УДК 81'33+811.161.1.28
DOI: 10.17223/19986645/67/3

С.С. Земичева, Е.В. Иванцова

ТЕМАТИЧЕСКАЯ РАЗМЕТКА ДИАЛЕКТНОГО КОРПУСА: ОПЫТ ТОМСКИХ ДИАЛЕКТОЛОГОВ[9]

Анализируются проблемы разметки по предметным областям в новых электронных ресурсах, отражающих данные народной речи. Обозначены задачи такой разметки, выявлены факторы, осложняющие её осуществление, описаны разработанные создателями Томского диалектного корпуса методические шаги, снижающие степень субъективности этой процедуры. Практический опыт внедрения тематической разметки томскими диалектологами осмысляется на фоне решений, принятых разработчиками других диалектных корпусов и Национального корпуса русского языка.

Ключевые слова: *диалектный корпус, тематическая разметка, русские говоры Сибири.*

Подходы к осуществлению тематической разметки в национальных корпусах

Синтез лингвистики и компьютерных технологий привёл к созданию множества новых электронных ресурсов, среди которых особое место занимают лингвистические корпуса. Существуют как национальные корпуса многих языков мира, так и корпусные ресурсы других типов, их общее количество насчитывает более 3 000 и постоянно растёт [1. С. 21].

Одним из ключевых параметров создания лингвистического корпуса является разметка, отражающая разнообразную информацию о представленных в нем языковых единицах. Тематическая разметка при создании национального корпуса проектируется одной из первых и рассматривается как часть метаразметки (т.е. экстралингвистической, разметки по «внешним» для текста параметрам). Разработаны международные стандарты корпусной разметки, направленные на унификацию представления материалов, в частности стандарт EAGLES (European Advisory Group on Language Engineering Standards), включающий общий перечень параметров метаразметки, в том числе тематических областей, а также их возможных комбинаций [2].

В национальных корпусах представлены разные варианты реализации тематической разметки. Предпринятый С.О. Савчук сравнительный обзор принципов зарубежных электронных ресурсов и Национального корпуса русского языка (НКРЯ) показывает, что итоговый список тем НРКЯ характе-

[9] Исследование выполнено при поддержке гранта РФФИ № 19-012-00320 «Томский диалектный корпус как новый ресурс для изучения народно-речевой культуры».

ризуется неполным совпадением с тематическим перечнем EAGLES, отличаясь степенью обобщённости некоторых предметных областей [3. С. 73–74].

Сложной задачей при создании национальных корпусов оказалась тематическая разметка текстов отдельных сфер речи – в первую очередь устной повседневной коммуникации и беллетристики, поскольку предметные области таких текстов чрезвычайно разнообразны. В связи с этим для многих видов разнородных текстов, широко представленных в национальных корпусах, разметка по предметным областям фактически отсутствует: газетно-публицистические тексты на морально-этические и бытовые темы, а также личная переписка и повседневная устная речь согласно рекомендациям EAGLES отнесены в общую группу «Life» / «Частная жизнь»; вся художественная литература либо также маркируется этой темой, либо, как в НКРЯ, не размечается по темам вообще (вместо этого указывается лишь хронотоп – время и место описываемых событий).

Говоря о принципах тематической разметки, разработчики НКРЯ отмечали: «При построении корпуса не слишком важна глубина кодирования предметной области, затрагиваемой текстом <…> поскольку корпус не является универсальной энциклопедией. <...> При построении корпуса можно иметь грубую классификацию, выделяющую естественные и общественные науки, политику и экономику, искусство и досуг» [4. С. 14].

Очевидно, эта установка на «грубую» разметку предметных областей связана с тем, что она должна способствовать созданию представительного корпуса. Функции метаразметки НКРЯ (в состав которой входит и тематическая) заключаются в том, что она «1) служит для формирования архитектуры корпуса; 2) позволяет контролировать процесс его пополнения; 3) обеспечивает возможность поиска текстов пользователями, составления подкорпусов с заданными параметрами» [3. С. 62].

Таким образом, тематическая разметка в национальных корпусах базируется на дедуктивном подходе и относительно унифицирована, хотя и отличается деталями. Однако с внедрением корпусной лингвистики в сферу диалектологии возникает вопрос о принципах и способах осуществления тематической разметки диалектного материала.

Задачи, проблемы, принципы тематической разметки в диалектном корпусе

Первый электронный ресурс с текстами русских народных говоров был включён в НКРЯ в качестве одного из его модулей. При этом разработчики диалектного подкорпуса (далее ДпНКРЯ) опирались на общие подходы, выработанные в ходе реализации этого проекта. В их числе были заявлены ориентация на морфологическую разметку и выдача для пользователей только кратких фрагментов текста [5].

Новый подход к отражению данных народно-речевой культуры в лингвистическом корпусе был предложен саратовскими лингвистами. Исследователи исходили из идеи В.Е. Гольдина о своеобразии диалектной речи,

репрезентирующей традиционную систему русского деревенского общения как особого коммуникативного феномена [6]. Такая установка стала точкой отсчёта при разработке концепции корпуса, представляющего собой модель традиционной сельской коммуникации на диалекте [7]. Она воплотилась в проекте мультимедийного Саратовского диалектологического корпуса [8], привела к частичному видоизменению концепции ДпНКРЯ [9], а также инициированию ряда проектов по созданию диалектных корпусов отдельных территорий. В их числе – Кубанский диалектный корпус [10, 11], диалектный корпус лингвокультуры Северного Приангарья [12], корпус народной речи Среднего Прииртышья [13] и др. С 2010 г. на основе архива экспедиционных записей среднеобского бассейна создаётся Томский диалектный корпус (ТДК).

Фокусировка таких корпусов на текстах как отражении диалектной коммуникации и народной лингвокультуры вызвала пересмотр принципов выдачи экспедиционных материалов и функций тематической разметки. Доступ к текстовой базе стал вариативным с учётом запроса пользователя: от небольшого фрагмента до полного текста. Что касается функций маркирования предметных областей, то из области метаразметки (экстралингвистической) разметка по темам сдвигается в область собственно лингвистическую. Отмечается, что при моделировании диалектной коммуникации «важно учесть и реальное тематическое разнообразие речи, и хотя бы примерное количественное соотношение различных тем и жанров в континууме сельского речевого общения» [7. С. 72]. Текстовая разметка в электронном корпусе начинает также рассматриваться в качестве инструмента этнолингвистической репрезентации диалектного дискурса [10]. С опорой на этот вид разметки изучаются трансформация тематики диалектного дискурса во времени [14], соотношение количества тематических фрагментов в разных социолингвистических группах информантов [15]. Предприняты первые попытки привлечения корпусных материалов для анализа диалектной концептосферы [16–19], в том числе сравнения репрезентации концепта в диалектном и литературном корпусах [20]. При создании новых словарей лингвокультурологического типа [21, 22] также привлекались материалы диалектного дискурса с учётом тематической выборки.

