ТОМСКИЙ ДИАЛЕКТНЫЙ КОРПУС: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ[[1]](#footnote-1)

Земичева С.С. ([optysmith@gmail.ru](mailto:optysmith@gmail.ru))

Национальный исследовательский Томский государственный университет, Томск, Россия

TOMSK DIALECTAL CORPUS: PROBLEMS AND PROSPECTS

Zemicheva S.S. ([optysmith@gmail.ru](mailto:optysmith@gmail.ru))

National Research Tomsk State University, Tomsk, Russia

The article deals with the conception of the corpus representing the Russian dialectal speech of the Middle Ob region. The problems encountered while creating this resource are discussed. The first problem is transcribing the texts. Orthographic representation is chosen as the main principle. Secondly, the process of the topic markup is described and some quantitative results are given. The author suggested that thematic organization of the text reflects the specifics of dialect communication. The third problem discussed is representativeness of the corpus. The markup by the spontaneity level of the text is proposed. Prospects of improving the corpus are described. They include: new parameters of metamarkup, genre markup (which should reflect some pragmatic aspects of the communication, such as its intention), morphological tagging and the integration of the corpus data with the created electronic dialect dictionaries (semantic markup). Prospects for usage the resource are proposed.

Key words: dialect corpus, Tomsk dialect corpus, Siberian dialects of Russian

Томский диалектный корпус[[2]](#footnote-2) создаётся на материале архива записей диалектной речи, сделанных с конца 40-х гг. ХХ в. до настоящего времени в Томской и Кемеровской областях. Цель проекта – презентация текстов народно-речевой культуры в электронном формате с широкими поисковыми возможностями.

Проект соответствует общей тенденции цифровизации существующих диалектологических источников и создания новых электронных ресурсов на их основе. Диалектные корпуса созданы для многих европейских языков, среди них можно упомянуть болгарский диалектный корпус[[3]](#footnote-3), чешский диалектный корпус[[4]](#footnote-4), корпуса немецких диалектов в рамках Мангеймского проекта[[5]](#footnote-5), базу данных баварских диалектов, также созданную на материале немецкого языка[[6]](#footnote-6), эстонский диалектный корпус[[7]](#footnote-7), диалектный корпус скандинавских языков[[8]](#footnote-8), португальский синтаксически размеченный диалектный корпус[[9]](#footnote-9).

На материале русского языка наиболее известным является диалектный подкорпус в составе Национального корпуса русского языка[[10]](#footnote-10), который, несмотря на представленность разных типов говоров (севернорусских, среднерусских, южнорусских) не может считаться репрезентативным для страны в целом (включает материалы из 22 регионов объёмом 285 281 слово). Наиболее репрезентативным русским диалектным корпусом сегодня является корпус говоров бассейна р. Устья Архангельской области[[11]](#footnote-11). Он насчитывает более 864 словоупотреблений информантов и более 1 100 000 всего (с учётом речи собирателей).

Создание диалектного корпуса предполагает решение ряда проблем на стыке корпусной лингвистики и диалектологии.

1. **Проблема унификации материалов и передачи звучащей речи.**

На первом этапе создания диалектного корпуса возникает проблема отражения фонетических особенностей диалектной речи. Существуют различные подходы к её решению. Так, при создании корпуса говоров р. Устья проблема фактически была снята, т.к. каждый текст снабжён аудиозаписью, синхронизированной с ним, в связи с чем разработчиками корпуса принято решение давать тексты в стандартной орфографии [Waldenfels et. all]. При создании диалектного подкорпуса НКРЯ каждый текст даётся в нескольких вариантах – транскрипции, облегчённой транскрипции, орфографической записи [Kachinskaya & Sichinava 2017].

