Современное состояние и перспективы развития электронных источников для исследования болгарской разговорной речи

Йовка Тишева, Марина Джонова

При представлении современного состояния и перспектив электронных источников для исследования болгарской разговорной речи (РР) важно уточнить два ключевых понятия - корпус и разговорная речь. Под *разговорной речью* понимается спонтанная, неподготвленная речь, сфера употребления которой не ограничена. А понятие *корпус* здесь употребляется в значении собрания, коллекции текстов (транскрибированные аудиозаписи спонтанной речи), которые организованы в зависимости от сферы и вида общения. Корпус должен быть представлен в форме, подходящей для э*лектронной обработки*, должен быть *представительным* (т.е. должен охватывать все сферы, в которых спонтанная речь используется как средство общения) и быть *доступным*, т.е. быть опубликованным в Интернете и быть доступным в исследовательских и образовательных целях.

Создание корпуса болгарской РР - долгосрочный проект, цель которого - исследование современного состояния болгарского языка, отраженного в РР. РР – более динамическая система по сравненинию с письменной, быстрее реагирует на социально-экономические перемены, поэтому дает ясное представление о современных развойных тенденциях в языке. РР отличается от литературно-письменной речи по своим фонетическим, лексическим, морфологическим и синтактическим особенностям. Их изучение может быть осуществлено единственно на основе достаточно большого объема спонтанной РР.

Сбор материала из разных сфер социальной интеракции в различных коммуникативных ситуациях (например, бизнес-встречи, отношения учитель – ученик в классе, врач – пациент, отношения в судебном зале и т.д.), будут представлены различные коммуникативные стили в рамках болгарской разговорной речи. Исследование их функций может быть использовано при изучении народопсихологических и культурных особенностей общения в болгарской среде в сопоставлении с другими странами.

При сборе и исследовании разговорной речи (в отличие от письменной речи) исключительное значение имеют обстановка, время, место общения, как и социальные и психологические особенности участнков в эксперименте. Учитываются некоторые характеристики исследуемых лиц, как возраст, пол, социальный статус, место рождения, образование и профессия – таким образом становится возможным проследить динамику языкового поведения. Изучение этих факторов позволит исследовать и объяснить “отклонения” от литературной нормы, которые наблюдаются в разговорной речи. Эти отклонения - необходимая основа проверки литературной нормы.

В этом смысле проект имеет бесспорный вклад в изучение динамических процессов в гражданском обществе (а следовательно, и в мышлении) и перемен, отраженных в речи. Модели речевого поведения будут интерпретированны как экспликацию социального поведения.

Наличие представительного для болгарской разговорной речи корпуса позволит исследовать некоторые особенности языкого общения, которые не отражаются в письменной речи. Например, можно исследовать различия между употреблением разговорной речи в официальной сфере и в неофициальных

контактах; наблюдать отношение между социальной ролью и адекватностью речевого поведения; описать национальный речевой этикет.

В процессе обработки разговорной речи возникает и много дополнительных проблем по сравнению с обработкой корпусов письменной литературной речи, например, наличието пауз, прерванных реплик и перебивания между говорящими, роль интонации и паралингвистики в осуществлении коммуникативных намерений говорящего. Необходимо изучение и систематизирование этих особенностей с целью создания программы для электронной обработки корпуса болгарской РР.

Материал, собранный в корпусе РР, может быть использован на пользу обучения иностранному языку, может дать возможность через знакомство с феноменами болгарской разговорной речи, а также и в обучении высококвалифицированных переводчиков.

На базе собранного корпуса впервые в высшем образовании будет возможность ввести обучение по корпусной лингвистике на материале болгарской разговорной речи. Участие студентов и докторантов в проекте по созданию корпуса позволит им применить на практике знания, полученные в высшем образованиии, что поможет им познакомиться с новыми информационными технологиями, которые могут применяться при сборе и обработке корпуса.

Что до сих пор сделано на Факультете славянских филологий? Основа разработки проекта – уже существующие корпуса затранскрибированной РР, состоявшиеся на Факультете спецкурсы по корпусной лингвистике и два лекционных курса в рамках магистрской программы по лингвистике, связанных с проблемами РР: *Вербальные коммуникативные стратегии* с руководителем доц. д-р Радка Влахова и *Проблемы болгарской разговорной речи* с руководителем доц. д-р Й. Тишева.

Во время своего обучения в бакалаврских программах студенты-филологи знакомятся с особенностями литературного болгарского языка, с нормативной системой письменных текстов. Кроме этой формы общенационального языка они знают и общие особенности болгарских диалектов. В сторону от их обучения остается болгарская РР. В этих целях в рамки магистрской программы по лингвистике на ФСФ был введен лекционный курс *Проблемы болгарской разговорной речи* с руководителем доц. д-р Й. Тишева с основной целью познакомить студентов с проявлениями этой формы болгарского языка. Описание РР невозможно с помощью методов традиционной грамматики, поэтому студенты должны знать прагматические и социолингвистические подходы в языковедческих исследованиях. Основная практическая цель курса – выработка студенческого корпуса разговорных текстов, который будет обработан и аннотирован самими студентами. Таким образом, они сами смогут идентифицировать проявления разговорного и подвижность литературно-нормативного в реальном устном общении.

