СЛОВАРНЫЕ ПРОЕКТЫ И ТРУДЫ

УДК 811.161.1; 81-25; 81'322
DOI: 10.17223/22274200/11/4

Е.В. Иванцова

ТОМСКИЙ ДИАЛЕКТНЫЙ КОРПУС: ОБОСНОВАНИЕ КОНЦЕПЦИИ И ПЕРСПЕКТИВЫ РАЗВИТИЯ[1]

В статье рассматривается реализуемый в томской диалектологической школе новый проект корпусного представления русских говоров Среднего Приобья. Обосновывается модификация первоначальной концепции корпуса. Освещаются общие установки его создания и ориентация электронного ресурса, определяющиеся современными задачами диалектологии, интересами томских диалектологов и характером имеющихся в их распоряжении материалов. Описаны принципы подачи и метаразметки диалектной базы данных, перспективы развития корпуса.
Ключевые слова: *русские говоры Сибири, Среднее Приобье, диалектный корпус, концепция.*

В современных научных исследованиях с каждым годом растет потребность в создании полноохватных, доступных и удобных для научного поиска электронных источников информации. Наряду с масштабными проектами национальных корпусов, ставящими цель репрезентативного представления всех типов дискурса того или иного языка, показателен и явно выраженный интерес лингвистов к проектам создания диалектных корпусов. Они созданы или создаются в Германии, Австрии, Испании, Китае, Португалии, Финляндии, Скандинавии, Польше, Литве, Грузии (см., например, CORDIAL-SIN; Helsinki corpus of English dialects; Freiburg English Dialect Corpus; The Nordic Dialect Corpus; Archiv fur gesprochenes Deutsch, Die bayerische Dialektdatenbank; LEXDIALGRAM и др.), на территории России – в научных центрах Москвы, Саратова, Казани, Славянска-на-Кубани, Вологды и ряда других городов (Диалектный подкорпус в составе Национального корпуса русского языка, Саратовский диалектологический корпус, Электронная библиотека русских народных говоров, Электронный корпус диалектной культуры Кубани и др.). Это явление закономерно отражает, с одной стороны,

[1] Исследование выполнено при поддержке гранта Российского научного фонда (проект № 16-18-02043).

понимание значимости местных народных говоров как первооосновы национального языка и национальной ментальности, с другой – насущную потребность в получении эффективных инструментов работы с диалектными материалами, в большинстве случаев доступными очень ограниченному кругу пользователей.

Создание диалектного корпуса входит в число актуальных проблем и для томской диалектологической школы, являющейся старейшим научным центром по изучению народно-речевой культуры Сибири. Значимость этого проекта определяется не только внутренними потребностями развития самой школы, связанными с необходимостью продуктивного использования экспедиционных и словарных данных в условиях интеграции гуманитарных и точных наук, но и важностью репрезентации сибирских материалов для научного сообщества. Хотя диалектологическая работа ведется во многих научных центрах Сибири (в их числе Тюмень, Омск, Новосибирск, Кемерово, Новокузнецк, Красноярск и др.), диалектные корпуса этих территорий пока не созданы; почти не представлены говоры Сибири и в региональном подкорпусе НКРЯ.

Несмотря на то, что в Томском государственном университете внедрение достижений машинной лингвистики в сферу изучения народной речи было начато Г.А. Раковым еще в 80–90-е гг. прошлого века[1], в силу многих обстоятельств вплотную к решению обозначенной задачи диалектологи подошли только сейчас.

В 2010 г. Е.А. Юриной была сформулирована идея создания Томского диалектного корпуса (ТДК) и обозначены первые шаги работы над ним [1]. Началось сканирование архива полевых записей, вырабатывались принципы графической передачи устной речи и метаразметки диалектных текстов [2]. Однако общий замысел был очерчен очень эскизно и в дальнейшем работа по его реализации не была продолжена.

[1] Группой молодых ученых под его руководством велись работы по составлению диалектного конкорданса на среднеобском материале. Г.А. Раковым также был разработан и реализован проект «человеко-машинного» идеографического словаря нарымского говора.

В условиях углубленной работы с материалом и смены рабочей группы[1] первичные идеи проекта получили развитие и подверглись корректировке. Целью данной публикации является освещение выработанной на сегодняшний день общей концепции ТДК, обоснование его установок и внесенных изменений, характеристика принципов метаразметки: намечены также задачи развития нового корпуса в ближайшей и отдаленной перспективе.

1. Общая установка создания корпуса

Частная, казалось бы, цель создания электронной базы данных одного из регионов связана с решением проблем, стоящих перед корпусной лингвистикой в целом.