Томский диалектный корпус (далее ТДК) создаётся на основе имеющихся архивных записей. При этом появляется необходимость унифицировать разнородные элементы архива (самые старые записи – 40-60 гг. даны в полной транскрипции, записи 70-90 гг. в «полутранскрипции», в них часто произвольно отражаются те или иные черты; расшифровки новых аудиозаписей даются по специальному стандарту). Качественные аудиоматериалы имеются только для записей последнего десятилетия; восстановить полностью произношение для текстов, выполненных в «полутранскрипции», не представляется возможным. Кроме того, изучение фонетических особенностей говора не является приоритетным для разработчиков корпуса. В связи с обозначенными причинами при создании ТДК принято компромиссное решение: тексты подаются в упрощённой транскрипции.

Общая идея состоит в том, чтобы отмечать только а) нелитературные и б) вариативные произносительные особенности. Написания, не отражающие реальных различий в произношении, по возможности унифицируются. Поэтому, например, при наборе тетради для корпуса *жызнь*заменяется на*жизнь*, *щитает* на *считает* (т.к. произносительный вариант, при котором сочетание звуков *сч* в начале слова произносится «так как пишется», в исследуемых говорах не зафиксирован), но при этом сохраняется написание *шшытает* (долгие твёрдые шипящие – характерная особенность среднеобских говоров)*.*При проверке набора*апеть*заменяется на*опе'ть*(т.к. аканье в безударной позиции – общерусское, а переход а в э под ударением – лексикализованная диалектная черта). В то же время при наличии реальных произносительных вариантов написание отражает особенности каждого из них (так, не приводятся к единой форме *кода'* и *когда, тысяча* и *тыща,* *грит, гыт* и *говорит,* *така'* и *такая, чё* и *что, этот* и *е'тот* и др.). Ударение ставится по возможности во всех нелитературных формах (на основе рукописи или аудиозаписи), а также в формах литературного языка, допускающих вариативность (*тво'рог – творо'г, о'тдал – отда'л*) и в тех случаях, когда оно играет смыслоразличительную роль (*за'пах – запа'х, дру'га – друга'* диал. *‘другая’*). При поиске словоформы с ударением и без него распознаются как идентичные.

1. **Проблемы тематической разметки.**

Специфика тематической разметки в диалектном корпусе состоит в том, что она осуществляется на уровне отдельного фрагмента, а не текста в целом. Цель такой разметки – отразить своеобразие диалектной коммуникации.

В первоначальной концепции корпуса перечень тем ТДК включал 52 пункта [Yurina 2010: 61]. Впоследствии этот список был значительно скорректирован. Сформирован иерархический список макротем, включающий разделы РАБОТА, БЫТ, ЧЕЛОВЕК ДУХОВНЫЙ и др.; устранён ряд частных тем, таких как «Картофель», «Ягоды», «Грибы» – вместо них выделена более широкая тема «Растения»; добавлен ряд новых тем, например, «Животные» с дифференциацией подтем «Домашние животные», «Дикие животные», «Вредители»; часть тем переименована, чтобы включать новое содержание, например, «Любовные связи» -> «Межличностные отношения», что подразумевает также дружбу, конфликты. [Ivantsova 2017]. Темы «оценка современности», «оценка человека», «приметы» были исключены из первоначального перечня, т.к. это, скорее, жанры коммуникации, и соответствующие фрагменты будут включены в жанровую разметку.

Хронотоп (место и время описываемых событий) также даётся как часть тематического описания. Указание места событий, если оно не совпадает с местом записи, маркируется темой «Страны, города, сёла». Время событий довольно редко можно определить с точностью даже до десятилетия, поэтому тематическая разметка указывает не только на срок, но одновременно и на содержание текстового фрагмента: «Период единоличного хозяйства», «Революция и гражданская война», «Колхозы и совхозы», «Репрессии», «Великая Отечественная война», «Перестройка и распад СССР», «Политика в современной России (с 90-х гг.).

