УДК 81'27
DOI 10.17223/19986645/33/4

З.И. Резанова

ЛИНГВИСТИЧЕСКИЙ КОРПУС «ТОМСКИЙ РЕГИОНАЛЬНЫЙ ТЕКСТ»: ТИПОЛОГИЧЕСКИ РЕЛЕВАНТНЫЕ ПАРАМЕТРЫ СБАЛАНСИРОВАННОСТИ И РЕПРЕЗЕНТАТИВНОСТИ[1]

В статье представлены типологически значимые характеристики создаваемого в настоящее время в Томском государственном университете корпуса региональных текстов, определяемые характером интерпретации сбалансированности текстового состава как отражения в нем структуры коммуникации в регионе в единстве и взаимодействии устной и письменной, литературной и нелитературных форм русского языка, жанров институционального и личностного общения, а также жанровых форм, репрезентирующих пересечение дискурсов и интерференцию русского языка с языками, функционирующими в регионе.
Ключевые слова: *корпусная лингвистика, сбалансированность корпуса, репрезентативность корпуса, региональный корпус, региональная лингвистика.*

В Лаборатории когнитивных исследований языка филологического факультета Томского государственного университета в настоящее время ведется работа по созданию корпуса текстов русского литературного языка, маркированных локальной (региональной) ограниченностью их существования: начата разработка его концепции, ведется первичный сбор материала, его лингвистическая разметка на базе учебных практик по направлению подготовки бакалавриата «Фундаментальная и прикладная лингвистика», ресурсов лаборатории, разрабатывается компьютерная программа платформы корпуса под руководством проф. В.В. Поддубного.

В предыдущих публикациях авторов проекта уже обсуждались отдельные аспекты концепции создаваемого корпуса – параметр региональности как базовый признак ограничения представленного материала [1], содержательное направление признаков сбалансированности и репрезентативности относительно данного типа корпуса [2]. В данной статье мы характеризуем создаваемый корпус по ряду типологически важных параметров, по которым в настоящее время оцениваются существующие корпуса текстов, и, в случае необходимости, сравниваем с соответствующими параметрами уже реализованных проектов корпусов.

Как известно, корпусная лингвистика – прикладное направление мирового языкознания, которое имеет уже достаточно длительную историю, начавшуюся с создания Brown Corpus в Брауновском университете (США) в 1963 г., и в настоящее время оно представлено значительным количеством реализованных проектов. Обобщение полувекового опыта создания лингвистически ориентированных корпусов текстов приводит к выявлению их разноаспект-

[1] Публикация подготовлена в рамках поддержанного РГНФ научного проекта №14-14-70010 «Лингвистический корпус «Томский региональный текст»: концепция и структура».

ной типологизации, противопоставлению по совокупности релевантных признаков. Так, например, В.П. Захаров выделяет в качестве таковых тип данных (письменные vs. устные vs. смешанные), язык текстов (русский vs. английский и т.д.), «параллельность» (одноязычные vs. параллельные), «литературность»/ специфичность (тексты одной из форм национального языка (литературные, диалектные и под.) vs. смешанные), жанр (ограничения жанрового типа); «общность» (общие vs. одного писателя), хронологический аспект (синхронические vs. диахронические); объем текстов (полнотекстовые vs. «фрагментнотекстовые), разметка (размеченные vs. неразмеченные), характер разметки (морфологические vs. синтаксические vs. семантические), динамичность (динамические vs. статические); доступность (свободно доступные vs. коммерческие vs. закрытые); назначение (исследовательские vs. иллюстративные) [3. С. 13].

Некоторые из перечисляемых В.П. Захаровым признаков являются автономными (например, язык текстов и принцип отбора материала – синхроничность vs. диахроничность), другие взаимосвязаны, находятся либо в отношениях взаимоподчинения, либо родовидовых. Например, жанровая представленность текстов корпуса является конкретизирующим параметром по отношению к параметру «литературность» vs. специфичность», и этот аспект квалификации применим как к корпусам литературных текстов, так и диалектным, просторечным. В свою очередь, параметр «общность» (тексты «общие» vs. одного писателя) является конкретизирующим к признаку «литературность».

Представляется, что ряд параметров может быть обобщен относительно противопоставления корпусов универсальных (ориентированных на представление максимально разнотипных текстов) и специфичных (ориентированных на ограничение в представлении текстов по каким-либо их значимым признакам).

В наблюдаемом поле типологически разнообразных корпусов в настоящее время активно проявляются две тенденденции в развитии корпусной лингвистики – создание глобальных национальных проектов ([4; 5; 6; 7; 8; 9; 10] и др.) и разного рода специализированных корпусов, например [11; 12; 13] и др. Эти тенденции не противоречат друг другу, но являются взаимодополняющими, о чем, в частности, свидетельствует развитие проекта Национального корпуса русского языка, как пополняющегося в настоящее время в основном фонде, так и развивающегося за счет разработки специализированных подкорпусов: мультимедийного, диалектного, поэтическиого, акцентологического, устной речи.