В идеале корпусное представление отдельных говоров должно быть максимально унифицировано для того, чтобы создаваемые в том или ином научном центре корпуса вливались в более общие, становясь подкорпусами национальных корпусов, а те, в свою очередь, соотносились между собой, давая материал для сопоставительных исследований. Вместе с тем думается, что интеграция ТДК в более крупные проекты, предполагавшаяся первоначально [1. С. 60], возможна только в достаточно отдаленной перспективе. Многочисленные факторы порождают различие концептуальных подходов к созданию таких продуктов. Среди этих факторов – и различия языковых подсистем, и имеющийся архив, на основе которого создается корпус, и первоочередные задачи, решение которых предполагается с опорой на новый ресурс. Созданные и создаваемые российские диалектные корпуса отражают это положение дел. Они очень разнородны по принципам представления собранных материалов: общей архитектонике корпуса, глубине разметки, поисковым возможностям и т.д. Программы работающих в данном направлении научных центров как в России, так и за рубежом слабо координируются. Очевидно, время для решения этой масштабной задачи еще не наступило; выработка единых принципов корпусной репрезентации диалектной речи – дело будущего[2].

---

[1] Работа над концепцией корпуса в последнее время осуществлялась группой молодых ученых-филологов (Н.А. Агапова, С.В. Волошина, А.А. Долганина, С.С. Кузнецова (Земичева)) и программиста (Н.Ф. Картофелева) под руководством автора настоящей статьи.

[2] Ср.: «…создание универсального языкового ресурса, который бы мог удовлетворить все запросы ученых, вряд ли возможно, так как сегодня лингвистика значительно расширила поле своих исследований. Поэтому исследователи, разрабатывая новый ис-

Исходя из сказанного выше, концепция ТДК разрабатывается с учетом как ключевых направлений исследования народной речи томской диалектологической школы, коррелирующих с достижениями современной науки о языке, так и характера материалов, имеющихся в распоряжении диалектологов.

2. Ориентация корпуса

Определяющей ключевые параметры концепции ТДК является его ориентация.

В созданных и создаваемых диалектных корпусах отражаются общие процессы развития языкознания. Наряду с традиционной системно-структурной парадигмой на рубеже веков утверждается парадигма антропоцентрическая; при этом она не отменяет предшествующих, придавая современной лингвистике полипарадигмальный характер [4. С. 228].

Структурные закономерности языковых подсистем местных говоров репрезентируют корпусы с фонетической или грамматической ориентацией (например, Диалектный корпус мандаринского китайского языка или Хельсинкский диалектный корпус британского английского языка), реже – лексической (например, Корпус грузинских диалектов). Акцентируется при этом своеобразие диалектной фонетики, грамматики или лексики на фоне кодифицированного языка. Эта особенность присуща и диалектному подкорпусу НКРЯ, где выбрана «морфологически ориентированная стратегия» и основное внимание при разметке уделяется соотнесению областных словоизменительных форм с литературными [5. С. 215]. Типичной в таких случаях является выдача минимальных контекстов.

В то же время в диалектных корпусах начинают находить отражение запросы, соотносимые с усилением антропоцентрического начала. Увеличивается число электронных баз данных, в которых возможен доступ к полным текстам. Наряду с библиотеками текстов (к ним фактически относятся, например, испанский и польский областные корпуса, среди отечественных – Электронная библиотека русских народных говоров), текстоцентрическая форма представления диалектных материалов появляется и в собственно корпусных продуктах (в их числе – болгарский, эстонский, скандинавский,

---

точниковый ресурс, учитывают специфику лингвистического направления, а также и природу самого объекта» [3].

шотландский, португальский корпуса). Показательным в этом отношении можно считать развитие концепции диалектного подкорпуса НКРЯ, в котором первоначально выдавались только фрагментарные контексты; недавно составителями было принято решение о возможности работы с целостными текстами по запросу диалектологов [6]. Разрабатываются проекты, имеющие лингвокультурологическую направленность: «Электронный корпус диалектной культуры Кубани» [7] и мультимедийный вологодский корпус текстов «Жизненный круг» [8]. Идеи формирующегося направления коммуникативной диалектологии находят воплощение в Саратовском диалектологическом корпусе ([9] и др.).

Ориентацию ТДК можно определить как лексико- и текстоцентрическую. Выдвижение в качестве основных объектов анализа лексикона и текста обусловлено и развитием диалектологии в целом, и интересами исследователей Томской диалектологической школы.

Изучение среднеобских русских старожильческих говоров было начато с их поярусного (фонетического, грамматического, словообразовательного, лексического) системно-структурного описания, решения проблем типологизации и исторического генезиса (см. обобщающее представление этих итогов в работе [10]). На рубеже 70–80-х гг. XX в. произошел переход от описания различных тематических групп лексики к изучению лексических явлений (синонимии, антонимии, варьирования, мотивированности), выявлению специфики лексико-семантических категорий (образности, интенсивности и др.) и детальному анализу их функционирования на диалектном материале. Начинается исследование метаязыкового сознания диалектоносителей, организации диалектного высказывания и текста, системы речевых жанров и концептосферы народной речи. Параллельно с этим многоаспектным обследованием среднеобского диалектного массива на протяжении всего периода существования школы осуществлялась лексикографическая деятельность. Томскими диалектологами создано около 30 словарей разных типов – дифференциальных и недифференциальных, толковых и аспектных, прямых и обратных, представляющих группу говоров, говор одного села и идиолект отдельный языковой личности (см. обзор в работе [11]).