Одной из главных проблем при осуществлении тематической разметки можно считать субъективность: «Определение тематики текста имеет субъективный характер… <…>. Трудно или даже невозможно составить идеальный перечень тематических областей» [3. С. 73].

Факторы, порождающие неточности и разночтения при разметке диалектных текстов по предметным областям, многочисленны и разнообразны.

К числу таких факторов можно отнести специфику диалектного дискурса, который характеризуется устным характером коммуникации, тесной взаимосвязью диалектного текста с ситуацией и более широким культурным контекстом его существования [23]. Если ручная запись (для ранних экспедиций) или расшифровка аудиофайлов не содержит восполнения ситуативных и культурных лакун, а выделение тематических фрагментов

осуществляется лицами, не участвовавшими в сборе материала, это нередко порождает сложности при осмыслении текста и его членении по темам.

Ещё одним фактором, осложняющим разметку, являются часто встречающиеся в устной речи переходы говорящего от темы к теме, тематические обрывы, наложения и/или пересечения двух и более тем.

Следует отметить также отсутствие единой методики осуществления тематического членения диалектных текстов, хотя в данной сфере имеются некоторые наработки. Так, Ю.В. Косициной на материале говоров Кемеровской области была создана модель тематической организации диалектного монологического текста [24], а А.И. Буранова на материале саратовских говоров обратилась к квантитативному анализу тематической организации диалектной речи [25]. Однако в первом исследовании моделирование носит теоретический характер и автор ставит своей целью выявление эмоционально-смысловых типов доминант текста. В связи с этим список выделенных тем ограничен, они формулируются предельно обобщённо («Жизненные трудности», «Жизнь села») и конкретные методические шаги по определению темы текста не прописаны. Во второй из названных работ были выделены ключевые слова в диалектных текстах и проанализирован количественный состав образуемых ими лексико-тематических групп. Такая методика представляется продуктивной, но группировка единиц в понятийные (тематические) области происходит на уровне слова, а не текстового фрагмента, т. е. фактически предполагается осуществление семантической, а не тематической разметки[10].

Преодоление обозначенных трудностей – сложная задача, решением которой в перспективе можно считать выработку методики тематической разметки, оптимальной для диалектных корпусов. Достижение этой цели не снимет полностью элемент субъективности в выделении тем, но позволит значительно снизить его «градус».

Пока можно говорить только о некоторых общих приёмах, применяемых на практике разработчиками областных корпусов при выделении предметных областей. Близки к ним и установки создателей Томского диалектного корпуса.

Во-первых, разметка осуществляется вручную. Она носит сплошной характер – размечаются все без исключения тексты.

Во-вторых, разметка осуществляется на уровне отдельных текстовых фрагментов. Тема всего текста не маркируется, поскольку он, как правило, «многотемен».

В-третьих, во всех рассмотренных диалектных корпусах используется «мягкая» разметка с возможностью присвоения одному и тому же фрагменту нескольких тематических меток и частичным наложением текстовых фрагментов.

[10] Основное отличие заключается в уровне разметки и маркируемых единицах: семантической разметкой принято называть разметку на уровне отдельного слова [26], тематической – на уровне текста или текстового фрагмента.

В то же время многие проблемы маркирования диалектных текстов по предметным областям ещё не имеют однозначных вариантов решения. К их числу относятся, прежде всего, состав тем и степень детализации тематического перечня.

Количество выделяемых в диалектных корпусах тем очень существенно различается. При этом число выделенных тем коррелирует с объёмом корпуса. Сведений об объёме Саратовского диалектологического корпуса в нашем распоряжении не имеется, в других случаях прослеживается определённая тенденция: чем больше объём корпуса, тем больше наименований включает тематический перечень. Из названных выше корпусов самый небольшой – диалектный корпус лингвокультуры Северного Приангарья (170 813 словоупотреблений), он же насчитывает минимальное число тем (39); ДпНКРЯ имеет несколько больший объём (285 281 слово), и его список тем полнее (58), Томский диалектный корпус имеет самый большой объём (1 787 416 словоупотреблений), тематический список насчитывает 77 наименований. Этот перечень не исчерпывающий и в ходе работы постепенно пополняется.

Касаясь состава тем, можно отметить, что многие выделяемые темы в созданных корпусах совпадают, но нередко имеются различия в формулировках их названия: «Работа» (приангарский корпус, ТДК) – «Трудовая деятельность» (саратовский корпус) – «Трудовая деятельность. Работа» (ДпНКРЯ); «Великая Отечественная война» (приангарский, ТДК) – «Война» (ДпНКРЯ); «Игры и развлечения» (приангарский) – «Развлечения» (саратовский) – «Досуг. Развлечения. Игры» (ДпНКРЯ) – «Досуг» (ТДК); «Семья» (саратовский, приангарский) – «Семья. Семейные отношения» (ДпНКРЯ) – «Семья и родственники» (ТДК) и т.д.

В некоторых случаях отмечаются уникальные темы, обусловленные региональными особенностями материала: так, только в корпусе лингвокультуры Северного Приангарья обозначены «Молевой сплав», «Сбор живицы» и «Богучанская ГЭС»; только в ТДК есть тема «Заготовка кедрового ореха». В других случаях выделение той или иной темы обусловлено интересом участников полевых экспедиций к определённым предметным областям, концепцией создателей корпуса, предполагающей степень детализации тем и принципы их маркирования. В ДпНКРЯ, например, много внимания уделено обрядности и мифологической составляющей народной культуры. В ТДК выделены темы, отражающие специфику устной диалектной коммуникации, – «О себе», «Прошлое и настоящее», отсутствующие в других диалектных корпусах. В числе прочих имеет место и человеческий фактор, усиливающий неоднородность разметки. Например, лишь в Саратовском диалектологическом корпусе выделена тема «Пьянство и наркомания», только в ДпНКРЯ – «Астрономия», «Рекрутский обряд и проводы в армию», «Народный этикет».

