

УДК 811.161.1; 81-25; 81'322

Е.А. Юрина

ТОМСКИЙ ДИАЛЕКТНЫЙ КОРПУС: В НАЧАЛЕ ПУТИ

В статье представлен проект создания машинного корпуса текстов говоров Среднего Приобья, нацеленный на перевод диалектного архива в формат электронного корпуса с возможностями разноаспектного поиска. Дается краткая характеристика диалектного архива; определяются актуальные задачи и этапы работы; рассматриваются параметры разметки текстового массива. Подробно характеризуются параметры метаразметки, включающие паспортизацию текста, тематические и жанровые определители.

Ключевые слова: Томский диалектный корпус, параметры метаразметки, говоры Среднего Приобья.

Подзаголовок «в начале пути» в контексте без малого семидесятилетней научной деятельности Томской диалектологической школы звучит парадоксально. Первые наблюдения относительно говоров Среднего Приобья были сделаны А.Д. Григорьевым в 1917–1918 гг. А начало их комплексного исследования датируется 1946 г. Именно в этом году состоялась первая послевоенная диалектологическая экспедиция, которую проводили томские лингвисты в соответствии с планом работ Института русского языка АН СССР по сбору материалов к «Диалектологическому атласу русского языка». Для обследования были выбраны старожильческие сёла, основанные первопоселенцами: Томский район – сёла Вершинино и Батурино; Туганский район – сёла Халдеево, Турунтаево, Подломск и др. Источником изучения выступала живая разговорная речь сельских жителей.

Со временем определились типичные фонетические, лексические и грамматические черты речи сибирских старожилов: аканье (*па нача`м*), иканье (*гри-чи`ха*), долгие твердые шипящие (*та`ишиыт*), стяжённые формы местоимений, прилагательных и глаголов (*така`хоро`ша*) и др. О.И. Блиновой была предложена трактовка термина «русский старожильский говор Сибири» с опорой на совокупность фонетико-грамматических и лексических черт [1]. Полная характеристика исследуемых говоров, представляющих собой особый тип вторичного сибирского старожильского говора с северной диалектной основой, изложена в коллективной монографии «Русские говоры Среднего Приобья» [2, 3]. В книге описаны типологические черты среднеобского диалекта и дана языковая характеристика каждой группы говоров в его составе: Нарымской, Прикетской, Приобской, Притомской, Причулымской [2. С. 30–33].

Первые тетради были записаны в транскрипции: «*па фс`о`j но`чи гул`а`jут / д`ефк`и да роб`а`ты...*» (1946 г.). Впоследствии речь информанта записывалась на магнитофон, а затем расшифровывалась в так называемой «полутранскрипции». В этих записях употреблялись буквы я, ю, е, ё, не обозначалась мягкость согласных перед гласным переднего ряда, не ставилось

ударение, если оно не отличалось от литературного, ставились знаки препинания, писались заглавные буквы.

В результате за 66 лет работы коллективом томских диалектологов было обследовано свыше 400 населенных пунктов Томской и Кемеровской областей. На сегодня архив представлен 1500 тетрадами с записями диалектной речи, отражающей русские старожильческие говоры Среднего Приобья. На базе данного материала были созданы картотеки диалектных словарей более чем на 1 000 000 единиц: «Среднеобского диалектного дифференциального словаря» [4, 5], «Словаря диалектно-просторечной лексики» [6], «Полного словаря говора села Вершинино» [7], «Полного словаря диалектной личности» [8] и целого ряда других словарей разного типа (мотивационного, образных слов и выражений, антонимов и др.). Библиотека среднеобских словарей насчитывает 46 томов.

Материалы архива являют собой уникальный памятник народной речевой культуры, бесценный источник исследования истории и современного состояния диалектной формы русского национального языка. На базе словарей, картотек и текстовых записей проводится широкий круг разноаспектных исследований, связанных не только с описанием процессов и феноменов диалектного дискурса, но и с решением важнейших теоретических и методологических вопросов лингвистики. Это такие проблемы, как мотивированность языкового знака, лексико-семантические категории (мотивированность, образность, интенсивность, оценочность, экспрессивность), категории традиционной народной и речевой культуры, языковое миромоделирование, жанры речи, языковая личность диалектоносителя, метаязыковое сознание личности, лексикографический метод, метод полевого исследования языковых категорий и др. (см. подробнее в [9]).

