УДК 811.112.2’28 (571.150)

МЕТОДЫ И ПОДХОДЫ КОРПУСНОЙ ЛИНГВИСТИКИ В ИССЛЕДОВАНИЯХ СЕМАНТИКИ ДИАЛЕКТНОЙ ЛЕКСИКИ

Москвина Т.Н.

ФГБОУ ВПО «Алтайская государственная педагогическая академия», Барнаул, Россия (656031, г. Барнаул, ул. Крупской, 108, Лингвистический институт), e-mail: moskvina@uni-altai.ru

В данной статье рассматриваются основные подходы к изучению семантики языковых единиц островных немецких говоров с привлечением методов корпусной лингвистики. Языковая система островных немецких говоров характеризуется значительной вариативностью на всех уровнях: фонетическом, морфологическом, лексико-семантическом, синтаксическом. Изучение спектра значений в синхронии и диахронии возможно лишь при привлечении большого корпуса диалектных текстов. Многие европейские корпусы разговорной и диалектной речи используют систему EXMARaLDA. Диалектный корпус представляет собой специфический массив данных, поскольку диалект обладает системными отличиями от литературного языка и является исключительно устной формой коммуникации. Электронные корпусы диалектных текстов являются принципиально новым источником, способствующим приобщению диалектологии к современной научной лингвистической парадигме, в которой изучение основных языковых черт диалекта было бы автоматизировано, обеспечивало бы перекрестные исследования в текстах различных говоров, облегчало бы поиск и выборку необходимых данных и позволяло бы проводить диахронические исследования на примере нескольких десятилетий.

Ключевые слова: диалектология, островные немецкие говоры, языковая вариативность, корпусная лингвистика, лингвистический корпус.

METHODES AND APPROACHES OF THE CORPUS LINGUSITICS IN SEMANTIC RESEARCH OF INSULAR GERMAN DIALECTS

Moskvina T.N.

Altay state pedagogical academy, Barnaul, e-mail: moskvina@uni-altai.ru

In this article we consider the main approaches to the study of language units semantics in the island German dialects using the methods of corpus linguistics. The language system of the island German dialects is characterized by a considerable variability at all levels: phonetic, morphological, lexical-semantic and syntactic. The synchronic and diachronic study of the meanings spectrum is possible only by looking at a large number of dialect texts. Many European corpora of colloquial and dialect speech use the EXMARaLDA system. The dialect corpus represents a specific data array because a dialect possesses some systematically distinctive features different from the literary language and it represents a pure oral communication form. Computer-based corpora of dialect texts are a fundamental new source which promotes introduction of dialectology into the modern scientific linguistic paradigm where the study of the main linguistic features of a dialect would be automated, would provide cross researches in the texts of different dialects, would facilitate search and selection of the necessary data and would make it possible to conduct diachronic researches using the example of several decades.

Keywords: dialectology, insular German dialects, language variations, corpus-based linguistics, linguistic corpus.

Семантические исследования диалектной лексики предполагают работу в нескольких направлениях: синхронное описание лексического состава диалекта, изучение семантических новаций и вариантов с учетом внутренних законов развития и языковых контактов в диахроническом аспекте, а также сопоставительный анализ с материнским диалектом или группой родственных диалектов для определения путей семантического развития. Особое место в группе диалектов занимают островные говоры, под которыми традиционно понимаются разновидности языка, длительное время существующие в окружении другого

языка, территориальная и культурно-языковая изоляция которых привела к появлению дивергентных языковых признаков или сохранению архаических черт.

Изучение проходящих в диалекте процессов важно не только для понимания развития диалекта как одной из подсистем языка, но и для понимания динамики языковых процессов в национальном языке в целом. Немецкие говоры на территории Алтайского края характеризуются значительным разнообразием их лексико-семантической системы, которая является составной частью единой языковой системы немецкого языка, но содержит множество лексических единиц, отличающихся от стандарта и локально ограниченных в своем употреблении. Исследование всех случаев и контекстов употребления того или иного слова в различных островных говорах позволяет сделать вывод о стабильности или изменчивости конкретной языковой единицы. Изучение спектра значений в синхронии и диахронии возможно лишь при привлечении большого количества языкового материала, подтверждающего узуальность и конвенциональность определенного значения. Исследование семантики лексических единиц в диахроническом аспекте подразумевает прослеживание способов/контекстов использования той или иной единицы в различных коммуникативных ситуациях и контекстах. Таким образом, для получения объективных результатов необходим достаточный корпус языкового материала, собранного из множества различных источников в различное время.