По-разному решается в областных корпусах и задача упорядочения выделенного списка тем. При обсуждении концепции диалектного корпуса В.Е. Гольдиным и О.В. Крючковой предлагался следующий подход: «Предметная специфика диалектного текста обусловливает целесообразность его двухуровневой тематической разметки – широкой и узкой. Ши-

рокую разметку имеет смысл максимально приблизить к тематическому кодированию, применяемому в поиске на массивах письменных текстов, что обеспечит сопоставимость различных корпусов и включённых в них текстов. Узкая разметка должна отражать тематическую структуру широкой предметной области, выявляя её специфику. Узкая тематизация послужит также базой для лексико-семантических и когнитивных исследований диалектной речи» [7. С. 73]. Этот принцип отчасти был воплощён в ДпНКРЯ: при полном или частичном совпадении общих тем (НКРЯ: «Природа» – ДпНКРЯ: «Природа», НКРЯ: «Искусство и культура» – ДпНКРЯ: «Духовная культура», НКРЯ: «Религия» – ДпНКРЯ – «Духовная культура. Религия», НКРЯ: «Здоровье и медицина» – ДпНКРЯ: «Здоровье и медицина. Народная медицина» и др.) диалектный подкорпус детализирует их через частные темы. Например, в теме «Природа» выделяются «Астрономия», «Животный мир», «Ландшафт», «Метеорологические явления, погода», «Растительный мир», в «Здоровье и медицина…» – «Болезни», «Заговоры», «Роды», «Смерть». В ТДК также использовался принцип двухуровневой разметки: выделялись макротемы «ПРИРОДА», «РАБОТА», «БЫТ» и т.п., а в их составе – частные темы, однако в других диалектных корпусах тематическая разметка является одноуровневой.

Одна из проблем, возникающих при осуществлении тематической разметки диалектного дискурса, – отграничение её от смежных типов разметки, в частности разметки концептов (имеется в корпусе Северного Приангарья, Кубанском диалектном корпусе) и жанровой разметки (представлена в ДпНКРЯ, саратовском и приангарском корпусах).

Так, в корпусе лингвокультуры Северного Приангарья перечни тем и концептов представляют собой два отдельных, но значительно пересекающихся списка, по каждому из которых возможен поиск [12], а в Кубанском диалектном корпусе концепты встраиваются в тематическую разметку как более частные её уровни. Например, в макротеме «обрядовая культура» выделяется тема «свадебный обряд», а внутри неё в конкретном тексте – макроконцепты («деятель», «время», «локус» и др.) и микроконцепты («дружки», «жених», «суббота», «дом», «воскресенье») [10]. К сожалению, ни в том, ни в другом проекте принципы разграничения тем и концептов не описаны.

В рассмотренных диалектных корпусах есть и некоторые следы пересечения тематической разметки с жанровой: в ДпНКРЯ в список тем включены «Фольклор», «Заговоры» и «Народный этикет», в саратовском корпусе есть тема «Обряды, обычаи, приметы», а также «Общая оценка жизни», в корпусе Северного Приангарья в тематический перечень включены «Автобиографические нарративы».

Приёмы осуществления тематической разметки в Томском диалектном корпусе

На первом этапе реализации проекта детализированная тематическая разметка в ТДК являлась центральной. В настоящее время тематическая разметка (наряду с жанровой) входит в текстовый модуль ТДК, который

рассматривается как один из основных, но не единственный (запланированы также модуль грамматической разметки, лексикографический и историко-географический модули).

Тематическая разметка ТДК осуществлялась в 2017–2020 гг. сотрудниками лаборатории общей и сибирской лексикографии и кафедры русского языка ТГУ А.А. Васильченко, Л.А. Ивановой, В.В. Галаниной, Н.А. Зюзьковой под руководством С.С. Земичевой. Исходная концепция корпуса, предложенная Е.А. Юриной [27] и более детально обоснованная Е.В. Иванцовой [28], новым коллективом молодых учёных была значительно трансформирована и расширена: в проект вошли новые модули ресурса и новые виды разметки; электронный корпус стал интенсивно наполняться. На данный момент Томский диалектный корпус насчитывает 1679 текстов, разбитых на более чем 20 000 тематических фрагментов, и является самым репрезентативным ресурсом такого рода в России.

Далее опишем методику осуществления тематической разметки, реализуемую в настоящее время его составителями.

Объектом разметки выступает текст как одномоментно сделанная собирателем запись речи диалектоносителя.

Задача первого этапа разметки – фрагментировать целостный текст. Разбиение на фрагменты, объединённые общим содержанием, происходит интуитивно. Границы тематических фрагментов определяются на этом этапе предварительно. Минимальной единицей разметки выступает высказывание. Рекомендуется выделять тему в тех случаях, когда она объединяет два и более высказывания.

На втором этапе выделяются ключевые слова фрагмента для выбора темы. «Тема текста находит своё выражение в референтно или сигнификативно объединённых группах лексики в его составе – в тематических группах, совокупность которых составляет текстовое поле тематической целостности» [29. С. 21]. На статус ключевого слова[11] могут претендовать единицы, которые встречаются в этом текстовом фрагменте несколько раз, сопровождаются однокоренными словами, выступают референтом местоимения, включаются в синонимические или родовидовые отношения с другими единицами.

Проиллюстрируем сказанное примером: *Здесь работали в колхозе, а родители весь век работали здесь в колхозе, но и в колхозе было работать чажело'. От, неправильно сделали, кто работал на производстве и кто в колхозе. Щас колхоз живёт богато. Щас в Красноярке придёшь – скотина ходит, свиньи, всё заполнено скотиной. В совхозе больше дают пенсию. Я уже двенадцать лет на пенсии, я в военкомате работала. Щас все совхозы, колхозы богато стали жить. И работать же легше тепе'ря.* (Зырянское, 1979). В данном фрагменте можно выделить ключевое слово *колхоз,* повторяющееся 6 раз, и существительное *совхоз*, повторяющееся

[11] При определении критериев ключевого слова использовались имеющиеся наработки, представленные в исследованиях Ю.В. Косициной [24], Т.В. Матвеевой [25] и др.

2 раза, что дало возможность выделить тему «Колхозы и совхозы». Внутри этого фрагмента наблюдаются тематические включения двух других тем – «О себе» (*Я уже двенадцать лет на пенсии, я в военкомате работала*) и «Страны, города, сёла» (*Щас в Красноярке придёшь – скотина ходит, свиньи, всё заполнено скотиной*).