Одной из проблем тематической разметки является степень детализированности тематического перечня. В ТДК список тем разрабатывался таким образом, чтобы избежать пересечений с другими реализованными и планируемыми видами разметки и поиска. Так, в составе темы «Обряды» решено не выделять специальные микротемы «Крещение», «Похороны» и др., т.к. практически все тексты, посвящённые данным темам, достаточно легко найти по соответствующему ключевому слову – их названию. (Ср. близкий подход, принятый создателями диалектного подкорпуса в составе НКРЯ [Kachinskaya & Sichinava 2017: 73]).

Другой проблемой является контекстная обусловленность темы. Для определения темы разметчик опирается на списки ключевых слов, представленные в специально разработанной инструкции, но во многих случаях опоры на ключевые слова оказывается недостаточно, приходится осуществлять поиск смысловой доминанты текста, выявлять акценты, сделанные говорящим. Полностью контекстно обусловленной является выделенная в ТДК тема «Чужие», предполагающая описание тех социальных групп, к которым не принадлежит информант (чаще всего – по национальному, религиозному, территориальному признаку или их комбинации). Так, описание жизни ссыльных маркируется данной темой только если сам говорящий принадлежит к группе старожилов: *Я примерно, конечно, в войну не голодова'ла, у меня муж рыбачил, а так вот, приходили люди, сядут на стул и встать не могут, то что из сил выбились – ссыльные. [Кто это ссыльные?] Ссылали их, а тут вот война начала'сь и их уже обратно, кото'ры бежали, кото'ры их ссылали.* (Парабель, 1985, жен.).

Количественные данные, собранные на сегодняшний день, показывают, что чаще всего в текстах упоминаются темы «Семья и родственники» (представлена в 748 текстах, что соответствует 74%), «О себе» (635, или 63%), «Еда» (558, или 55%), «Язык и речь» (516, или 51%), «Выращивание животных» (514 текстов, или 51%), «Болезнь» (483, или 48%), «Одежда и обувь» (470, или 47%). Эти цифры отражают типичные «зоны актуального внимания» сельского жителя.

Наиболее редки темы «Мужские работы по дому» (38 текстов, или 4 %), «Экология» (43 текста, или 4%), «Перестройка и распад СССР» (23 текста, или 2%), «Политика в современной России» (12 текстов, или 1%), а также макротемы «Природа» (55 текстов, или 5%), «Животные» (24 текста, или 2%). «Непопулярность» той или иной темы может объясняться либо нехваткой текстов определённого типа (такова, на наш взгляд, причина редкости тем, касающихся событий недавней политической жизни), либо её обобщённостью (это касается макротем «Природа», «Животные», которые пополняются по остаточному принципу, т.е. при невозможности определить тему более детально; отметим, что здесь сказывается и специфика диалектной коммуникации, где «абстрактные» рассуждения вообще довольно редки). В других случаях невостребованность темы может служить аргументом для её исключения из тематического списка.

1. **Проблема репрезентативности корпуса.**

«Репрезентативность– свойство корпуса, заключающееся в статистически достоверном представлении языка или его части и достигаемое за счет необходимого объема и жанрового разнообразия текстов». [Kopotev 2014: 10].

ТДК обладает достаточным объёмом (более 1000 текстов, около 1 500 000 словоупотреблений). Однако нужно отметить, что записанные тексты, как и в других диалектных корпусах, по большей части являются «полуаутентичными». Ситуация сбора материала в экспедиции предполагает общение малознакомых людей, различных по социальному статусу. При этом диалектоносителям не всегда ясна цель коммуникации, и само присутствие собирателя может влиять на их речевое поведение [Tubalova 2016]. В то же время экспедиционные материалы всё же близки к беседе и являются более спонтанными видами речевой деятельности, чем чтение и пересказ (применявшиеся при создании речевых корпусов, в частности петербургского «Звукового корпуса русского языка» [Bogdanova-Beglaryan et.all 2017]). Жанровым разнообразием тексты ТДК также не отличаются (см. об этом ниже).

