка	коллекция сбалансированных и специализированных корпусов (1990–2004)	500	+	а (весь корпус)/р (0,08)**	–
	ЧНК — подкорпуса устной речи	записи устной речи из разных регионов Чехии	2,3	–		–
	PDT	газеты и журналы (1990–95)	2	+	р	+ (1,5)**
словацкий	СНК	письменные тексты разных типов (1955–2006)	339	+	а (весь корпус)/р (0,5)**	–
польский	IPI PAN	несбалансированная коллекция текстов нескольких типов	250	+	а (сохраняются все варианты разбора)	–
	PELCRA	письменные и устные тексты разных типов (1989–2003)	93	–		–
	PWN	фрагменты письменных текстов различных типов, устная речь (1903–2005)	22/ 3,7*	только лемматизация	–	–

Типы разметки		Поисковые возможности: поиск по:						Параметры выдачи				
семантическая	метаразметка	словоформе	лексеме	последовательности словоформ	грамматическим признакам	синтаксическим структурам	семантическим признакам	максимальный контекст	ограничения на количество контекстов	сортировка выдачи	фильтрация выдачи (поиск в найденном)	статистическая обработка запроса
–	+	+	+	+	+	–	–	≈1000 знаков/ 100 слов/ 3 предл.	нет	+	+	+
–	+	+	–	+	–	–	–		нет	+	+	+
+ (0,8)**	–	+	+	+	+	+ (1,5)**	+ (0,8)**	1 предл.	нет	–	+	–
–	+	+	+	+	+	–	–	≈200 слов	нет	+	+	+
–	+	+	+	+	+	–	–	200 слов	нет	+	–	–
–	+	+	–	+	–	–	–	3 абзаца	250	+	–	+
–	±	+	+	+	–	–	–	не ограничен	нет	+	–	–