Перемещение в фокус исследований функционального, текстоцентрического, когнитивного подходов к постижению сущности на-

родно-речевой культуры закономерно вызвало необходимость обращения исследователей к диалектному дискурсу и тем его составляющим, которые позволяют ставить вопрос о специфике коммуникации сельского социума, мировидения и миропонимания диалектоносителей, своеобразия языковой картины мира. Этими составляющими являются прежде всего различные виды текстовой организации дискурса (речевые жанры, метатексты / тексты спонтанной речи, монологическое / диалогическое общение, прецедентные высказывания и т.д.) и лексико-фразеологический слой языковой системы, также выступающий в качестве «когнитивного ключа» для постижения народной ментальности. Практика создания диалектных словарей тоже требует внимания как к единицам лексикона, так и к тексту, в котором реализуются семантика и функциональные свойства лексем. Таким образом, ориентация корпуса как тексто- и лексикоцентрическая закономерна.

3. Принципы представления материалов

Характер подачи и метаразметки материалов в создаваемом корпусе определяется особенностями данных, которыми располагают томские диалектологи, и задачами нового электронного ресурса.

3.1. Подача материалов

Массив записей диалектной речи среднеобского региона, имеющийся в распоряжении ученых, формировался в течение столетия[1]. Это обстоятельство, несомненно, позволяет рассматривать созданный архив как ценный источник изучения системы сибирских говоров в диахронии, но вместе с тем создает сложности представления ресурсов, в различные периоды полученных в различавшихся условиями проведения экспедиций, характером применявшихся при записи средств, целями экспедиционного обследования территории, способами передачи звучащей речи.

Наиболее ранние материалы 1920–1950-х гг. зафиксированы от руки, в транскрипции. Блокнотные записи 1960–1970-х гг. также ручные, но в них уже вырабатываются принципы полуорфографической передачи диалектных особенностей звучащей речи, принятые впоследствии в томской диалектологической школе.

С конца 1970-х – начала 1980-х гг. в экспедициях начинают использоваться магнитофоны. Архив этого периода представляет со-

[1] Первые экспедиции были проведены в 1917–1922 гг. А.Д. Григорьевым.

бой рукописную расшифровку записей на магнитной ленте и отчасти ручные записи, которые еще встречаются в связи с недостаточной обеспеченностью техникой. Сами аудиоматериалы сохранились фрагментарно.

Экспедиции последних лет, оснащенные техническими средствами нового поколения, позволили формировать фонотеку оцифрованных аудио- и видеофайлов (около 190 часов). Они расшифровываются сразу в виде электронного набора. Фотографии немногочисленны и также относятся главным образом к Новейшему времени. В рукописных тетрадях раннего периода имеются рисунки предметов народного быта, промысловых инструментов и т.п.

Разнородность материала потребовала выработки принципов его единообразной подачи в корпусе.

Все имеющиеся блокнотные записи переводятся сейчас в компьютерный набор; на начало 2017 г. это свыше 1,5 млн словоупотреблений.

В качестве основной формы представления среднеобских говоров избран текст с орфографической передачей отдельных особенностей устной речи: твердых долгих шипящих (*шшука, ташшыт*), элементов цоканья (*цясто*), стяжения в формах прилагательных и глаголов (*больша', хоро'ша; знат, понима'шь*) и некот. др. «Полуорфографическая запись», утвердившаяся в начале 1960-х гг. при подготовке первого выпуска «Словаря русских старожильческих говоров средней части бассейна р. Оби» [12], последовательно проведена во всех словарных изданиях томских диалектологов; она же используется в иллюстративном материале научных публикаций[1]. Опора на этот принцип позволит унифицировать репрезентацию разнородного архива – от первых экспедиционных тетрадей до цифровых аудиозаписей последних лет. Избранный способ передачи звучащей диалектной речи можно считать универсальным для лексикологических, лингвокультурологических, дискурсивных и лексикографических исследований.

Правила однотипного отражения на письме устной речи в целом уже были сформированы на первом этапе работы над корпусом [1.

[1] Исключение составляют только работы по фонетической характеристике говоров Среднего Приобья К.М. Браславца, Г.А. Садретдиновой, О.А. Любимовой, В.А. Сенкевича 1950–1970-х гг., где упрощенная транскрипция применяется для передачи произношения звуков в словоформах; фразы и связные тексты в транскрипции в них не встречаются.