Полагаем также, что корпуса противопоставляются и по тому, как их создателями интерпретируются признаки универсальности и сбалансированности, в каком направлении и на каком основании осуществляется спецификация материала корпуса. Именно эти признаки при характеристике корпуса представляются нам чрезвычайно важными и, как было отмечено в [2], коррелируют с важнейшим характеристиками, по которым оценивается корпус, а также с принципами разметки и метаразметки корпуса. Их значимость при оценке источниковедческих возможностей корпусов неоднократно отмечалась в обзорах ([14. С. 10–38; 15. С.402–461; 16. С. 31–61; 17. С. 9–17; 18] и

др.). Сбалансированные и с достаточной репрезентативностью корпуса текстов (в настоящее время к таковым относятся те, текстовый объем которых превышает 100 млн словоупотреблений) позволяют разрешать многие исследовательские языковедческие задачи на новом уровне. Так, использование данных РНКЯ, что неоднократно отмечалось, позволяет внести существенные коррективы относительно утвердившихся мнений о так называемой «языковой реальности», прежде всего о сочетаемости или об управлении конкретных лексем, о лексических и грамматических значениях, особенностях синтаксических конструкций и под. [19. С. 318–331; 20. С. 227–245].

Как следует из названия представляемого проекта – «Лингвистический корпус. Томский региональный текст», его базовым дифференциальным признаком является региональность как принцип ограничения материала. Язык региона включает сложно организованное единство разных форм национального языка: нормированного литературного языка, диалектов, социолектов, городского и сельского просторечия. Авторы проекта ставят своей целью репрезентативное и сбалансированное представление в корпусе регионального варианта русского литературного языка. Отметим, что проблема регионального варьирования литературного языка в настоящее время активно дискутируется, вследствие чего этот аспект концепции корпуса нуждается в отдельном обсуждении, что и было предпринято в статье соавтора проекта Н.А. Мишанкиной [1]. Мы же, присоединяясь к высказанным позициям автора по данной проблеме, считаем необходимым подчеркнуть, что создание корпуса направлено на формирование эмпирической базы такого рода исследований, которая при обсуждении дискуссионной проблемы позволит исследователям опереться на данные лингвистически размеченного сбалансированного и репрезентативного корпуса текстов.

С одной стороны, в сравнении с корпусами национальных языков разрабатываемый корпус является специализированным, так как в задачи авторов корпуса не входит репрезентативное текстовое представление всех особенностей и сфер общения на русском языке; цель его создателей – собрание и лингвистически релевантная разметка текстов *регионально ограниченного варианта* русского литературного языка с выдвижением на первый план задач представления аспекта его локального варьирования. Такая направленность создаваемого корпуса непосредственно коррелирует с «идеологическими предпочтениями авторов НКРЯ» – «внимание к **синхронной вариативности** языка» и принципиальная «нелитературоцентричность» [21. С. 7–20]. Последний принцип заключается в том, что при включении текстов художественной литературы в корпус их место должно отражать роль данного типа текстов в структуре региональных коммуникативных практик в сочетании с текстами специального делового, научного, обыденного и т.д. общения.

Принципы отбора материала в Лингвистический корпус «Томский региональный текст» – **репрезентативность и сбалансированность** – находятся в определенной корреляции с принципами НКРЯ, который относится к репрезентативным и сбалансированным корпусам, однако следует отметить ряд их особенностей. Во-первых, как отмечалось, эти принципы применяются к материалу, ограниченному локально, – текстам, созданным в Томском регионе. Во-вторых, если НКРЯ относится к сбалансированным и репрезентативным

относительно письменных форм русского языка («содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т.п.), и… все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода» [7]), то авторы проекта Лингвистический корпус «Томский региональный текст» ставят задачу создать корпус, включающий совокупность текстов – *устных и письменных,* представляющих структуру общения в регионе.

В российской лингвистике региональные корпуса представляют, как правило, диалектную форму национального языка (назовем здесь наиболее известные и отрефлексированные: Саратовский диалектологический корпус и диалектный подкорпус НКРЯ [22. С. 215–232; 23. С. 114–128; 24, С. 359–367; 25. С. 620–629], а также находящийся на этапе разработки Томский диалектный корпус [26]).

Как известно, эти проекты авторами разрабатывались на разных теоретических основаниях и ориентированы они на выполнение разных задач (см. об этом: [24. С. 359–367]). Сравнивая существующие региональные корпуса, отметим существенное различие в характере представления диалектных текстов: в Диалектном подкорпусе НКРЯ диалектные материалы представляются на фоне литературной нормы как совокупность отклонений от нее; реализованный проект Саратовского диалектологического корпуса и планируемый к осуществлению Томский диалектный корпус строятся с ориентацией на системное представление диалекта в его дискурсивно-жанровом разнообразии [24; 26]. Представляемый в данной статье корпус текстов выстраивается также на принципах недифференциального, но системного представления системы общения в регионе.