Для большинства традиционных исследовательских задач (таких как описание морфологических характеристик говора, семантизация и лексикографирование диалектных слов, описание системных связей лексических единиц, исследование отдельных семантических полей говора) тип текста не имеет существенного значения. Однако выдвижение в центр исследовательского интереса проблем когнитивно-дискурсивного моделирования диалектной коммуникации требует опоры на достаточно объёмные связные тексты. Фиксация такого рода материалов началась с к. 70-х гг. ХХ в. в связи с развитием технических средств записи устной речи. Спорадически тексты, близкие к спонтанной связной речи, встречаются и в более ранних записях (как правило, они анонимны и маркированы по месту записи: «в магазине», «на лавочке»). Ценность таких записей в том, что они отражают естественную, немоделируемую ситуацию общения и соответствующие языковые средства.

В ТДК проводится разметка по типу текста, нацеленная на маркирование, с одной стороны, текстов, близких к естественной коммуникации (диалог, полилог, ситуативное вкрапление), с другой – текстов, организованных как ответы на вопросники. В ходе работы к этому перечню был добавлен ещё один тип текста – реплика, представляющая собой часть диалога или полилога в тех случаях, когда коммуникантов несколько, но ответ собеседника не зафиксирован, либо собеседником выступает собиратель материала. Например: *В баню потошшышь воду? Уху будешь есь?*

На данный момент маркировано по типу около 60% всех введённых в корпус текстов. Первые количественные результаты свидетельствуют о следующем соотношении типов: ответы на вопросники – 303 фрагмента, реплика – 325, ситуативное вкрапление – 113, диалог – 62, полилог – 14. Таким образом, можно констатировать преобладание «естественных» текстов.

1. **Перспективы ресурса. Введение новых типов разметки.**

Особый характер диалектного корпуса и актуальные потребности диалектологической науки требуют маркировать тексты не только по экстралингвистическим параметрам, но и по прагматическим характеристикам. Для достижения данной цели необходимо разработать систему жанровой разметки диалектного корпуса, которая пока находится в стадии обсуждения. Отметим недостаточность имеющихся схем данного вида разметки в существующих русских диалектных корпусах.

В ТДК жанровую разметку планируется сделать уровневой, на первом уровне разделив фольклорную и нефольклорную речь (аналогично тому, как это реализовано в диалектном подкорпусе НКРЯ). За основу жанрового деления нефольклорной речи планируется взять перечень речевых жанров, предложенный Т.В. Шмелёвой. Отметим, что данная концепция была успешно апробирована на материале диалектных записей [Kazakova 2007]. Определённый интерес представляет также разметка речевых актов, представленная в региональном корпусе «Один речевой день» и частично описанная в работе Т.Ю. Шерстиновой [Sherstinova 2018], а также разметка речевых действий в корпусе МУРКО [Grishina 2009].

Проблема реализации жанровой разметки состоит в том, чтобы соблюсти баланс между информативностью, с одной стороны, и свободой пользователя-исследователя – с другой. Выделение макрожанров (таких как «воспоминание» или «автобиографический рассказ») малоинформативно, т.к. покрывает 80-90% всех текстов и не позволяет дифференцировать их по жанру (в пределе практически все материалы могут быть сведены к макрожанру «беседа с диалектологом»). Однако осуществление детализированной жанровой разметки фактически представляет собой самостоятельную исследовательскую задачу.

В планы исполнителей проекта входит также внедрение морфологической разметки. В качестве первого шага к достижению этой цели в корпусе осуществлена лемматизация слов и реализован поиск по лемме на основе анализатора phpMorphy, который планируется использовать и в дальнейшем. Отметим, что данный ресурс позволяет лемматизировать не только общерусские слова, но и единицы, имеющие отклонения от литературной нормы (так, автоматически связываются формы существительных – *зыбка, пашеница, лисопед*, глаголов – *висеться, базлать, бузовать*, прилагательных – *большенький, бусый* и др.). В то же время лемматизация некоторых других слов и словоформ, отличных от литературного стандарта, требует ручной корректировки (существительное *ти'тя* ‘грудь’, возведённое к форме гипотетического глагола *титить,* существительное *зубомо'й* ‘насмешник’, автоматический определяемое как форма существительного *зубома*, существительное *абармо'*, ‘мясо убитой или павшей скотины, употребляемое как приманка для зверей’, автоматически возводимое к форме прилагательного *абармый* и др.).