Задача создания электронного варианта текстового архива говоров Среднего Приобья, а также компьютерного словаря среднеобского диалекта назрела давно. Её актуальность была осознана уже двадцать лет назад, в период становления отечественной компьютерной лингвистики. Именно в конце 1980-х гг. Г.А. Раков, заведовавший кафедрой русского языка ТГУ с 1990 по 1994 г., начал исследовательскую деятельность, связанную с созданием машинного идеографического словаря среднеобских говоров [10–12]. Им была организована работа исследовательской группы, в которую входили молодые преподаватели кафедры М.В. Курышева, Г.В. Калиткина, В.Г. Наумов, аспиранты С.Г. Комков, студенты С. Родионова, И. Колмыкова. Велись работы по компьютерному набору диалектных текстов, автоматическому созданию конкорданса, семантической разметке словарного массива. К сожалению, разработки были прерваны в связи с трагическим уходом из жизни этого замечательного человека и ученого, а созданные в то время электронные материалы утрачены.

За последние десятилетия компьютерные технологии шагнули далеко вперед. Мы наблюдаем расцвет корпусной лингвистики на базе интернет-технологий. В 2004 г. начал функционировать Национальный корпус русского языка [13]. Работа над созданием электронных корпусов русской диалектной речи ведется во многих научных центрах России: Институте русского языка РАН в рамках проекта «Национальный корпус русского языка» [14], в

Институте лингвистических исследований РАН, Санкт-Петербургском, Саратовском [15, 16] и других государственных университетах.

Создание диалектного корпуса на материале говоров Среднего Приобья крайне необходимо. Уже имеется некоторый задел в виде базы электронных текстов (около 10% от имеющегося рукописного архива), компьютерный набор которых осуществлен диалектологами-исследователями и студентами, проходившими диалектологическую практику. И если 20 лет назад, в период становления компьютерной лингвистики, актуальность исследований была связана с поиском и разработкой методов компьютерной обработки текстовых массивов, то сегодня перед Томской диалектологической школой стоят не менее актуальные задачи прикладного характера. Они состоят в эффективном использовании уже имеющихся в арсенале корпусной лингвистики технологий, позволяющих сохранить архив, наиболее ранние записи которого находятся в аварийном состоянии; создать максимально полную текстовую и словарную базы данных и осуществлять их ведение; оптимизировать поиск данных, их статистическую обработку и анализ при помощи современной программной платформы (составление конкордансов, словников, подкорпусов, соответствующих исследовательским задачам). Применение имеющихся в арсенале корпусной лингвистики методов позволит вывести исследования Томской диалектологической школы на новый уровень, соответствующий современным техническим возможностям.

Для решения поставленной задачи в 2010 г. группой томских диалектологов и программистов (Е.А. Юрина, В.В. Поддубный, С.В. Волошина, М.А. Толстова, О.Г. Шевелев) разработан проект создания диалектного Корпуса говоров Среднего Приобья. В результате работы планируется полный перевод рукописных текстов на электронные носители, проведение лингвистической разметки текстов с целью последующей машинной обработки, создание программ по автоматизированной разметке текстов, машинному поиску, анализу результатов поиска, создание лингвистических баз данных и программной оболочки корпуса. Основным результатом проекта станет доступный для широкого круга исследователей диалектный корпус, который впоследствии может быть интегрирован в качестве подкорпуса в более крупные национальные проекты.

Предполагается поэтапно осуществить следующие виды работ:

1. Перевод рукописных записей диалектных текстов в электронный формат методом компьютерного набора и создания электронных текстовых файлов.

Для решения этой задачи проводится оцифровка рукописных тетрадей и сохранение изображений в формате DJVU. Более 50% диалектных тетрадей уже оцифровано. Проводится набор текстов и редактирование ранее набранных текстов для унификации графических символов.

С этой целью разработана единая система графической передачи особенностей диалектной фонетики и морфологии в тексте с использованием символов юникода. Общерусские слова и словоформы, не имеющие отличий от литературного языка, передаются в орфографической записи, а диалектные, диалектно-просторечные единицы и диалектные варианты общерусских слов записываются с передачей фонетических и морфологических особенностей. Долгие твердые шипящие передаются написанием двух символов **шш**, **жж**

(доиши, дожжа, ешио); для обозначения мягкости согласных используется символ ь (зерькало); мягкое [ш'] передается буквой щ (тащмя-то); у неслоговое передается символом ъ (война когда быѡа); г фрикативное обозначается как ґ; твердость согласных перед гласными переднего ряда передается буквами, обозначающими соответствующие гласные непяреднего ряда: э, ы, а (вобшэ-то). В словах с диалектными особенностями проставляется ударение. Например: «Иґры вся`ки ра`зны были. Пляса`ля ши`бко. У нас был Введе`нье [праздник]. Отку`пишь дом на три-четыре дня. Деньги соберёшь с молодёжья. Отдашь хозяину. Поу`жинашь вечером и на вечёрку. В дома`-то таки`. Для молодёжу» (с. Вершинино Том. обл., Вершинин И.В., 1909 г.р.).

2. Проведение метаразметки массива диалектных текстов, задающей параметры паспортизации текстов, квалификации тематики и жанровой специфики.