С. 60–61], хотя в них внесен ряд уточнений. Отказ от транскрипции отчасти восполняется возможностью обращения к имеющимся звуковым материалам; сохранившиеся кассетные и катушечные аудиозаписи предстоит перевести в цифровой формат. Предполагается также возможность доступа к дополняющим данные основного корпуса сканированным ручным записям, в том числе транскрибированным; их оцифровка (свыше 1000 единиц хранения) в основном уже завершена благодаря сотрудничеству с Научной библиотекой Томского государственного университета. В перспективе эти ресурсы могут составить единую систему.

3.2. Принципы метаразметки

Принципы метаразметки новой базы данных сочетают традиционные и новые для диалектных корпусов особенности, они также определяются ориентированностью корпуса.

За единицу метаразметки в данном корпусе принят текст, понимаемый как фрагмент диалектного дискурса, записанный от отдельного информанта и отличающийся признаками единства хронотопа (время, место записи) и условий фиксации речи.

Разработаны основные виды метаразметки вводимых в корпус текстов: паспортная, тематическая и разметка по типу текста.

Паспортная метаразметка включает экстралингвистические данные о введенном в корпус тексте. Ее параметрами являются указания на место и время произведенной записи, фамилию, имя и отчество информанта, его пол и год рождения, дополнительные данные о нем (при наличии таковых приводятся сведения о родителях, образовании, роде занятий, местах длительного проживания и др.), тип записи (от руки / с магнитной ленты / диктофона), наличие / отсутствие аудио- и видеофайлов, архивный номер тетради, на основе которой производился компьютерный набор; там, где это возможно, размещается фотография диалектоносителя.

Метаразметка по обозначенным в речи темам важна для выявления специфики диалектного дискурса, «зон актуального внимания» сельского социума, изучения концептосферы народной речи; она может стать полезной и при выборке лексических единиц для диалектных словарей.

Тематическая разметка присутствует в диалектном подкорпусе НКРЯ и Саратовском диалектологическом корпусе, однако методологические проблемы ее разработки по-прежнему являются актуаль-

ными для корпусной лингвистики. Существует мнение, что в диалектах «набор тем текстов мало отличается от литературного, но, естественно, гораздо более ограничен», а «диалектные тексты посвящены почти исключительно быту и обычаям» [5. С. 230], однако эти утверждения требуют проверки. Тематика диалектного дискурса еще только начинает изучаться [13, 14, 15]. Политематичность разговорной речи, с одной стороны, и достаточно высокая степень субъективности при интерпретации текста – с другой, осложняют процедуры выделения тем.

Первоначально сформированный перечень размечаемых в Томском диалектном корпусе тем [1. С. 61] был существенно переработан. Тематическое членение текста осуществлялось исходя из следующих посылок:

– вычленение тем среднеобского диалектного дискурса производится индуктивным путем, с опорой на реальную дискурсивную практику старожилов, а не на изначально составленную идеографическую схему;

– выделение тем идет в направлении от частного к общему. Составление перечня тем, отражаемых в разметке корпуса, начиналось с обозначения частных тем. В формирующийся список включались те из них, которые регулярно повторялись в записях; на этом же основании выделялись и подтемы. Далее частные темы обобщались до макротем, например: «Дом и усадьба»; «Одежда и обувь»; «Домашние вещи»; «Покупки и продажа»; «Условия жизни» → БЫТ;

– ни слишком обобщенное, ни слишком дробное выделение тем неудобно как при разметке, так и при пользовании материалами корпуса. В связи с этим единично встречающиеся частные темы не вносились в список и возводились к макротемам; углубление каждой темы не превышало трех уровней;

– номинации тем по возможности соотносились с лексиконом рядового носителя языка. Несколько исключений составляют случаи, когда книжные слова не имеют лаконичных аналогов в народной речи (темы «Сбор дикоросов», «Экология», «Мораль», «Досуг», «Репрессии»);

– упорядочение тем с логических позиций в составленном для корпуса перечне, как показал анализ, возможно лишь отчасти: многие темы пересекаются или синтетичны. В связи с этим список тем в корпусе дан по частотности: от наиболее частотных к редким;

– тема выделяется не в отдельном высказывании, а во фрагментах текста, обладающих признаком связности; прочие относятся к атематическим фрагментам;

– разметка опирается на принцип «мягкого» тематического членения зафиксированного речевого потока с возможностью частичного наложения границ вычленяемых текстов. Один фрагмент текста также может маркироваться как одновременно принадлежащий к нескольким темам.

В настоящее время на основании предварительного анализа части набранных текстов список макротем насчитывает 16 пунктов; наиболее дробный перечень тем второго порядка выявлен в макротеме РАБОТА (13 подтем, обозначающих виды работ). При разработанности отдельных подтем выделялись темы 3-го уровня (так, в «Женских работах по дому» маркирована тематика «Рукоделие»). Некоторые макротемы (ОБРАЗОВАНИЕ, ТЕХНИКА, ТРАНСПОРТ и др.) не членились на более частные в силу их факультативности в дискурсе.