1. **Перспективы ресурса. Интеграция корпуса и словарей.**

Перспективы развития Томского диалектного корпуса связаны также с интеграцией данного ресурса и системы словарных источников, созданных на материале говоров Среднего Приобья. Предполагается, что такая интеграция позволит решить несколько задач. Во-первых, возможность специально маркировать в корпусе диалектные слова и давать ссылку на их толкование. Отметим, что подобная возможность пока не реализована ни в одном из русских диалектных корпусов, хотя она осуществлена в основной части НКРЯ. Для достижения данной цели планируется использовать дифференциальный толковый словарь говоров среднеобского региона в 7 томах [Palagina et al. 1964–1967, Zakharova et al. 1975, Palagina 1983–1986]. К настоящему моменту словарь переведён в электронный формат, ведётся работа по созданию базы данных на его основе.

Во-вторых, существующие словари могут положить начало осуществлению семантической разметки (в соответствии с принципами, предложенными в [Rakhilina et.all 2009]). Для этого могут быть полезны как вышеупомянутый дифференциальный словарь, так и словари полного типа [Blinova 1998–2002, Ivantsova 2006–2012]. Эти источники дают возможность составить перечни единиц, относящихся к определённым лексико-грамматическим категориям (диминутивы, аугментативы, феминитивы, семельфактивы и др.) и в дальнейшем осуществлять поиск по ним. Значительным потенциалом для осуществления семантической разметки обладает также новый идеографический словарь, включающий как диалектные, так и общерусские единицы [Berzin et.all 2018]. В частности, в словаре выделены классы «цвет», «звук», «запах», «вкус»; «животные»; «растения»; «семья и родственники»; «движение»; «положение в пространстве»; «оценка» и др.; представлены также некоторые типы отношений (мереология – «части тела животных»; «части тела человека»; множество – «совокупность предметов»; «совокупность людей»).

В заключение кратко охарактеризуем аспекты применения нового корпуса в лингвистических исследованиях. Пословный поиск на основе большого массива текстов создаёт новые возможности для исследования диалектной концептосферы [Demeshkina 2018], а также открывает перспективу сопоставительных исследований бытовой устной речи сельских жителей и горожан [Ivantsova 2018]; метаразметка корпуса позволяет изучать взаимосвязь экстралингвистических параметров диалектных текстов [Zemicheva 2018]. В будущем на основе реализации морфологической разметки и усовершенствования метаразметки появится возможность исследовать влияние социолингвистических факторов на грамматические особенности речи. Наконец, корпус служит источником для новых словарей, особенно – лингвокультурологической направленности, что обеспечивается благодаря полнотекстовому формату ресурса. Так, его материалы активно задействовались при составлении «Словаря детства» (на материале говоров Среднего Приобья) и «Словаря сибирского свадебного обряда».

**References**

Berzin A.K., Boltova V.S., Zemicheva S.S. (2018) Thesaurus of the dialect language personality, available at: <http://losl.tsu.ru/?q=node/65>

Blinova, O.I. (ed.) (1998–2002) Vershininsky dictionary [Vershininskiy slovar'], Tomsk State University Publ., Tomsk, vol. 1-7.

Bogdanova-Beglaryan N.V., Sherstinova T.Yu., Zaides K.D. (2017) Corpus “Balanced Annotated Text Library”: a technique for multilevel analysis of Russian monologue speech [Korpus Sbalansirovannaya Annotirovannaya Tekstoteka metodika mnogourovnevogo analiza russkoj monologicheskoj rechi], Analysis of colloquial Russian speech (AP3–2017): Proceedings of Seventh Interdisciplinary Workshop [Analiz razgovornoj russkoj rechi AR3 2017 trudy Cedmogo mezhdisciplinarnogo seminara], Polytechnic-print, St. Petersburg, pp. 8–13.