Паспортная часть включает сведения об информантах, о времени и месте записи, собирателе, номере тетради. Далее параметризацию метаразметки продолжает квалификация тематики и жанровой специфики диалектного текста.

Тематическая характеристика включает указание на время описываемых событий (двадцатые 20в; тридцатые 20в; сороковые 20в и т.д.) и собственно тематику, которая задается через указание на ключевой концепт, определяющий смысловое единство текста (Война, Семья, Огород, Колхоз, Дом т.д.).

Анализ тематики диалектных записей выявил ключевые понятия традиционной крестьянской культуры, актуализированные в дискурсивной практике сибирских старожилов. Номинации этих концептов составляют основной список маркеров тематической метаразметки, в который также включаются определения тем, связанных с оценкой жизни и оценкой человека, с происшествиями, и другими явлениями, составляющими быденную жизнь крестьянина.

Приведем полный список тематических определителей: Семья. Односельчане. Любвные связи. Чужие (бураты, хакасы, татары, ссыльные, городские и т. д.). История села. Город. Дом. Постройки. Труд. Колхоз. Орудия труда. Профессия. Огород. Сенокос. Картофель. Лён. Животноводство. Пчеловодство. Кедровый промысел. Рыбалка. Охота. Лесозаготовка. Строительство. Ремесло. Ягоды. Грибы. Цветы и травы. Болезни и лечение. Еда. Посуда. Одежда. Пьянство. Церковь. Колдовство и суеверия. Приметы. Обряды. Сватовство. Свадьба. Похороны. Крестины. Праздники. Развлечения. Погода. Природа. Политика. Исторические личности. Война. Революция. Коллективизация. Происшествия. Оценка современности. Оценка человека.

Жанровая характеристика текста задается через указание на жанр текста, обусловленный типом речи и характером речевого произведения в целом. Опираясь на результаты исследований О.А. Казаковой [16] и С.В. Волошиной [17], мы выделили такие жанры, как биографический рассказ, сюжетный рассказ, описание, рассуждение, интервью, сказка, песня, частушка, пословица.

Завершает жанровую квалификацию указание на реализованный в тексте субжанр (речевой жанр). В метаразметке используется следующая параметризация реализованных в диалектном тексте субжанров:

Информативные: сообщение о событии, сообщение о намерении, сообщение о мнении, сообщение-цитирование, предположение, объяснение, жалоба, предупреждение.

Императивные: просьба, распоряжение, поручение, приказ, предложение, совет.

Ритуальные: приветствие, прощание, извинение, благодарность, приглашение, угощение, пожелание.

Оценочные: похвала, осуждение, самооценка, оценка.

Образец записи: Информ_сообщение о намерении, Импер_поручение, Ритуал_извинение, Оцен_похвала.

Паспортная часть и другие параметры метаразметки предваряют диалектный текст в следующей записи:

#МестоЗаписи с. Вершинино, Томский район, Томская область

#МестоЗаписиКоротк Верш

#ДатаЗаписи 1991

#Информант Вершинин Иван Васильевич

#ИнформантКоротк ВершининИВ

#ИнформантПол муж

#ИнформантГодРожд 1906

#ИнформантМестоРожд Неизв_информантместорожд

#ИнформантОбразов Неграмотн

#Собиратель МихалеваЕВ

#Тетрадь 1168

#Тематика Семья, Война, Свадьба, Праздники, Развлечения, Кедровый промысел, Труд

#ВремяСобытий двадцатые 20в, сороковые 20в

#Жанр Биографический рассказ

#Субжанр Информ_сообщение о событии

Далее за паспортной частью следует запись диалектного текста. Приведем в качестве примера короткий фрагмент записи речи указанного диалектоносителя:

Я-то на Курской дуге воевал. Орёл – Ку`рска дуга. У-у, деточки мое`, всего там насмотрелись. Чего только не ви`дывали. Взя`тый в сорок втором году был. А так оставляли все по семейному положению. Дети были. Чё. Пришел в ноябре сорок пятого. Везде были, все видели. В Литве был, в Германии был, в Норвегии был. В лагере, в плену. Ехали через Швецию, Финляндию. Старуха не ждала уж. Всё, grit, по`мер, наверно. ...Пой там ей-ей. Было. Самолеты, орудия, минометы. Дело было горя`че. Орёл... Ку`рска дуга.

3. Создание сводного электронного словаря говоров Среднего Приобья. Решение этой задачи предполагает формирование базы электронных версий всех диалектных среднеобских словарей, а также создание и разметку сводного словника. Сводный словарь будет включать список всех лексем и их вариантов (акцентологических, фонематических, словообразовательных, морфологических); список морфологических форм лексем; указание постоянных и изменяемых морфологических признаков; толкование лексических значений; квалификацию единицы с точки зрения соотношения с формами национального языка (общерусское, диалектное, диалектно-просторечное, диалектный вариант общерусского слова); квалификацию семантических свойств слова (мотивированное/немотивированное, экспрессивное/нейтраль-

ное, образное, оценочное, интенсивное и т.д.); стилистическую маркированность (народно-поэтическое, сниженное, бранное).