Подспорьем в выделении тем является опора на списки ключевых слов, представленные в инструкции для разметчиков ТДК. Так, для темы «Выращивание животных» в состав маркеров входят следующие единицы (приведены в алфавитном порядке): *бо'тало, водопой, вымя, выпас, высокодойная, доенье, доить, доиться, дойная, доярка, за'пуск, инкубатор, кастрировать, кастрированный, комбикорм, колоть, конюх, молозиво, ночное, пастбище, пасти, пастись, пастух, подоить, поить, пойло, поярка, надоить, надаивать, низкодойная, осеменять, отдаивать, свинарка, сдаивать, стадо, стайка, табун, телиться, убой, фи'рма/ферма*[12]. Список ключевых слов открыт и может пополняться (в инструкцию включались, прежде всего, однозначные слова и единицы, связанные только с одной конкретной темой, а не с несколькими).

Сравним два текстовых фрагмента с учетом их ключевых слов.

Было по двадцать пять свиноматок. Я за откормом ходила, и за… Ро'стим, вырастим поросят… я как-то вырастила хорошего… поросята обошли по двадцать два килограмма всем весом. Ну вот пове'шали там… хорошо денег получила, на деньги работали. <…> Потом на почётной доске была, поросят вырастила. А как вырастила? Брала муку в колхозе да, и таскала домой, и в своей печке пекла хлеб, чтобы поросят-то хорошенько выкормить, чтобы хлебом их подкормить надо. А потом на фи'рме сделали… фи'рма у нас там была. На фе'рма там сделали русскую печку, и вот мы по очереди… Сёдня я хлеб стряпаю поросятам, на второй день втора' свинарка стряпает, настряпает булок десять… <…> Кормили хлебом, чтобы поросят вырастить. Вот и дешёво было мясо. Работали, трудились очень даже, очень. И щас работают, конечно, люди, вся'ко работают (Алаево, 2008).

Свинья, она родилась свиньёй, так и есть. Называют и «чушка», это матку, а боровок – мужука'. А «хряком» у нас не зовут. Чаще зовут примерно «чушка», это когда одна чушка, а когда много держишь – то свиньи (Томская обл., 1979).

В обоих текстах можно выделить ключевое слово *свинья*, его гипонимы (1-й текст: *поросёнок, свиноматка*; 2-й: *матка, боровок, хряк*) и синоним (*чушка*). В первом случае выстраивается ряд ключевых слов *свиноматка, откорм, ро'стить, вырастить, кормить, выкормить, поросята, фи'рма/ферма, стряпать, свинарка, работать*. Акцент в рассказе информанта делается на уходе за животными, что позволяет отнести данный фрагмент к теме «Выращивание животных» (макротема «РАБОТА»). Во втором случае ряд ключевых слов выглядит иначе: *свинья, чушка, боровок, хряк,*

[12] Через слеш даны формальные варианты одной лексемы.

называть, звать; представляется возможным выделить тему «Домашние животные» (с одновременным наложением темы «Язык и речь»).

В некоторых случаях, однако, опоры на принцип ключевых слов оказывается недостаточно. Разметчику необходимо осуществлять поиск смысловой доминанты текста, выявлять акценты, сделанные говорящим.

Сравним два текста, сходных по содержанию.

У нас в деревне дак воруют! Вот эти дачники, избушки. Громят только так. Вот у одной семьи уташшыли всю технику. И пахать была земли, это, землю какой-то там трактори'шка. И чё-то ешо, у их и пила была электро и… хапану'ли всё на свете. Да и не только у их. [Со двора прям вынесли?] Да, да. Да не со двора. Это, конечно, было наверно там де-нить в кладовке или как ли. (Вершинино, 2013).

А то, есть, в карман лезут. Да что говорить. Много молодежи сидит сейчас. А тода' ведь не сидели. Такие вот. Холостяки. Никода' ни слуху, ни духу не было, чтоб он в тюрьме сидел. А сейчас отчего в тюрьме? В город пойдёт учиться. Он там один, хозяин сам себе. Разбалывается в артеле там. И девки таки' есть вольные. Да что говорить. Ой, рассказывать это. Раньше вот. Попутал в кармане. Я его в сторону отвёл, чтоб народ не мешал мне. Как дам ему, и пустил. Он ушёл и больше не лезет никому. Боится. А другой не поверит первому побою, обратно залез. Ещё пуще получит. А сейчас лезет тебе в огород и в карман залезет, ворует. И ещё пойдёт жаловаться; ну, конечно, я не боюсь, я знаю, что говорить. (Зырянское, 1988).

Отметим, что выделение ключевых слов в данных фрагментах затруднительно. В первом случае это могут быть *воровать, громить, хапануть,* во втором – *(за)лезть в карман, тюрьма, сидеть (в тюрьме), воровать,* что позволяет выделить тему «Воровство». В тематическом перечне ТДК такая тема отсутствует. С точки зрения идеографической классификации «Воровство» является частью темы «Криминал». Как представляется, данная тема может быть выделена в первом тексте, так как там речь идёт только о материальном ущербе. Однако во втором случае превалирует установка говорящего на этическое осуждение такого поведения молодёжи, отсутствуют детали конкретного преступления, текст носит условно-обобщённый характер, что позволяет отнести его к теме «Мораль».

Не всегда в текстовом отрезке возможно выделить ключевые слова по описанным выше критериям. Рассмотрим ещё один фрагмент текста: *Ну, ой, ну от упала! Сломала шейку этого, бедра. И вот теперь, и вот теперь вот такая вот. И вот три года, девочки, лежу вот на этой уже койке. Ну я говорю ешшо, спасибо, вот, маленько, на ведро встаю, ауа* (Колпашево, 2019). В данном случае нет повторов или синонимических рядов лексем, гиперо-гипонимических отношений, однако выделяется ряд слов и словосочетаний, принадлежащих к общему семантическому полю (*упасть, шейка бедра, лежать, вставать*), что позволяет выделить макротему «ЧЕЛОВЕК ФИЗИЧЕСКИЙ».

На третьем этапе работы разметчик, проанализировав содержание текстового отрезка, окончательно определяет границы фрагмента и выбирает

подходящую тему из имеющегося перечня. При этом необходимо иметь в виду следующие нюансы.

Во-первых, при осуществлении разметки необходимо помнить об иерархическом устройстве тематического списка, учитывая не только тему, но и макротему. Поэтому выделение, скажем, темы «Местность» во фрагментах, посвящённых описанию частей села, было бы ошибочным, так как данная тема входит в состав макротемы «ПРИРОДА».