Demeshkina T.A. (2018) “Exile” as a phenomenon of the Siberian linguaculture [Ssylka kak fenomen sibirskoj lingvokultury], Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya], no 56, pp. 34–46.

Grishina E.A. (2009) Multimodal Russian Corpus (MURСO): Problems of Annotation [Multimedijnyj russkij korpus MURKO: problemy annotacii], The National Corpus of the Russian Language: 2006-2008. New results and perspectives [Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy], St. Petersburg, pp. 150–174.

Ivantsova E.V. (2018) The variety of the key concept “Bread” in different types of Russian speech culture [Variativnost realizacii klyuchevogo koncepta hleb v raznyh tipah russkoj rechevoj kultury], Current Trends and Future Perspectives in Russian Studies. Proceedings of the International Conference on Russian Studies at the University of Barcelona [Aktualnye problemy i perspektivy rusistiki. Materialy po itogam Mezhdunarodnoj konferencii rusistov v Barselonskom universitete], Barcelona, pp. 1164–1173.

Ivantsova E.V. (ed.) (2006–2012) Complete dictionary of the dialect language personality [Polnyy slovar' dialektnoy yazykovoy lichnosti], Tomsk State University Publ., Tomsk, vol. 1-4.

Kachinskaya I.B. & Sichinava D.V. (2017) On the corpus of dialectal texts in the Russian national corpus [O korpuse dialektnyh tekstov v nacionalnom korpuse Russkogo yazyka], Russian Journal of Lexicography [Voprosy leksikografii], no 11, pp. 71–85.

Kazakova O.A. (2007) Dialectal language personality in the genre aspect [Dialektnaya yazykovaya lichnost v zhanrovom aspekte], Tomsk State University Publ., Tomsk.

Kopotev M. (2014) Introduction to corpus linguistics. Manual for students of philological and linguistic specialties of universities [Vvedenie v korpusnuyu lingvistiku. Uchebnoe posobie dlya studentov filologicheskih i lingvisticheskih specialnostej universitetov], Praha, Animedia Company.

Palagina V.V. (ed.) (1983-1986) The Middle Ob Dictionary (supplement) [Sredneobskiy slovar' (dopolnenie)], Tomsk State University Publ., Tomsk, vol.1-2.

Palagina V.V., Blinova O.I., Yantsenetskaya M.N., Sokolov O.M., Ivanova F.P. (1964-1967) The Dictionary of the Russian Old-Timer Dialects of the Central Basin of the Ob River [Slovar' russkikh starozhil'cheskikh govorov Sredney chasti basseyna r. Obi], Tomsk State University Publ., Tomsk, vol.1-3.

Rakhilina E.V., Kustova G.I., Lyashevskaya O.N., Reznikova T.I., Shemanaeva O.Yu. (2009) Tasks and principles of semantic markup of vocabulary in the NCRL [Zadachi i principy semanticheskoj razmetki leksiki v NKRYA], The National Corpus of the Russian Language: 2006-2008. New results and perspectives [Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy], St. Petersburg.

Sherstinova T.Yu. (2018) The Structure of Everyday Dialogue as the Sequence of Speech Acts [Struktura povsednevnogo dialoga kak posledovatelnost rechevyh aktov], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” 2018 [Kompyuternaya lingvistika i intellektualnye tekhnologii. Po materialam ezhegodnoj mezhdunarodnoj konferencii Dialog 2018], Issue 17, Moscow, pp.638–651.

Stolyarova E.A. (2000) Types of lexico-semantic fields in Russian colloquial speech [Tipy leksiko-semanticheskih polej v russkoj razgovornoj rechi], Russian language today [Russkij yazyk segodnya], vol. 1, Moscow, pp. 433–443.