Понятие корпуса является в лингвистике неоднозначным и даже многогранным. Так, «Словарь лингвистических терминов» дает следующие трактовки.

Корпус (массив, текст)

1. Примерная совокупность высказываний, отобранных для анализа и представленных в виде письменного текста, аудиозаписи и т.п.
2. Вся сумма (совокупность) произведений речи, созданных коллективом носителей данного языка [1, с. 209].

Такое классическое понимание лингвистического корпуса принципиально важно для диалектологических исследований, первым этапом которых всегда является запись речи носителей диалекта (как правило, аудио-и/или видеозапись с последующей письменной фиксацией в виде транскрипции или в орфографии).

Научно-исследовательской группой Лингвистического института Алтайской государственной педагогической академии под руководством проф. Л.И. Москалюк в течение нескольких десятилетий накоплен огромный языковой диалектный материал. Значительная часть аудиозаписей уже расшифрована и затранскрибирована, данный языковой материал уже частично лингвистически обработан и исследован.

Однако такой текстовый (в широком смысле) корпус представляет собой базу, но не инструмент исследования. Традиционно исследователь вручную проводил выборку отдельных явлений (лексических, грамматических, синтаксических и др.), исходя из целей и гипотез проводимого им исследования, и лишь потом проводил анализ полученного языкового материала. Такая технология исследования всегда является достаточно трудоемкой, требует много времени. Следует отметить и определенную долю субъективизма исследователя при отборе материала, при которой статистически возможны погрешности.

Но развитие науки и техники открывает новые возможности для исследователей, современные информационные технологии способствуют более быстрой и объективной обработке языковых данных. В лингвистических исследованиях все более широкое применение находят электронные ресурсы различных типов: электронные словари, базы данных, текстовые корпусы. Как отмечают создатели Саратовского диалектного корпуса русского языка, наличие электронных автоматически обрабатываемых лингвистических баз данных не только значительно ускоряет и оптимизирует трудоемкий процесс сбора языкового материала, но и ведет к смене научной парадигмы в лингвистике [4].

Эти задачи успешно решает корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Это позволяет в реальном времени получать результаты, требующие обработки таких массивов текстов, для получения и обработки которых ранее требовались месяцы. Корпус не просто позволяет ускорить исследования языка и многократно повысить их эффективность, достоверность и проверяемость – он позволяет решать такие задачи, которые лингвистика предыдущих эпох практически не ставила в силу их трудоемкости или невыполнимости. К таким задачам относятся, например, многие виды статистических и других квантитативных исследований языка. Корпусная лингвистика при этом не только измерительный и статистический инструмент, но и своеобразная «стратегия, методология исследования» [7, с. 19]. Примат объективных количественных данных, требование большого массива примеров, а также необходимость относительно широкой «географии» источников предполагает и совершенно иной методологический подход к решению задачи. Корпусная лингвистика исходит из того, что исследователь занимает, с одной стороны, позицию стороннего наблюдателя над языковыми явлениями, с другой стороны, произвольно задает параметры для выборки и анализа данных корпуса, т.е. корпусная лингвистика объединяет в себе теоретические и эмпирические принципы лингвистики.

В настоящее время существует множество определений понятия «лингвистический корпус». В качестве базового можно принять определение В.П. Захарова: «под

лингвистическим, или языковым, корпусом текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [3, с. 8].

Различают различные виды корпусов: иллюстративные, исследовательские, динамические, статистические. Исследовательский корпус предназначен преимущественно для изучения различных аспектов функционирования языковой системы. Этот тип корпусов данных, как правило, ориентирован на широкий класс лингвистических задач. В качестве основных требований, предъявляемых исследователями к подобного рода корпусу, выделяются репрезентативность, полнота, экономичность, самодостаточность, компьютерная поддержка, структуризация материала [2, с. 118-119].

Главная характеристика корпуса, отличающая его от простых коллекций текстов, заключается в наличии дополнительной информации о свойствах входящих в него текстов (разметки, или аннотации). Каждый текст должен иметь лингвистическую и экстралингвистическую разметку. В информацию о тексте необходимо включить сведения об информантах, о времени, месте записи, о конкретной ситуации общения, сведения о диалекте (говоре). Метатекстовая информация должна быть универсальной, типичной для лингвистических корпусов различного типа, чтобы не ограничивать параметры поиска, а, наоборот, сделать корпус доступным для многих исследователей с их различными целями, подходами и исходными гипотезами.