Во-вторых, макротемы в ТДК пополняются контекстами по остаточному принципу. Текстовый фрагмент относится к макротеме также в тех случаях, когда в коротком отрезке текста (1–2 предложения) упоминаются ключевые слова, которые можно отнести к разным темам общей макротемы: *Я и на лошади ездила, навоз возила. На пашне, раньше же пашни были, это называли пашни, там сеяли, пахали, сеяли* (Тогур, 2016). Одновременное наличие в контексте обозначений разных видов трудовой деятельности (*возила, пахали, сеяли*) позволяет причислить его к макротеме «РАБОТА». Кроме того, фрагмент может маркироваться как относящийся к определённой макротеме, если определить частную тему не удаётся или она отсутствует в перечне. Например, контекст *Я к сыну хожу, там вымоюсь и домой ужинать иду. Казенная дале'ко баня* (Зырянское, 1988) причисляется к макротеме «БЫТ», поскольку тема «баня» в списке не представлена.

Время от времени проводится «ревизия» текстов внутри макротемы. Если при этом внутри размеченного текстового массива выделяется регулярно повторяющаяся в текстах тема, она приобретает статус самостоятельной и вносится в список. Такой «челночный» принцип тематической разметки позволяет брать за основу выделяемых в корпусе предметных областей материалы диалектного дискурса и постепенно расширять тематический перечень.

В-третьих, существуют внутригрупповые договорённости, которые необходимо учитывать. В ходе коллективной работы создатели корпуса принимают решения о том, к какой «крупной» теме отнести те или иные «мелкие» фрагменты. Так, при разметке ТДК описание русской печи решено включить в тему «Дом и усадьба» (а не к смежной теме «Домашние вещи»), описание женских украшений (серьги, бусы и др.) отнести к теме «Одежда и обувь» (а не выделять как отдельную тему), упоминания бартера – прямого обмена без использования денежных средств – к теме «Покупка и продажа», описание состояния дорог – к макротеме «ТРАНСПОРТ» и т.п. Принятые решения фиксируются в памятке, а совместное обсуждение частично помогает снять проблему субъективности разметки. Пользователь корпуса имеет возможность ознакомиться с данными допущениями, обращаясь к размещённой на сайте инструкции.

Технически тематическая разметка осуществляется следующим образом: разметчик выделяет в корпусе текстовый фрагмент, относящийся к определённой теме, затем выбирает соответствующую ему тему или макротему из выпадающего списка. В перечне сначала даётся макротема, затем все частные темы, относящиеся к ней, затем следующая макротема

и т.д. Для маркирования тем внутри макротемы используется графическое выделение, характерное для уровневых списков (абзацный отступ для макротемы, дефис с абзацного отступа для темы, двойной дефис после отступа для микротемы) (рис. 1).

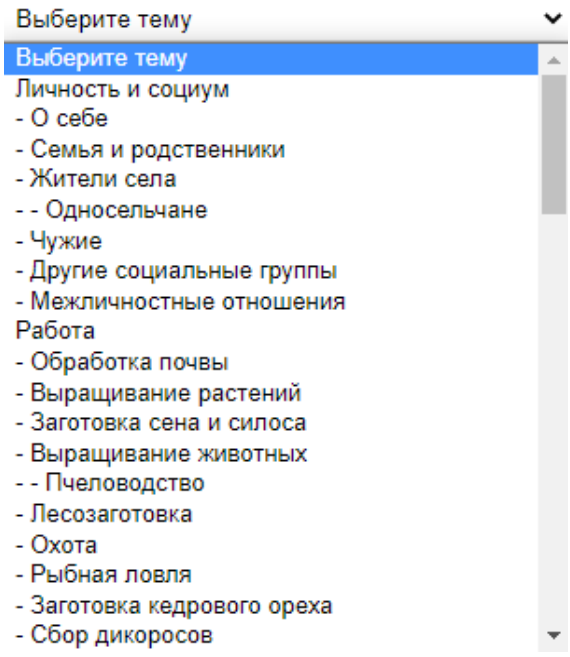

Рис. 1. Скриншот тематического списка (фрагмент)

В базе данных каждая тема и макротема имеет числовой идентификатор (тег). Теги помещаются в квадратные скобки, каждый из них состоит из латинской буквы и цифр. Латинские буквы указывают на тип разметки (t – тематическая разметка, k – разметка по типу текста, g – жанровая разметка). Различаются «открывающие» теги (начало темы) и «закрывающие» (конец темы), в последних используется дополнительно косая черта (слеш). Теги могут расставляться и вручную.

Размеченный текст внутри системы выглядит следующим образом: [t35] *Играли вечёрки, песни пели, в «суседи» играли. Король у их выбранный. Посадят девку на коленки к парню и говорят: «С коленки на коленку, повидаться помаленьку». Тапе'рь такого нету.* [/t35] – выделена тема «Досуг».

Взаимное наложение тем (присвоение нескольких тематических меток одному фрагменту) допустимо в тех случаях, когда тему текста невозможно определить однозначно. Технически возможность добавления нескольких тем не ограничена, но рекомендуется на одном и том же отрезке выделять не более трёх тем. Границы двух или нескольких тем могут полностью совпадать, при этом размеченный текст выглядит так: [t6] [t39]

У тунгусов копьё называется. С одной стороны как ножик. Оно как улета'т, втыка'т. Тунгусы или э'венки. [/t6] [/t39] – выделены темы «Язык и речь» [t39], «Чужие» [t6]. При наложении тем теги даются в произвольном порядке.

Темы могут совпадать частично, т.е. пересекаться: [t29] [t3] *У нас была така' семья, жили раньше плохо, не в чем было ходить в школу, братовья' были в школе, а я всё водилась с маленькими, не в чем было ходить, плохо жили.* [/t29] *Семья больша' была, девять человек живых осталось, а было тринадцать. У меня было девять человек детей, живых осталось пять.* [/t3]. В данном фрагменте более широкой является тема «Семья и родственники» – её границы обозначены тегом [t3], внутри фрагмента выделена тема «Условия жизни» – тег [t29].

Наконец, возможно включение инотематических фрагментов в состав той или иной темы: [t13] *[Давно скотину не держите?] Я скотину не стала держать, как дед умер, так не стала держать. [До этого держали?] До эт… ешо' после его я держала год три поросёнка.* [t3] *Дочке, сыну и себе, ну вырастила их; сын приехал, заколол, дочка взяла одного, они одного взяли, одного мне, а мне чё, куда мне одной поросёнка?! Я его половину им же отдала, ну а потом я говорю: «Санька, давай купим (дочка умерла), – я грю, – Санька, давай купим, это, два поросёнка, – я грю, – тебе поросёнка на твою се'мью, ну и мне поросёнка. Я, – грю, – дам внучатам маленько мяса», а он меня заругал: «Мама, гыт, тебе это надо? Ты всю жись со скотиной да в труде, да всё. Хватит, отдыхай».* [/t3] *Вот с тех пор я не держу, пять лет не держу никого, даже куриц не держу.* [/t13]. Внутри темы «Выращивание животных», обозначенной тегом [t13], выделена тема «Семья» [t3].