Tubalova I.V. (2016) Polyphonic text in oral personal-oriented discourses [Polifonicheskiy tekst v ustnykh lichnostno-orientirovannykh diskursakh], Tomsk State University Publ., Tomsk.

Yurina E.A. (2011) Tomsk dialect corpus: the starting point [Tomskij dialektnyj korpus: v nachale puti], Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya], no 2 (14), pp. 58–63.

Waldenfels R., Daniel M., Dobrushina N. (2014) Building the Ustya River Basin Corpus, an online corpus of a Russian dialect, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, Issue 13, Moscow, pp. 270–278.

Zakharova L.A., Sypchenko S.V. et al. (1975) The Dictionary of the Russian Old-Timer Dialects of the Central Basin of the Ob River (supplement) [Slovar' russkikh starozhil'cheskikh govorov Sredney chasti basseyna r. Obi (dopolnenie)], Tomsk State University Publ., Tomsk, Pt.1, 2.

Zemicheva S.S. (2018) The correlation between the topic of a dialect text and the speaker’s gender (based on the materials of Tomsk dialect corpus) [Vzaimosvyaz tematiki dialektnogo teksta i pola govoryashchego na materiale Tomskogo dialektnogo korpusa], Current Trends and Future Perspectives in Russian Studies. Proceedings of the International Conference on Russian Studies at the University of Barcelona [Aktualnye problemy i perspektivy rusistiki. Materialy po itogam Mezhdunarodnoj konferencii rusistov v Barselonskom universitete], Barcelona, pp. 483–492.

1. Исследование выполнено при поддержке гранта РФФИ № 19-012-00320 «Томский диалектный корпус как новый ресурс для изучения народно-речевой культуры», и Программы повышения международной конкурентоспособности Томского государственного университета на 2013-2020 гг. [↑](#footnote-ref-1)
2. Tomsk dialect corpus [Electronic resource] URL: http://losl.tsu.ru/corpus/demo(access date: 29.03.2019). [↑](#footnote-ref-2)
3. Bulgarian Dialectology as Living Tradition [Electronic resource]. URL: http://bulgariandialectology.org/ (access date: 05.03.2019). [↑](#footnote-ref-3)
4. Korpus DIALEKT [Electronic resource]. URL: <http://wiki.korpus.cz/doku.php/cnk:dialekt> (access date: 05.03.2019). [↑](#footnote-ref-4)
5. Herzlich willkommen in der Datenbank für Gesprochenes Deutsch (DGD) [Electronic resource]. URL: <https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome> (access date: 05.03.2018). [↑](#footnote-ref-5)
6. BayDat [Electronic resource]. URL: <http://www.baydat.uni-wuerzburg.de:8080/cocoon/baydat/> (access date: 29.03.2019). [↑](#footnote-ref-6)
7. Estonian Dialect Corpus [Electronic resource]. URL: <http://www.murre.ut.ee/estonian-dialect-corpus/> (access date: 29.03.2019). [↑](#footnote-ref-7)
8. The Nordic Dialect Corpus [Electronic resource]. URL: <http://www.tekstlab.uio.no/nota/scandiasyn/> (access date: 28. 03.2019) [↑](#footnote-ref-8)
9. CORDIAL-SIN – Syntax-oriented Corpus of Portuguese Dialects [Electronic resource]. URL: http://www.clul.ulisboa.pt/en/10-research/695-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects (access date: 28. 03.2019). [↑](#footnote-ref-9)
10. Russian dialect corpus [Electronic resource]. URL: <http://www.ruscorpora.ru/search-dialect.html>

    (access date: 28. 03.2019). [↑](#footnote-ref-10)
11. Ustja River Basin Corpus. Statistics [Electronic resource]. URL: <http://parasolcorpus.org/Pushkino/stats.php> (access date: 28. 03.2019). [↑](#footnote-ref-11)