Самыми большими и полными являются корпусы национальных языков, например Национальный корпус русского языка, Брауновский корпус американского варианта английского языка, Британский национальный корпус и др. Во многих странах ведутся работы по созданию корпусов по разновидностям языка (корпус диалектов, устной или письменной речи, корпус смс-сообщений, детской речи, публицистических текстов и др. [6, с. 113-123]). Обширной информационной системой является корпус разговорного немецкого языка (Datenbank Gesprochenes Deutsch (DGD des DSAv)), разрабатываемый Институтом немецкого языка (г. Мангейм). Диалектный корпус является особым видом лингвистического корпуса, отличным от корпуса разговорной речи национального языка, поскольку диалектная языковая система по многим параметрам отличается от стандартной литературной и даже разговорной нормы (многочисленные фонетические варианты одной лексемы, уникальная, собственно диалектная лексика, не поддающаяся простому переводу на литературный язык, и др.). Создание диалектного электронного корпуса сопряжено с целым рядом сложностей:

- системные языковые отличия от литературного языка;

- исключительно устный характер диалектной коммуникации, как следствие – невозможность опереться на письменные источники;
- вариативность на всех уровнях, затрудняющая идентификацию единиц в корпусе;
- отсутствие единообразия при фиксации диалектной речи и различные способы организации информации.

Именно эти сложности и определяют еще незначительное количество диалектных и региональных корпусов как в русском, так и в немецком языковом пространстве. Работа над большинством корпусов еще не закончена. Технические и методологические проблемы во многом схожи. Так, разработчики Саратовского диалектного корпуса определяют необходимые параметры для четкой концепции корпуса. К числу таких параметров относятся, по их мнению, следующие:

1) принципы отбора диалектного материала и критерии репрезентативности диалектного корпуса;
2) принципы членения речевого континуума в корпусе;
3) параметры выдачи текстовых фрагментов;
4) формы представления диалектных текстов в корпусе;
5) виды и правила аннотирования текстовой базы корпуса;
6) параметры метаразметки диалектных текстов;
7) представление в диалектном корпусе нелингвистической информации;
8) оптимальные для диалектологических исследований возможности пользовательских запросов [4].

Остановимся кратко на каждом параметре. Для создания репрезентативного корпуса диалектных текстов необходимо большое количество лингвистически валидных и аутентичных записей и их транскрипций. При этом преимущество должно отдаваться записям реальной, не моделируемой исследователем коммуникации. Текст понимается максимально широко как любое речевое действие различной протяженности во времени. Учитывая наличие различных немецких говоров на территории Алтайского края, необходимо представить тексты всех диалектных областей. Все это обеспечивает объективность и надежность представленных лингвистических данных. В семантических исследованиях с помощью корпуса репрезентативность понимается не только количественно, но и качественно. Такой корпус должен покрывать большое количество тематических «проблемных областей». Под «проблемной областью» понимается «область реализаций языковой системы, содержащая феномены, подлежащие лингвистическому описанию» [2, с. 114]. Учитывая преимущественно бытовую (не профессиональную) и

семейную сферу употребления диалекта, необходимо включить в состав корпуса тексты различной тематики.

Как правило, диалектные текстовые корпусы значительно меньше по объему корпуса национального языка. Это обусловлено исключительно устной формой общения носителей диалекта, отсутствием письменных текстов на диалекте и ограниченностью тем личной и бытовой сферы общения. Кроме того, сложность лингвистической обработки таких текстов (расшифровка, разметка, аннотирование, семантический и структурный анализ) замедляет работу над пополнением корпуса и требует привлечения достаточного количества исследователей для его создания.