На странице текста в ТДК представлен перечень затронутых в нём тем (в порядке их появления в речи диалектоносителя), при выборе конкретной темы из списка соответствующий фрагмент целостного текста будет подсвечен (рис. 2).

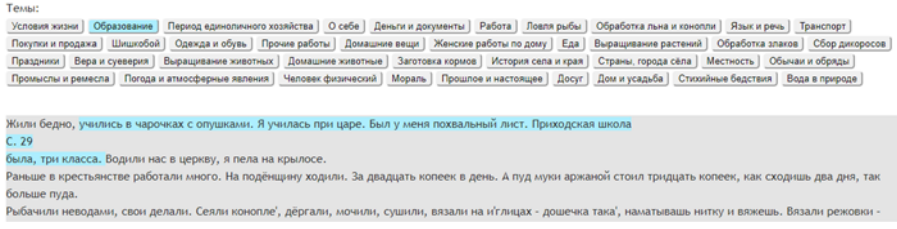

Рис. 2. Скриншот текста (фрагмент) с выделением конкретной темы

При поиске пользователь получает по запросу все тексты, где есть интересующая его тема. Пользователь видит тот же перечень, что и разметчик (с уровневым маркированием с помощью абзацных отступов и дефисов). При поиске по макротеме на данный момент можно найти только тексты, отнесённые к ней по остаточному принципу. В перспективе реали-

зация уровневой разметки предполагает, что на запрос макротемы при необходимости будут выдаваться все тексты, отнесённые как к макротеме собственно, так и к входящим в неё темам.

Таким образом, совершенствование диалектных электронных ресурсов, в том числе связанное с вопросами их тематической разметки, будет способствовать исследованию народно-речевой культуры с опорой на новые эффективные инструменты научного поиска. Разметка по темам в диалектных корпусах позволяет рассматривать её не столько как экстралингвистический, «внешний» по отношению к тексту (служащий для балансировки вводимых материалов), сколько как собственно лингвистический параметр, открывающий перспективы анализа содержательной специфики диалектного дискурса.

Представленный опыт создателей Томского диалектного корпуса может быть использован при разработке других корпусных ресурсов, а также в сфере теоретического изучения диалектной речи с позиций дискурсивного анализа.

Литература

1. *Копотев М.В.* Введение в корпусную лингвистику : учебное пособие для студентов филологических и лингвистических специальностей университетов. Прага : Anime-dia Company, 2014. 195 с.

2. Topic // EAGLES. Preliminary Recommendations on Text Typology. EAG---TCWG---TTYP/P. Version of Jun 1996. URL: http://www.ilc.cnr.it/EAGLES96/texttyp/node21.html#SECTION00070000000000000000 (дата обращения: 15.05.2020).

3. *Савчук С.О.* Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 62–88.

4. *Шаров С.А.* Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Серия «Информационные процессы и системы». 2003. № 6. С. 9–18. URL: http://lamb.viniti.ru/sid2/sid2free?sid2=J0338267X35 (дата обращения: 10.05.2020).

5. *Летучий А.Б.* Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. М., 2005. С. 215–232.

6. *Гольдин В.Е.* Теоретические проблемы коммуникативной диалектологии : дис. … д-ра филол. наук в виде науч. докл. Саратов, 1997. 52 с.

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

8. *Саратовский* диалектный корпус: новый научный и образовательный ресурс: Концепция, методические материалы / сост. Крючкова О.Ю., Гольдин В.Е. Саратов, 2010. 39 с.

9. *Качинская И.Б.* Диалектный подкорпус НКРЯ: Новый стандарт подачи. Новое рабочее место // Русская устная речь: материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения» и межвузовского совещания «Проблемы создания и использования диалектологических корпусов», Саратов, 15–17 ноября 2010 г. Саратов, 2011. С. 239–248.

10. *Трегубова Е.Н.* Многоуровневая тематическая разметка как инструмент этнолингвистической репрезентации диалектного дискурса в электронном текстовом корпусе // Вестник Томского государственного университета. Филология. 2015. № 1 (33). С. 66–77.

11. *Диалектный* корпус // Региональная этнолингвистика. URL: https://ethnolex.ru/kubdk/ (дата обращения: 12.03.2020).

12. *Диалектный* подкорпус. // Электронный текстовый корпус лингвокультуры северного Приангарья. URL: http://angara.sfu-kras.ru/?page=dialect# (дата обращения: 02.05.2020).

13. *Лавров Д.Н., Харламова М.А., Костюшина Е.А.* Представление разметки корпуса народной речи Среднего Прииртышья // Математические структуры и моделирование. 2018. № 4 (48). С. 85–91.

14. *Земичева С.С.* Новые темы диалектного дискурса (на материале Томского диалектного корпуса) // Труды международной конференции «Корпусная лингвистика-2019». СПб., 2019. С. 280–287.

15. *Земичева С.С.* Взаимосвязь тематики диалектного текста и пола говорящего (на материале Томского диалектного корпуса) // Актуальные проблемы и перспективы русистики: материалы по итогам Международной конференции русистов в Барселонском университете, 20–22 июня 2018. Barcelona, 2018. С. 491–500.

16. *Волошина С.В., Толстова М.А.* Репрезентация концепта «Богатство» в диалектном дискурсе: константы и трансформации // Вестник Томского государственного университета. Филология. 2018. № 55. С. 17–28.

17. *Демешкина Т.А.* «Ссылка» как феномен сибирской лингвокультуры // Вестник Томского государственного университета. Филология. 2018. № 56. С. 34–46.

18. *Демешкина Т.А.* Мир природы в зеркале диалекта (на материале концепта «Болото») // Вестник Томского государственного университета. Филология. 2019. № 62. С. 85–103.

19. *Смирнов Е.С.* Ценностные доминанты ангарцев в устных текстах о «своих» // Известия Волгоградского государственного педагогического университета. 2019. № 6 (139). С. 140–143.

20. *Иванцова Е.В.* Вариативность реализации ключевого концепта ХЛЕБ в разных типах русской речевой культуры // Актуальные проблемы и перспективы русистики: материалы по итогам Международной конференции русистов в Барселонском университете, 20–22 июня 2018. Barcelona, 2018. С. 1172–1181.

21. *Словарь* детства: говоры Среднего Приобья (с лингвокультурологическим комментарием) / под ред. М.М. Угрюмовой. Томск : Изд-во Том. ун-та, 2018. 200 с.

