Международная конференция «Slavicorp»

22–23 ноября 2010 года в Варшаве прошла конференция «Slavicorp», собравшая специалистов, занимающихся реализацией корпусных проектов, в той или иной степени включающих славянский языковой материал. Это одна из первых конференций по славянским корпусам, довольно ясно обозначившая современные тенденции в корпусной славистике. Естественно, что большинство участников представляло славянские страны, но заметную роль играли также и корпусные лингвисты из Норвегии и Швейцарии. За два дня работы можно было познакомиться с пленарным докладом и принять участие в работе двух параллельных секций. Заключительное заседание было совмещено с рабочим совещанием корпусной комиссии при Международном комитете славистов.

Пленарное выступление Ф. Чермака (Чехия) вписывало масштабный чешский корпус InterCorp в общую панораму современного состояния корпусной лингвистики. Во-первых, не случайно именно материал о параллельном корпусе стал темой главного доклада конференции, речь об этом типе корпусных инструментов в Варшаве заходила неоднократно. В той или иной степени почти все положения доклада вызывали живое сочувствие или полемический отклик. В частности, уже ближе к концу выступления докладчик обозначил

свою позицию по довольно болезненному вопросу авторских прав, с которым в явном или неявном виде неизбежно сталкивается любой создатель современного корпуса. Ф. Чермак постарался сгладить эту проблему, представив ее наименее острые стороны.

Дальнейшая работа конференции в первый день являла собой своего рода парад больших и малых славянских монолингвальных корпусов, которые чередовались с находящимися на том или ином этапе реализации проектами параллельных корпусов. Последние, как можно судить по активности в этой области (по крайней мере, в среде славянских корпусных лингвистов), особенно востребованы и в ближайшее время будут приоритетным направлением корпусостроения. В.А. Плунгян (Россия) представил «Национальный корпус русского языка», который уже довольно хорошо известен и является своего рода авторитетом (но все же пока еще не образцом) в корпусной сфере. Русский корпус упоминался в дискуссиях к докладам конференции как пример необычной для корпусной практики простоты и дружественности интерфейса. Почти одновременно были представлены два больших и тоже по-своему авторитетных хорватских корпуса: в докладах М. Тадич (Хорватия) «Хорватский национальный корпус» и Д. Брозович, Д. Чавар (Хорватия) «Ризница – хорватский языковой корпус». Объем корпусов, тщательность разметки, внимание к жанровой сбалансированности текстов в обоих корпусах свидетельствуют о высоком уровне корпусной лингвистики в Хорватии, что, увы, не всегда можно сказать о других славянских странах. А. Грён н, Ш. Ро Хауге, Е. Хачатурян, Л. Шарич (Норвегия) представили путь развития уже существующего параллельного русско-норвежского корпуса RuN, в который добавляются другие европейские (главным образом, славянские) языки: «От RuN к Run-Euro: мультилингвальный параллельный корпус университета Осло». Кроме собственно презентации корпуса, в докладе внимание было уделено идее создания единого банка стандартизированных текстов, которые при необходимости создатели корпусов могли бы включать в свои проекты, затрачивая при этом минимум усилий по обработке. Вопросы стандартизации тоже стали одним из ключевых мотивов конференции и еще не раз будут обсуждаться в приложении к системам разметки, интерфейса и т. д.

На фоне таких масштабных докладов несколько выделялось выступление И. Косема (Словения), в котором поднимался вопрос о дружественном к пользователю интерфейсе корпусных инструментов. Опыт словенских разработчиков подсказал стандарты в этой области, до которых, как обозначила дискуссия после доклада, современным корпусам (может быть, за исключением Национального корпуса русского языка) еще очень далеко. Стандарты эти задаются «большими» поисковыми системами Интернета, в частности, лидером интернет-поиска Google. Простота, очевидность результата и, напротив, неочевидные для незаинтересованного взгляда технические детали работы системы – это те параметры, которыми пока что может похвастаться далеко не каждый корпус.

Своим значительным техническим и лингвистическим уровнем впечатляет Болгарский национальный корпус. Его представили в своем докладе Д. Благоева, С. Колковска, Ц. Григорисва и Н. Костова (Болгария), специально указав, что сейчас корпус является важной основой для лексикографической деятельности.

Гораздо более скромно на этом фоне выглядят достижения сербских ученых. Д. Витас и М. Утвич (Сербия) описали сравнительно небольшой (25 миллионов слов), морфологически не размеченный и, что тоже важно, недоступный для свободного использования Корпус современного сербского языка, которым, однако, пользуется около 250 славистов по всему миру.

Заключительное секционное заседание первого дня работы конференции в значительной степени вновь было посвящено параллельным корпусам. Д.В. Сичинава (Россия) рассказал о параллельных корпусах в составе Национального корпуса русского языка – как об уже доступных, так и о планируемых, при этом очевидно, что планы эти в первую очередь связаны именно со славянскими параллельными корпусами. М. Лазинский (Польша), М. Куратчек (Польша), Б.В. Орехов (Россия / Норвегия) и Е.А. Слободян (Россия) представили демонстрационную версию русско-польского параллельного корпуса, который стал результатом сотрудничества больших польского и русского корпусных проектов. Когда работа будет завершена, корпус будет доступен одновременно на сайте НКРЯ и на сервере Варшавского университета. Л. Грабовски (Польша) выступил с докладом об использовании корпуса при составлении «Польско-русского словаря устойчивых выражений», попутно высказав рабочие соображения об удобстве интерфейса корпуса.

Второй день работы конференции прошел под знаком более частных аспектов построения корпусов, а также основанных на корпусном материале собственно лингвистических исследований. Так, Ш. Ро Хауге (Норвегия)

представил доклад об использовании параллельного корпуса в изучении дискурсивных элементов, а М. Копеч (Польша) рассказал о снятии омонимии (дезамбигуации) в Национальном корпусе польского языка. Хозяева конференции поделились опытом в еще нескольких выступлениях, с разных сторон представив Национальный корпус польского языка, в частности, К. Гловиньска прочитала доклад о синтаксической разметке в польском корпусе.

Однако и во второй день состоялось несколько презентаций славянских корпусных проектов. Н. Дарчук (Украина) рассказала о Корпусе украинского языка, а Р. фон Вальденфельс (Швейцария) представил довольно масштабный по охвату материала параллельный корпус славянских (и некоторых других) языков ParaSol, включающий тексты на 20 языках общим объемом в 16 миллионов слов. При этом структура корпуса довольно сложна и не каждый текст на одном языке имеет пару на каждом другом, заявленном в корпусе. Однако количество таких парных соответствий уже тоже довольно внушительно: 167.

В заключении конференции Н. Коцыбой (Польша) был сделан обзор польско-украинского параллельного корпуса, работа над которым все еще далека от завершения, и инструмент пока не доступен для использования в Интернете.

На заседании корпусной комиссии при Международном комитете славистов обсуждались организационные вопросы о составе и порядке работы этой структуры. Было принято предварительное решение о приглашении в члены самых авторитетных специалистов в данной области и о том, что до выборов, которые пройдут на Международном съезде славистов в Минске, исполняющим обязанности председателя будет Марек Лазинский (Польша).

Б.В. Орехов

Сведения об авторе:
Борис Валерьевич Орехов
Университет Осло /
Башкирский гос. пед. университет
им. М. Акмуллы
boris.orekhov@ilos.uio.no