При создании корпуса и работе с ним наряду с репрезентативностью и полнотой данных методологически важен также параметр аутентичности и валидности текстов. Приоритет должен отдаваться записям естественной, спонтанной и неконтролируемой исследователем речи носителей диалекта. Однако даже сам факт присутствия наблюдателя, даже не участвующего в беседе, накладывает отпечаток на ход беседы. Методика сбора диалектного материала, как правило, не дает возможности получать по-настоящему естественный диалог, поскольку ситуация общения искусственно конструируется: эксплицитно задается тема коммуникации, участники диалога информированы о целях опроса и т.д. Такого рода тексты создатели корпусов диалектной речи относят к так называемым полуаутентичным, «провоцированным», контролируемым текстам (evozierte Daten: halbkontrollierte Texte (evokative Feldexperimente und aufgabenorientierte Kommunikation) [8]. Поэтому необходимы метаданные о характере протекания разговора и условиях записи. Это еще одно преимущество обработки диалектных текстов с помощью автоматического текстового корпуса, что позволяет дифференцировать полученные в процессе выборки и анализа результаты.

Наиболее надежной формой хранения диалектных текстов и оптимальной формой для проведения лингвистического анализа на примере большого массива данных является программно обеспеченный электронный текстовый корпус. Электронная форма представления диалектных текстов повышает сохранность этого уникального материала, создает возможность для более свободного доступа лингвистов различной специализации к первичному диалектному материалу, позволяющему анализировать различные явления в речи носителей немецких диалектов. Это определяет параметры 2-7, которые взаимосвязаны и их соблюдение возможно только при правильном подборе компьютерной программы для создания корпуса. Программное обеспечение электронного корпуса позволяет каждому исследователю при минимальных затратах усилий самостоятельно создавать на основе

корпуса полные базы данных в соответствии со своими исследовательскими задачами, систематизировать данные по различным заданным параметрам.

Многие европейские корпусы разговорной и диалектной речи используют систему EXMARaLDA (Extensible Markup Language for Discourse Annotation), т.е. расширенную маркированную систему лингвистической аннотации разговорной речи. Это система программ и инструментов для создания, управления, аннотирования и обработки корпуса разговорной речи. Базовой программой для первичного создания корпуса текстов и их аннотирования является Partitur Editor, название которой уже само определяет тип ввода информации: партитурная нотация. В отличие от так называемой драматургической нотации, предполагающей вертикальное расположение текста, партитурная нотация, считающаяся более удачной, строится как музыкальная партитура, но вместо инструментов выступают участники коммуникации. Это позволяет более точно отразить процесс общения в абсолютном измерении (на временной оси) и в относительном измерении, характеризующем речь участников коммуникации в сравнении друг с другом (одновременное говорение, паузы, вставки). Чисто технически партитурная нотация требует большей точности и более сложна в написании. Однако использование специальных средств компьютерной поддержки позволяет упростить создание партитурных транскриптов речи [2, с. 124].

Программный пакет EXMARaLDA позволяет членить речевой поток в корпусе различными способами, предполагает лингвистическое, метаязыковое и внелингвистическое аннотирование как отдельных единиц текста, так и его фрагментов, содержит метаданные, релевантные для автоматической обработки диалектных текстов. Важным преимуществом этой программы являются ее технические характеристики, возможность конвертирования в другие часто используемые компьютерные форматы и совместимость с другими приложениями и операционными системами. Она позволяет также настраивать формат выдачи текстовых фрагментов от одного слова и предложения до текста, в зависимости от целей исследования (параметр 3). Регулируемые параметры выдачи единиц корпуса и возможность лингвистического аннотирования принципиально важны именно для синтаксических и семантических исследований. Если для первых релевантным будет являться предложение или даже абзац, то для вторых во многих случаях важен более широкий контекст, чем отдельное предложение или словоупотребление.

Следующий параметр (4) определяет формы представления диалектных текстов в корпусе. В большинстве корпусов диалектные тексты представлены только в виде полуорфографической/полутранскрипционной записи. Такая фиксация диалектной речи не позволяет изучать ее фонетическую сторону, в этих условиях бóльшую актуальность приобретает вопрос о включении в корпус аудио- и видеозаписей диалектной коммуникации

и формах их соотнесения с символьной расшифровкой. Это позволяет программа EXMARaLDA Partitur-Editor, однако процесс синхронизации каждого отрезка речи (как правило, реплики говорящего) является достаточно сложным с технической точки зрения. Тем не менее наличие аудиозаписи делает корпус более интересным и полноценным. Достичь достаточного единообразия отображения диалектных текстов различных диалектных групп и систем и передать основные диалектные признаки в письменной форме позволяет система HIAT (halbinterpretative Arbeitstranskription – полуинтерпретативная рабочая транскрипция), которая позволяет с помощью традиционных орфографических знаков передать особенности звучания, опираясь на традиционные буквенно-звуковые соответствия. Использование системы транскрибирования HIAT в EXMARaLDA Partitur-Editor позволяет также аннотировать каждый элемент текста не только с лингвистической точки зрения (грамматические категории, формы слова, его стандартное литературное соответствие), но и сопроводить транскрипцию внелингвистическим комментарием (мимика, действия респондента (смех, ироничный тон и т.д.), длительность неразборчивых фрагментов) и синхронизировать ее с аудио- или видеозаписью. Система позволяет также фиксировать параллельную, синхронную речь нескольких говорящих, что очень важно при изучении разговорной речи.