4. Разработка и создание программной платформы корпуса, позволяющей осуществлять автоматизированную разметку текстов, хранить лингвистические базы данных, осуществлять многоаспектный поиск по заданным параметрам языковых единиц, проводить статистическую обработку материала, сохранение результатов поиска.

Решение намеченных в начале исследовательского пути задач связано с осуществлением большого объема практической работы. И тот факт, что к её осуществлению с большим энтузиазмом подключаются студенты и аспиранты филологического факультета ТГУ, внушает разработчикам чувство оптимизма и уверенность в успехе.

Литература

1. Блинова О.И. О термине «старожилческий говор Сибири» // Вопросы языкознания и сибирской диалектологии. Томск, 1971. Вып. 2. С. 3–8.
2. Русские говоры Среднего Приобья / О.И. Блинова, Л.Г. Гынгазова, Л.А. Захарова и др.; ред. В.В. Палагина. Ч. 1. Томск: Изд-во Том. ун-та, 1985. 205 с.
3. Русские говоры Среднего Приобья / О.И. Блинова, О.И. Гордеева, М.Н. Янценецкая и др.; ред. В.В. Палагина. Ч. 2. Томск: Изд-во Том. ун-та, 1989. 324 с.
4. Словарь русских старожилческих говоров средней части бассейна р. Оби / сост. О.И. Блинова, Ф.П. Иванова, В.В. Палагина, О.М. Соколов, М.Н. Янценецкая; ред. В.В. Палагина. Томск: Изд-во Том. ун-та, 1964. Т. 1: А–Е. 143 с.; 1965. Т. 2: Ж–О. 233 с.; 1967. Т. 3: П–Я. 249 с.
5. Среднеобский словарь: (Дополнение) / сост. Л.Г. Гынгазова, Т.А. Демешкина, Н.Г. Нестерова и др.; ред. В.В. Палагина. Томск: Изд-во Том. ун-та, 1983. Ч. 1: А–К. 180 с.; 1986. Ч. 2: П–Я. 212 с.
6. Словарь просторечий русских говоров Среднего Приобья / Сост. О.И. Блинова, В.В. Палагина, С.В. Сыпченко и др.; ред. О.И. Блинова. Томск: Изд-во Том. ун-та, 1977. 183 с.
7. Вершининский словарь / сост. Т.Б. Банкова, О.И. Блинова, Е.В. Иванцова, В.В. Палагина и др.; Гл. ред. О.И. Блинова. Томск: Изд-во Том. ун-та, 1998. Т. 1: А–В. 308 с.; 1999. Т. 2: Г–З. 309 с.; 2000. Т. 3: И–М. 318 с.; 2001. Т. 4: Н–О. 368 с.; Т. 5: П. 512 с.; Т. 6: Р–С. 454 с.; Т. 7: Т–Я. 524 с.
8. Полный словарь диалектной языковой личности / сост. О.И. Гордеева, Л.Г. Гынгазова, Е.В. Иванцова и др.; под ред. Е.В. Иванцовой. Томск: Изд-во Том. ун-та, 2006. Т. 1: А–З. 358 с.; Т. 2: И–О. 338 с.; Т. 3: П–Р. 324 с.
9. Томская диалектологическая школа: историограф. очерк / под ред. О.И. Блиновой. Томск: Изд-во Том. ун-та, 2006. 392 с.
10. Раков Г.А. Принципы отбора языковых данных при составлении словаря народного говора на машинном носителе // II Всесоюз. конф. по созданию Машинного фонда русского языка. М., 1987.
11. Раков Г.А. Автоматический словарь народного говора как фрагмент Машинного фонда русского языка // III Всесоюзн. конф. по созданию Машинного фонда русского языка. М., 1989.
12. Раков Г.А. Диалектные словари и ЭВМ: Подготовка данных и некоторые результаты сравнения // Русские говоры Сибири: Семантика. Томск, 1995. С. 53–62.
13. <http://www.ruscorpora.ru>
14. Летучий А.Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 215.
15. Гольдин В.Е. К проекту текстового диалектологического подфонда Машинного фонда русского языка // Докл. III Всесоюз. конф. по созданию машинного фонда русского языка. М., 1990.
16. Казакова О.А. Диалектная языковая личность в жанровом аспекте. Томск: Изд-во Том. политехн. ун-та, 2007. 196 с.
17. Волошина С.В. Речевой жанр автобиографического рассказа (на материале диалектной речи) // Вестн. Том. гос. ун-та. Филология. 2010. № 2 (10). С. 5–10.