На факультете славянских филологий Софийского университета до сих пор проведены два курса по компютерной корпусной лингвистике:

Первый из них *Анализ мультимодального корпуса разговорной речи* с лекторами Биляна Мартиновски и Лейф Грьонквист из Гетеборгского университета (Швеция) состоялся в 2000 г. в рамках болгарско-шведского проекта при финансовой поддержке фонда “Открытое общество” (“Multimedia and Research on Multimodal Social Interaction - STAGE 1” – все материалы курса – лекции, транскрипции, сделанные участниками, инструкции работы с

программами доступны на адресе http://www.ling.gu.se/~leifg/sofia/). В рамках этого проекта была проведена и рабочая встреча на тему: *Основные принципы сбора, транскрибирования и кодирования стандартизированного мультимодального языка корпуса*. На этом курсе были представлены постижения шведских ученых в области корпусной лингвистики и были отмечены особенности болгарской РР, которые требуют специального аннотирования и кодирования в связи с использованием программ электронной обработки корпуса.

Второй спецкурс *Компьютерная корпусная лингвистика* с руководителями гл.ас. д-р Кр. Алексова, ас. д-р П. Осенова (СУ "Св. Климент Охридски") и К. Симов (Лаборатория лингвистеского моделирования, БАН; руководитель проекта BulTreeBank) читался в летнем семестре учебного 2000/2001 г., и закончился сбором и транскрибированием участниками итого 150 минут РР. Этот материал будет опубликован на интернет-адресе http://www.hf.uio.no/east/bulg/mat и будет обогащен результатами аналогичного спецкурса в летнем семестре учебного 2001/2002 г., и материалы, собранные студентами магистрской программы по лингвистике в связи с их курсовыми и магистрскими работами.

В рамках спецкурса *Компьютерная корпусная лингвистика* студенты знакомятся с уже существующими корпусами текстов болгарской литературы, административных документов и прессы, а также и с существующими уже корпусами болгарской разговорной речи: корпус Цветанки Николовой (создан на основе аудиозаписей разговорной речи 70-х годов и послужил основой Частотного словаря болгарской РР) и корпус Красимиры Алексовой (созданный на базе аудиозаписей разговорной речи, сделанных в периоде 1989-1993 гг.). Эти корпуса опубликован в Интернете на адресе http://www.hf.uio.no/east/bulg/mat. И хотя они не аннотированны, они вызывают интерес множества лингвистов, которые обращаются к их составителям и используют данные корпусов в своих научных целях. Эти корпуса болгарской разговорной речи содержат около 160 000 словоформ, которые, однако, недостаточны для представительной выборки культурной ситуации, в которой развивается современная болгарская разговорная речь, и не включают эмпирического материала всех сфер социального взаимодействия (записана преимущественно разговорная речь в неофициальной сфере при общении между близкими и друзьями).

Будущей целью проекта очерчивается создание корпуса болгарской разговорной речи, который: 1) должен быть *представительным*, т.е охватывать все сферы употребления разговорной речи и иметь достаточно большой объем; 2) обязательной его характеристикой будет использование *единой транскрипции* во всех текстах; 3) будет *аннотированным и кодированным* в виде, подходящем для электронной обработки. Корпус будет опубликован в Интернете для нужд исследователей и преподавателей. Нераздельной частью корпуса будет *дигитализированный аудиоматериал*, который затранскрибирован в корпусе. Это поволит исследователям проследить итонационные особенности высказывания и характеристики диалогического общения (одновременное говорение обоих собеседников, прерывания, паузы и т.д.).

Публикация корпуса в Интернете позволит более широкому кругу исследователей и преподавателей в Болгарии и за рубежом использовать собранный материал в преподавании болгарского языка как иностранного и для

верифицирования лингвистических исследований, так как очень часто из-за нехватки представительного для всех сфер корпуса употребления болгарской РР, делаются осследования, которые не всегда доказаны эмпирическим материалом.

В создании корпуса участвуют специалисты из области лингвистики и из области иформатики и обработки естественного языка (NLP - Natural Language Processing). В будущей работе по собранному корпусу будут использованы постижения в области морфологии, синтаксиса и прагматики, методы корпусной лингвистики, а также и статистические методы, которые применяются при электронной обработке корпусов.

Создание корпуса болгарской разговорной речи и архива дигитализированных аудиоматериалов даст возможность болгарским ученым проводить сопоставительные межъязыковые и межкультурные исследования, что улучшает межкультурную коммуникацию и поднимает престиж болгарской лингвистики.

Корпус болгарской разговорной речи будет типа *monitor corpus*, т.е. не будет включать конечный набор текстов, а будет пополняться периодично. Это предопределяется особенностями РР – а именно, ее динамичность.

В более долгосрочном плане предвидится и сбор видеозаписей вербальной коммуникации и создание мультимедийного корпуса болгарского языка.