Текстоцентрическая ориентированность Томского диалектного корпуса вызвала потребность в осуществлении нового типа разметки, отражающего характер организации текста в диалектном дискурсе. Он условно назван **разметкой по типам текста.**

Поскольку текстовые особенности диалектного дискурса еще недостаточно хорошо изучены, в ТДК на данном этапе производится маркирование наиболее определенно выделяемых в речевом потоке типов текста, имеющих регулярные вербальные маркеры. В перечень данной разновидности разметки включены:

а) указания на разновидности текста, различающиеся по степени спонтанности речевых проявлений:

– диалоги между диалектоносителями – самая типичная для диалектной коммуникации форма речи, демонстрирующая непринужденное речевое общение представителей народно-речевой культуры[1];

– ситуативные вкрапления**,** возникающие при отклонениях от целенаправленной беседы на определенную тему с диалектологами.

[1] Диалогическая речь редко фиксируется в условиях традиционного сбора диалектного материала, однако широко представлена при включении диалектолога в языковое существование членов сельского сообщества (см., например, тексты в сборнике «Живая речь русских старожилов Сибири» [16], где диалог явно преобладает над монологом).

Вкрапления могут представлять собой обращение к собирателям (*Что вам рассказать?; Кушайте, угощайтесь*), вербально выраженные отвлечения от рассказа на какое-либо событие (*Ой, внучка проснулась*) либо развернутые рассказы информантов при изменении темы беседы по их инициативе;

– метатексты как «вербализованные суждения о языке как результат осознания языковой действительности» [17. С. 55], встречающиеся и в спонтанной речи диалектоносителей, и при целенаправленном сборе материала, когда диалектологи в процессе общения с информантами эпизодически задают вопросы, касающиеся языка и речи;

– ответы на вопросники, отражающие главным образом сбор материала для толковых диалектных словарей с целью уточнения семантики и особенностей употребления единиц словарного состава среднеобских говоров. В отличие от предыдущего типа текстов это не единичные, а серийные метатекстовые фрагменты, часто имеющие общую тематическую направленность («Названия растений», «Обряды» и т.д.).

Два первых из перечисленных случаев представляют собой тексты, максимально близкие к естественной коммуникации, остальные – экспериментальные материалы. Вероятно, между теми и другими можно найти и разнообразные переходные случаи, но их выявление и типологизация – дело будущего;

б) речевые жанры. В первоначальной концепции ТДК предполагалась полная разметка материалов в соответствии с типологией жанров по характеру интенции (информативные, императивные, ритуальные и оценочные); в каждом случае указывался также конкретный жанр (в группу информативных включены разные виды сообщений, предположение, объяснение, жалоба, предупреждение; императивных – просьба, распоряжение, поручение, приказ, предложение, совет; ритуальных – приветствие, прощание, извинение, благодарность, приглашение, угощение, пожелание; оценочных – похвала, осуждение, самооценка, оценка). Кроме того, были выделены биографический рассказ, сюжетный

рассказ, описание, рассуждение, интервью, сказка, песня, частушка, пословица[1] [1. С. 61–62].

В уточненной концепции принято решение отказаться от детального маркирования всех речевых жанров, поскольку оно является сложной задачей в связи с недостаточно разработанной теорией генристики и наличием множества жанровых образований с комбинаторными характеристиками. Размечаются наиболее частотные речевые жанры:

– автобиографический рассказ, в свободной, неофициальной форме содержащий историю жизни информанта, рассказанную им самим;

– воспоминание – речевой жанр, отражающий в монологическом нарративе повествование о событиях человеческой жизни в прошлом;

– рассказы о других людях – вариант автобиографического рассказа с иным объектом повествования;

– рассказ о случае – повествование о ярком, экстраординарном жизненном событии, пережитом лично информантом или его близкими.

Разметке подлежали также значимые для народно-речевой культуры фольклорные жанры – вкрапления в устную бытовую речь частушек, сказок, примет, песен, пословиц, поговорок и др.

В перспективе, возможно, будут размечены оценочные речевые жанры, дающие богатую информацию для исследования мировосприятия и миропонимания носителей традиционных говоров.

Как и в случае тематического членения материала, используется «мягкая» разметка, допускающая отнесение того или иного текста более чем к одному типу (например, текст в речевом жанре «воспоминание» может быть маркирован также как «ситуативное вкрапление») и частичное наложение границ размечаемых текстов.

Параметры **лексической разметки** практически не разработаны как в общих, так и в диалектных корпусах. В НКРЯ введена семантическая разметка [18], но поиск по ней пока осуществляется только для литературных текстов; другие лексические параметры (напри-

[1] Представляется, что названные виды рассказа и интервью являются информативными либо информативно-оценочными, а повествование, описание и рассуждение находятся вне жанровой системы, традиционно рассматриваясь как функционально-смысловые типы речевого изложения.