Вторым этапом создания корпуса является объединение затранскрибированных, размеченных и аннотированных текстов в корпус. Для этого используются программы корпусного менеджера, например EXMARaLDA CoMa (Corpus Manager). Этот инструмент в полной мере соответствует требованиям, предъявляемым к корпусным менеджерам: корпусный менеджер должен: строить полные конкордансные списки; искать не только отдельные слова, но и словосочетания; осуществлять поиск по шаблонам (сложные запросы); сортировать списки по нескольким критериям, выбираемым пользователем; давать возможность отображать найденные словоформы в расширенном контексте; давать статистическую информацию по отдельным элементам корпуса; отображать леммы, морфологические характеристики словоформ и метаданные (библиографические, типологические) и пр. Объем возможностей по выдаче данных зависит от первоначального аннотирования, однако этот инструмент позволяет работать как с целым корпусом, так и с его разделами по выбору исследователя.

Как уже отмечалось, оптимальным размером выдачи данных для семантических исследований является микроконтекст (хотя бы на уровне абзаца). Для верификации данных и правильной их лингвистической интерпретации исследователь имеет возможность полнотекстового доступа. Таким образом, исходной предпосылкой создания корпуса является наличие некоторого количества текстов, различные по объему фрагменты которых

в последующем являются центральной единицей анализа в лингвистическом корпусе. И такое смещение акцентов в понимании текстового корпуса (от простого собрания) до объекта анализа с помощью автоматизированных систем определяет сферу применения методов и инструментов корпусной лингвистики.

Таким образом, электронные корпусы диалектных текстов являются принципиально новым источником, способствующим приобщению диалектологии к современной научной лингвистической парадигме, в которой изучение основных языковых черт диалекта было бы автоматизировано, обеспечивало бы перекрестные исследования в текстах различных говоров, облегчало бы поиск и выборку необходимых данных и позволяло бы проводить диахронические исследования хотя бы на примере нескольких десятилетий.

Исследование выполнено при финансовой поддержке РГНФ в рамках научно-исследовательского проекта № 12-04-00360 «Текстовый корпус немецких диалектов на Алтае».

Список литературы

1. Ахманова О.С. Словарь лингвистических терминов. – М. : КомКнига, 2007. – 576 с.
2. Баранов О.Н. Введение в прикладную лингвистику. – М. : Едиториал УРСС, 2003. – 360 с.
3. Захаров В.П., Богданова С.Ю. Корпусная лингвистика : учебник для студентов гуманитарных вузов. – Иркутск : ИГЛУ, 2011. – 161 с.
4. Крючкова О.Ю., Гольдин В.Е., Сдобнова А.П. Корпус русской диалектной речи: концепция и параметры оценки. – URL: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf.
5. Юрина Е.А. Томский диалектный корпус: в начале пути // Вестник Томского государственного университета. - 2011. – № 2 (14). - С. 58-63.
6. Lemnitzer L., Zinsmeister H. Korpuslinguistik. Eine Einführung. – Tübingen : Narr Verlag, 2010. – 214 s.
7. Perkuhn R., Keibel H., Kupietz M. Korpuslinguistik. – Paderborn : Wilhelm Fink Verlag, 2012. - 144 s.
8. Schmidt Th. Grundzüge von EXMARaLDA – einem System zur komputergestützten Erstellung und Auswertung von Korpora gesprochener Sprache. – URL: http://www1.uni-hamburg.de/exmaralda/files/Backstein.pdf.

Рецензенты:

Москалюк Л.И., д.фил.н., профессор ФГБОУ ВПО «Алтайская государственная педагогическая академия», г. Барнаул;
Колесов И.Ю., д.фил.н., профессор ФГБОУ ВПО «Алтайская государственная педагогическая академия», г. Барнаул.