22. *Банкова Т.Б.* Словарь сибирского свадебного обряда. Томск : Изд-во Том. ун-та, 2018. Т. 1. 198 с.

23. *Крючкова О.Ю., Гольдин В.Е.* Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», Бекасово, 25–29 мая 2011 г. М., 2011. Вып. 10 (17). С. 359–367.

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

25. *Буранова А.И.* Тематическая организация диалектной речи : квантитативный анализ // Известия Саратовского университета. Новая серия. Сер. Филология. Журналистика. 2012. Т. 12, вып. 3. С. 35–38.

26. *Задачи* и принципы семантической разметки лексики в НКРЯ / Е.В. Рахилина [и др.] // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009. С. 215–239.

27. *Юрина Е.А.* Томский диалектный корпус: в начале пути // Вестник Томского государственного университета. Филология. 2011. № 2 (14). С. 58–63.

28. *Иванцова Е.В.* Томский диалектный корпус: обоснование концепции и перспективы развития // Вестник Томского государственного университета. Филология. 2017. № 11. С. 54–70.

29. *Матвеева Т.В.* Функциональные стили в аспекте текстовых категорий: синхронно-сопоставительный очерк. Свердловск : Изд-во Урал. ун-та, 1990. 170 с.

30. Текстовая разметка Томского диалектного корпуса // Томский диалектный корпус: Инструкция для пользователя. URL: http://losl.tsu.ru/sites/default/files/docs/Topics_result.docx (дата обращения: 15.06.2020).

Dialect Corpus Thematic Markup: The Experience of Tomsk Dialectologists

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2020. 67. 45–61. DOI: 10.17223/19986645/67/3

Svetlana S. Zemicheva, Ekaterina V. Ivantsova, Tomsk State University (Tomsk, Russian Federation). E-mail: optysmith@gmail.com / ekivancova@yandex.ru

Keywords: dialect corpus, thematic markup, Russian dialects of Siberia.

The study is supported by the Russian Foundation for Basic Research, Project No. 19-012-00320.

The article discusses an urgent problem in the field of corpus linguistics: the implementation of text markup by topic. It presents the experience of implementation of such a markup in the Tomsk dialect corpus. The research team has achieved impressive results: at the moment, the Tomsk dialect corpus contains 1,600 texts, divided into more than 20,000 thematic fragments, and is the most representative resource of this kind in Russia. The authors of the article interpret the practical experience of Tomsk researchers in a broad context against the background of decisions made by the developers of the Russian National Corpus and the emerging Russian dialect corpora. The authors identify factors that give rise to difficulties in the implementation of thematic markup of dialect texts by subject areas. The factors include: the oral nature of dialect communication (thus, a close connection of the text with the situation of its generation, overlapping or intersection of themes); in some cases, a weak degree of coherence of texts due to the peculiarities of fixing the material, difficulties in understanding the texts of local culture "from outside", lack of a unified methodology for thematic markup. An analysis of the available developments in the field of creating regional corpora makes it possible to identify general techniques used in practice. The techniques include: manual thematic markup; hierarchy of the thematic list, which generally includes two levels of generalization; markup of the topic of separate text fragments, not the text as a whole; use of "soft" markup with the ability to assign several thematic labels to the same fragment and partial overlap of text fragments. The developers of the Tomsk dialect corpus propose specific methodological steps to implement thematic markup. The markup includes 3 stages: the person doing the markup (1) intuitively breaks the text into fragments united by a common content and determines these fragments' boundaries preparatively; (2) determines the keywords of the fragment (based on the lists of keywords from the instructions) and, in some cases, the semantic dominant of the text; (3) identifies the final boundaries of the fragment and the choice of a topic from the available list. The list of topics for the markup in the Tomsk dialect corpus currently includes 77 items; it is not exhaustive and is gradually updated in the course of work. The potential content of the texts on each topic and the thematic belonging of the "controversial" fragments are determined as a result of group discussions. The user of the corpus can learn the details of these discussions by referring to the instructions posted on the website. The authors also briefly describe the technical side of thematic markup, provide samples of marked-up text fragments. The presented experience can be applied to create other corpus resources and used in the field of theoretical studies of dialect speech from the standpoint of discourse analysis.

References

1. Kopotev, M.V. (2014) *Vvedenie v korpusnuyu lingvistiku* [Introduction to Corpus Linguistics]. Prague: Animedia Company.

2. Istituto di Linguistica Computazionale "A. Zampolli". (1996) Topic. *EAGLES. Preliminary Recommendations on Text Typology. EAG---TCWG---TTYP/P*. [Online] Available from: http://www.ilc.cnr.it/EAGLES96/texttyp/node21.html#SECTION00070000000000000000. (Accessed: 15.05.2020).

3. Savchuk, S.O. (2005) Metatekstovaya razmetka v Natsional'nom korpuse russkogo yazyka: bazovye printsipy i osnovnye funktsii [Metatext markup in the Russian National

Corpus: basic principles and main functions]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003–2005. Results and Prospects]. Moscow: Indrik. pp. 62–88.

4. Sharov, S.A. (2003) Predstavitel'nyy korpus russkogo yazyka v kontekste mirovogo opyta [Representative corpus of the Russian language in the context of world experience]. *Nauchno-tekhnicheskaya informatsiya. Seriya "Informatsionnye protsessy i sistemy"*. 6. pp. 9–18. [Online] Available from: http://lamb.viniti.ru/sid2/sid2free?sid2=J0338267X35. (Accessed: 10.05.2020).

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* [Russian National Corpus: 2003–2005. Results and prospects]. Moscow: Indrik. pp. 215–232.

6. Gol'din, V.E. (1997) *Teoreticheskie problemy kommunikativnoy dialektologii* [Theoretical problems of communicative dialectology]. Philology Dr. Diss. Saratov.

7. Gol'din, V.E. & Kryuchkova, O.Yu. (2006) [Thematic markup and thematic analysis of the dialectal text corpus]. *Yazykovaya lichnost' – tekst – diskurs: teoreticheskie i prikladnye aspekty issledovaniya* [Linguistic Persona – Text – Discourse: Theoretical and Applied Aspects of Research]. Proceedings of the International Conference. Part 1. Samara. 3 October 2006. Samara: Samara State University. pp. 71–80. (In Russian).

8. Kryuchkova, O.Yu. & Gol'din, V.E (eds) (2010) *Saratovskiy dialektnyy korpus: novyy nauchnyy i obrazovatel'nyy resurs. Kontseptsiya, metodicheskie materialy* [Saratov Dialect Corpus: A new scientific and educational resource. Concept, methodological materials]. Saratov: [s.n.].