мер, стилевая принадлежность лексем) не отражаются. Предоставление пользователям корпуса сведений о значениях областных слов, как отмечают И.Б. Качинская и Д.В. Сичинава, сдерживает отсутствие общедоступных электронных версий большинства диалектных словарей [6. С. 158].

Томичи располагают богатой словарной базой, репрезентирующей лексикон носителей говоров Среднего Прибья; сейчас начата работа по оцифровке опубликованных диалектных словарей. В ТДК в качестве первого шага на пути к лексической разметке планируется обеспечить возможность выдачи толкования значений нелитературных единиц, включенных в толковые дифференциальные среднеобские словари – трехтомный «Словарь русских старожильческих говоров средней части бассейна р. Оби» и четыре тома дополнений к нему (1964–1975 гг.). В дальнейшем, очевидно, необходимы расширение связки «корпус – словари» за счет отражения толкований из других словарных источников, введение функциональных характеристик лексем (устаревающее, новое, детское, грубое, бранное и др.) и апробация семантической разметки. Эта информация позволит использовать новый корпус как эффективный инструмент изучения диалектной лексики и создания новых лексикографических трудов.

Концепция нового корпусного проекта опирается на итоги работы нескольких поколений ученых томской диалектологической школы, принимавших активное участие в сборе диалектных материалов, составлении областных словарей, разработке различных аспектов изучения речи сибирских старожилов. Их труд – основа разрабатываемого корпуса, его фундамент. Создание задуманного ТДК, в свою очередь, будет стимулировать экспедиционную работу, способствовать совершенствованию лексикографической продукции, развитию новых направлений исследования народно-речевой культуры, задавая вектор деятельности школы в будущем.

Литература

1. *Юрина Е.А.* Томский диалектный корпус: в начале пути // Вестн. Том. гос. ун-та. Филология. – 2011. – №2 (14). – С. 58–63.

2. *Юрина Е.А., Толстова М.А.* Проект диалектного корпуса старожильческих говоров Среднего Приобья // Русская устная речь: материалы междунар. науч. конф. «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения» и межвуз. совещания «Проблемы создания и ис-

пользования диалектологических корпусов», Саратов, 15–17 ноября 2010 г. – Саратов, 2011. – С. 269–276.

3. *Трегубова Е.Н., Емельянова М.В.* Региональный лингвокультурологический корпус как электронный ресурс изучения народной аксиологии // Россия и славянский мир в контексте многополярности: материалы VII междунар. науч. конф., 6–9 августа 2010 г., Славянск-на-Кубани. – Ч. 2, разд. 2. – Славянск-на-Кубани, 2010. – С. 142–150. – URL: http://www.ethnolex.ru/2014-11-12-19-24-30/75-2014-10-09-20-33-55.html

4. *Кубрякова Е.С.* Эволюция лингвистических идей во второй половине XX века (опыт парадигмального анализа) // Язык и наука конца XX века. – М., 1995. – С. 144–238.

5. *Летучий А.Б.* Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005: Результаты и перспективы. – М., 2005. – С. 215–233.

6. *Качинская И.Б., Сичинава Д.В.* Диалектный подкорпус сегодня // Тр. Ин-та русского языка им. В.В. Виноградова РАН. – 2015. – № 6 (6). – С. 142–163.

7. *Трегубова Е.Н.* Многоуровневая тематическая разметка как инструмент этнолингвистической репрезентации диалектного дискурса в электронном текстовом корпусе // Вестн. Том. гос. ун-та. Филология. – 2015. – № 1 (33). – С. 66–77.

8. *Задумина П.Н.* О некоторых особенностях создания мультимедийного корпуса региональных текстов // Молодые исследователи – регионам: материалы междунар. науч. конф. – 2004. – Т. 3. – С. 194–196. – URL: http:// sno.vstu.edu.ru/wp-content/uploads/2014/09/t-3.pdf

9. *Крючкова О.Ю., Гольдин В.Е.* Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог–2011». – С. 359–367. – URL: http://www.dialog-21.ru/digests/dialog2011/materials/html/36.htm

10. *Русские* говоры Среднего Приобья / ред. В.В. Палагина. – Томск: Изд-во Том. ун-та, 1985–1989. – Ч. 1–2.

11. *Томская* диалектологическая школа: историографический очерк / под ред. О.И. Блиновой. Томск: Изд-во Том. ун-та, 2006. – 392 с.

12. *Словарь* русских старожильческих говоров средней части бассейна р. Оби / ред. В.В. Палагина. – Томск: Изд-во Том. ун-та, 1964. – Т. 1. – 143 с.

13. *Гольдин В.Е., Крючкова О.Ю.* Тематическая разметка и тематический анализ диалектного текстового корпуса // Языковая личность – текст – дискурс: теоретические и прикладные аспекты исследования: материалы междунар. науч. конф.: в 2 ч. – Ч. 1. – Самара, 2006. – С. 71–80.

14. *Буранова А.И.* Тематическая организация диалектной речи: квантитативный анализ // Изв. Сарат. гос. ун-та. – Нов. сер. – Сер. Филология и журналистика. – 2012. – Т. 12, вып. 3. – С. 35–38.

15. *Косицина Ю.В.* Статико-динамическая модель тематической организации диалектного монологического текста: автореф. дис. … канд. филол. наук. – Кемерово, 2013. – 26 с.

16. *Иванцова Е.В.* Живая речь русских старожилов Сибири: сб. текстов. – Томск, 2007. – 104 с.

17. *Ростова А.Н.* Метатекст как форма экспликации метаязыкового сознания (на материале русских говоров Сибири). – Томск: Изд-во Том. ун-та, 2000. – 194 с.

18. *Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л.* и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005: Результаты и перспективы. – М., 2005. – С. 193–214.

TOMSK DIALECT CORPUS: SUBSTANTIATION OF THE CONCEPT AND PROSPECTS OF DEVELOPMENT

Voprosy leksikografii – Russian Journal of Lexicography, 2017, 11, pp. 54–70.
DOI: 10.17223/22274200/11/4

Ekaterina V. Ivantsova, Tomsk State University (Tomsk, Russian Federation). E-mail: ekivancova@yandex.ru

Keywords: Russian dialects of Siberia, Middle Ob region, dialect corpus, concept.

The creation of a dialectal corpus is one of the topical problems for the Tomsk Dialectology School, the oldest research center for studying the folk speech culture of Siberia. The paper describes the general concept of the corpus, the substantiation of its purposes, the characteristics of the principles of meta-markup: the objectives of developing the new resource in the near and distant future are outlined.

The concept of the Tomsk Dialect Corpus is developed taking into account the key directions of the school on the study of folk speech that correlate with the achievements of the modern science of language, and with the nature of the materials available to dialectologists. The orientation of the new electronic resource can be defined as lexis- and text-centric.

The main form of representation of the Middle Ob dialects in the corpus is a text with an orthographic representation of separate features of oral speech. Reliance on this principle will allow to unify the representation of the diverse archive: from the first manuscript expedition notebooks to digital audio recordings of recent years. The chosen method of representation of the sounding dialect speech can be considered universal for lexicological, linguocultural, discursive and lexicographic research. Refusal from transcription is partly compensated by the possibility of accessing the existing audio records and scanned manual records of early expeditions.

The main types of meta-markup of texts entered into the corpus are developed: passport, thematic and markup by type of text.

Passport meta-markup includes extra-linguistic data about the texts entered in the corpus: instructions on the place and time of the recording, information about the informant, the type of recording (by hand / from the tape / recorder), the presence / absence of audio and video files, etc.

Thematic meta-markup is made on the basis of an inductive analysis of the discursive practice of old-timers, with the identification of particular topics and their generalization to macro-topics. Each topic is three levels deep maximum. The principle of "soft" thematic division of the fixed speech stream is used with the possibility of overlapping the boundaries of the extracted texts and/or simultaneous attribution of one fragment of the text to several topics.

Markup by type of text at this stage implies: a) indications of text varieties that differ in the degree of the spontaneity of speech manifestations (dialogues between dialect speakers, situational inclusions arising from deviations from a purposeful conversation with

dialectologists, episodic metatexts, answers to questionnaires); b) the most frequent speech genres (autobiographical story, recollection, stories about other people, stories about an event, folklore genres).

The first step on the way to lexical marking will be an opportunity to give an interpretation of the meaning of nonliterary units included in the differential dictionaries of the Middle Ob region.

Prospects for the development of the corpus include development of the indicated types of meta-markup, introduction of lexical markup, the integration of its data with the created electronic library of dialect dictionaries and other auxiliary resources.

References

1. Yurina, E.A. (2011) Tomsk dialectal corpora: the starting point. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 2 (14). pp. 58–63. (In Russian).

2. Yurina, E.A. & Tolstova, M.A. (2011) [Project of the dialect corpus of old-timer dialects of the Middle Ob region]. *Russkaya ustnaya rech'* [Russian oral speech]. Proceedings of the conference. Saratov. 15–17 November 2010. Saratov: Saratov State University. pp. 269–276. (In Russian).

3. Tregubova, E.N. & Emel'yanova, M.V. (2010) [Regional linguocultural corpus as an electronic resource of studying national axiology]. *Rossiya i slavyanskiy mir v kontekste mnogopolyarnosti* [Russia and the Slavic world in the context of multipolarity]. Proceedings of the conference. Slavyansk-na-Kubani. 6–9 August 2010. Vol. 2:2. Slavyansk-na-Kubani. pp. 142–150. [Online] Available from: http://www.ethnolex.ru/2014-11-12-19-24-30/75-2014-10-09-20-33-55.html. (In Russian).