9. Kachinskaya, I.B. (2011) [Dialectal subcorpus of the RNC. New filing standard. New workplace]. *Russkaya ustnaya rech'* [Russian Oral Speech]. Proceedings of the International Conference "Barannikovskie chteniya. Ustnaya rech': russkaya dialektnaya i razgovorno-prostorechnaya kul'tura obshcheniya" [Barannikov Readings. Oral Speech: Russian Dialectal and Colloquial Culture of Communication] and Interuniversity Meeting "Problemy sozdaniya i ispol'zovaniya dialektologicheskikh korpusov" [Problems of Creating and Using Dialectological Corpora]. Saratov. 15–17 November 2010. Saratov: Nauka. pp. 239–248. (In Russian).

10. 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). DOI: 10.17223/19986645/33/6

11. Regional'naya etnolingvistika [Regional Ethnolinguistics]. (n.d.) *Dialektnyy korpus* [Dialectal corpus]. [Online] Available from: https://ethnolex.ru/kubdk/. (Accessed: 12.03.2020).

12. The Electronic Text Corpus of the Linguistic Culture of North Priangarye. (n.d.) *Dialektnyy podkorpus* [Dialect subcorpus]. [Online] Available from: http://angara.sfu-kras.ru/?page=dialect#. (Accessed: 02.05.2020).

13. Lavrov, D.N., Kharlamova, M.A. & Kostyushina, E.A. (2018) Representation of the corpus of medium Irtysh folk dialect. *Matematicheskie struktury i modelirovanie – Mathematical Structures and Modeling*. 4 (48). pp. 85–91. (In Russian). DOI: 10.25513/2222-8772.2018.4.85-91

14. Zemicheva, S.S. (2019) [New topics of the dialect discourse (based on the Tomsk dialect corpus material)]. *Korpusnaya lingvistika-2019* [Corpus linguistics-2019]. Proceedings of the International Conference. Saint Petersburg. 24–28 June 2019. Saint Petersburg: Saint Petersburg State University. pp. 280–287. (In Russian).

15. Zemicheva, S.S. (2018) [The relationship between the topic of the dialect text and the speaker's gender (based on the Tomsk dialect corpus)]. *Aktual'nye problemy i perspektivy rusistiki* [Actual Problems and Prospects of Russian Studies]. Proceedings of the International Conference Barcelona. 20–22 June 2018. Barcelona: Trialba Ediciones. pp. 491–500. (In Russian).

16. Voloshina, S.V. & Tolstova, M.A. (2018) Representation of the concept "Wealth" in the dialect discourse: constants and transformations. *Vestnik Tomskogo gosudarstvennogo*

universiteta. Filologiya. – Tomsk State University Journal of Philology. 55. pp. 17–28. (In Russian). DOI: 10.17223/19986645/55/2

17. Demeshkina, T.A. (2018) "Exile" as a phenomenon of the Siberian linguaculture. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. – Tomsk State University Journal of Philology*. 56. pp. 34–46. (In Russian). DOI: 10.17223/19986645/56/3

18. Demeshkina, T.A. (2019) The world of nature in the mirror of the dialect (a case study of the concept "swamp"). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. – Tomsk State University Journal of Philology.* 62. pp. 85–103. (In Russian). DOI: 10.17223/19986645/62/6

19. Smirnov, E.S. (2019) Value dominants of the Angara basin residents in oral texts about "locals". *Izvestiya Volgogradskogo gosudarstvennogo pedagogicheskogo universiteta – Ivzestia of the Volgograd State Pedagogical University*. 6 (139). pp. 140–143. (In Russian).

20. Ivantsova, E.V. (2018) [Variation in the implementation of the key concept of BREAD in different types of Russian speech culture]. *Aktual'nye problemy i perspektivy rusistiki* [Actual Problems and Prospects of Russian Studies]. Proceedings of the International Conference. Barcelona. 20–22 June 2018. Barcelona: Trialba Ediciones. pp. 1172–1181. (In Russian).

21. Ugryumova, M.M. (ed.) (2018) *Slovar' detstva: govory Srednego Priob'ya (s lingvokul'turologicheskim kommentariem)* [Dictionary of Childhood: Dialects of the Middle Ob region (with linguoculturological commentary)]. Tomsk: Tomsk State University.

22. Bankova, T.B. (2018) *Slovar' sibirskogo svadebnogo obryada* [Dictionary of the Siberian Wedding Ceremony]. Vol. 1. Tomsk: Tomsk State University.

23. Kryuchkova, O.Yu. & Gol'din, V.E. (2011) [A corpus of Russian dialectal speech: the concept and parameters of evaluation]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational Linguistics and Intellectual Technologies]. Proceedings of the International Conference "Dialog" [Dialogue]. 10 (17). Bekasovo. 25–29 May 2011. Moscow: Russian State University for the Humanities. pp. 359–367. (In Russian).

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

25. Buranova, A.I. (2012) Thematic Organization of Dialect Speech: Quantitative Analysis. *Izvestiya Saratovskogo universiteta. Novaya seriya. Ser. Filologiya. Zhurnalistika – Izvestiya of Saratov University. New Series. Series: Philology. Journalism*. 3 (12). pp. 35–38. (In Russian).

26. Rakhilina, E.V. et al. (2009) Zadachi i printsipy semanticheskoy razmetki leksiki v NKRYa [Tasks and principles of semantic markup of vocabulary in the RNC]. In: *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New Results and Prospects]. Saint Petersburg: Nestor-Istoriya. pp. 215–239.

27. 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).

28. Ivantsova, E.V. (2017) Tomsk dialect corpus: substantiation of the concept and prospects of development. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. – Tomsk State University Journal of Philology*. 11. pp. 54–70. (In Russian). DOI: 10.17223/22274200/11/4

29. Matveeva, T.V. (1990) *Funktsional'nye stili v aspekte tekstovykh kategoriy: sinkhronno-sopostavitel'nyy ocherk* [Functional Styles in the Aspect of Textual Categories: A Synchronic-Comparative Sketch]. Sverdlovsk: Ural Federal University.

30. Tomsk Dialect Corpus. (2019) Tekstovaya razmetka Tomskogo dialektnogo korpusa [Text markup of the Tomsk Dialect Corpus]. *Instruktsiya dlya pol'zovatelya* [User manual]. [Online] Available from: http://losl.tsu.ru/sites/default/files/docs/Topics_result.docx. (Accessed: 15.06.2020).