4. Kubryakova, E.S. (1995) Evolyutsiya lingvisticheskikh idey vo vtoroy polovine XX veka (opyt paradigmal'nogo analiza) [Evolution of linguistic ideas in the second half of the 20th century (experience of paradigm analysis)]. In: Stepanov, Yu.S. (ed.) *Yazyk i nauka kontsa XX veka* [Language and science of the end of the twentieth century]. Moscow: RSUH.

5. Letuchiy, A.B. (2005) Korpus dialektnykh tekstov: zadachi i problemy [Corpus of dialect texts: tasks and problems]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian language: 2003–2005. Results and prospects]. Moscow: Indrik.

6. Kachinskaya, I.B. & Sichinava, D.V. (2015) Dialektnyy podkorpus segodnya [Dialect subcorpus today]. *Trudy instituta russkogo yazyka im. V.V. Vinogradova RAN*. 6 (6). pp. 142–163.

7. Tregubova, E.N. (2015) Multilevel thematic marking as an ethnolinguistic tool of dialectal discourse representation in digital text corpora. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 1 (33). pp. 66–77. (In Russian).

8. Zadumina, P.N. (2004) [On some features of creating a multimedia corpus of regional texts]. *Molodye issledovateli – regionam* [Young researchers to regions]. Proceedings of the conference. [Online] Available from: http://sno.vstu.edu.ru/wp-content/uploads/2014/09/t-3.pdf. (In Russian).

9. Kryuchkova, O.Yu. & Gol'din, V.E. (2011) Korpus russkoy dialektnoy rechi: kontseptsiya i parametry otsenki [The Corpus of Russian dialect speech: the concept and

parameters of evaluation]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computer Linguistics and Intellectual Technologies]. Proceedings of the international conference "Dialog–2011". pp. 359–367. [Online] Available from: http://www.dialog-21.ru/digests/dialog2011/materials/html/36.htm. (In Russian).

10. Palagina, V.V. (ed.) (1985–1989) *Russkie govory Srednego Priob'ya* [Russian dialects of the Middle Ob region]. Vols 1–2. Tomsk: Tomsk State University.

11. Blinova, O.I. (ed.) (2006) *Tomskaya dialektologicheskaya shkola: Istoriograficheskiy ocherk* [Tomsk School of Dialectology: A Historiographical Sketch]. Tomsk: Tomsk State University.

12. Palagina, V.V. (ed.) (1964) *Slovar' russkikh starozhil'cheskikh govorov sredney chasti basseyna r. Obi* [Dictionary of Russian old-timer dialects of the middle part of the basin of the river Ob]. Vol. 1. Tomsk: Tomsk State University.

13. Gol'din, V.E. & Kryuchkova, O.Yu. (2006) [Thematic markup and thematic analysis of the dialectal textual corpus]. *Yazykovaya lichnost' – tekst – diskurs: teoreticheskie i prikladnye aspekty issledovaniya* [Language personality – Text – Discourse: Theoretical and Applied Aspects of the Study]. Proceedings of the international conference. Vol. 1. Samara. pp. 71–80. (In Russian).

14. Buranova, A.I. (2012) Tematicheskaya organizatsiya dialektnoy rechi: kvantitativnyy analiz [Thematic organization of dialect speech: quantitative analysis]. *Izvestiya Saratovskogo gos. un-ta. Novaya seriya. Seriya Filologiya i zhurnalistika – Izvestiya of Saratov University. New Series. Series: Philology. Journalism*. 12:3. pp. 35–38.

15. Kositsina, Yu.V. (2013) *Statiko-dinamicheskaya model' tematicheskoy organizatsii dialektnogo monologicheskogo teksta* [A static-dynamic model of the thematic organization of the dialect monological text]. Abstract of Philology Cand. Diss. Kemerovo.

16. Ivantsova, E.V. (2007) *Zhivaya rech' russkikh starozhilov Sibiri: sbornik tekstov* [Live speech of Russian old-timers in Siberia: a collection of texts]. Tomsk: Tomsk State University.

17. Rostova, A.N. (2000) *Metatekst kak forma eksplikatsii metayazykovogo soznaniya (na materiale russkikh govorov Sibiri)* [Metatext as a form of explication of metalanguage consciousness (on the basis of Russian dialects of Siberia]. Tomsk: Tomsk State University.

18. Apresyan, Yu.D. et al. (2005) Sintaksicheski i semanticheski annotirovannyy korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy [Syntactically and semantically annotated corpus of the Russian language: current state and prospects]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian language: 2003–2005. Results and prospects]. Moscow: Indrik.