Мы полагаем, что создание регионального подкорпуса национального языка – это весьма актуальная задача, связанная как с дальнейшим развитием корпусной лингвистики, так и с активизацией проблем регионального варьирования литературного языка в русистике. В российском языкознании проблемы регионального языкового варьирования осмыслялись прежде всего в рамках диалектологии. Региональная вариативность литературного языка в русистике только начинает исследоваться, и создание такого корпуса, представляющего обширные лингвистически размеченные текстовые массивы, как мы уже отметили ранее, явится надежным основанием для статистически достоверных исследований о характере и направлениях территориального варьирования русского литературного языка, характере его контактирования и взаимовлияния с другими формами национального языка, другими языками в многообразии их форм, функционирующих в регионе.

Определение параметров сбалансированности текстов – одна из сложных задач, решаемая применением ряда методик, в том числе с опорой на результаты социолингвистического анкетирования (ср. оценку деятельности авторов Чешского национального корпуса в этом отношении: [15. С. 405]). Имеющиеся данные о соотношении типов текстов в корпусах, определяемых как сбалансированные, свидетельствуют о том, что «литературоцентризм», как правило, преодолевается за счет увеличения доли публицистики в его составе; художественные тексты при этом занимают второе место в общем

объеме корпуса, третье – так называемые специализированные тексты их внутренней дополнительной специализацией. Так, Словацкий национальный корпус, относимый к сбалансированным корпусам, объемом 339 млн словоупотреблений включает тексты в следующей пропорции: публицистика (60,6 %), художественная литература (17,5 %), специализированные тексты (11,6 %), другое (10,3 %). Два варианта Словенского национального корпуса FIDA и FidaPLUS содержат тексты «в следующих соотношениях: художественные тексты (6 vs. 3,47 %), научные (18,5 vs. 10 %), другие (75,5 vs. 86,34 %); книги (22,7 vs. 8,74 %), газеты (46,6 vs. 65,26 %), журналы (23,9 vs. 23,26 %), тексты из Интернета (электронные тексты) (0,02 vs. 1,24 %), другое (в том числе незначительная доля устной речи – стенограмм парламентских слушаний) (6,78 vs. 1,5 %). В то время как, например, корпус Боснийских текстов, не относимый к сбалансированным, включает тексты в таком соотношении: художественная литература (43 %), эссеистика (29,6 %), публицистика (16,9 %), книги для детей (6 %), религиозные тексты (2,8 %), юридические тексты (1,5 %), фольклор (0,2 %) (данные приводятся по: [15. С 413]).

Составители НКРЯ дифференцируют тексты, «представляющие современный русский *литературный* **(письменный)** язык» (шрифтовое выделение в тексте наше. – *З.Р.*), следующим образом: современная художественная проза разных жанров и направлений, современная драматургия, мемуарно-биографическая литература, журнальная публицистика и литературная критика, газетная публицистика и новости, научные, научно-популярные и учебные тексты, религиозные и религиозно-философские тексты, производственно-технические тексты, официально-деловые и юридические тексты, бытовые тексты (в том числе тексты, не предназначенные для публикации: личная переписка, дневники и т.п.). При этом авторы НКРЯ подчеркивают, что «тексты представлены в определенной пропорции, отражающей их долю в общем массиве современных текстов. Так, доля художественных текстов (включая драматургию и мемуары) составляет не более 40% и «все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода» [7].

Итак, общей чертой современных национальных лингвистических корпусов является непредставленность или слабая представленность в них разговорной речи. Как видим, данные корпуса репрезентативно представляют письменные тексты, включая транскрипты устной речи, относящиеся только к институциональному общению, к публичным жанрам устной официальной коммуникации (FIDA и FidaPLUS), или фольклору – корпус Боснийских текстов. Устная коммуникация может включаться в состав национального корпуса в статусе самостоятельного подкорпуса, как в НКРЯ (подкорпус устной речи и диалектный). Очевидно, во-первых, что их отношение к основному корпусу различно. Так, НКРЯ позиционирует себя как корпус, представляющий «современный русский литературный (письменный) язык», соответственно, устный подкорпус представляет другую подсистему литературного языка, диалектный – репрезентирует другую форму национального языка. Во-вторых, существенно, что устная коммуникация представлена в виде самостоятельных подкорпусов с частично вариативной разметкой, что не позволяет делать непосредственные выборки по одному заданному параметру.

Таким образом, современные сбалансированные корпуса, как правило, отражают структуру письменной коммуникации, часто – письменного институционального общения. Значимость активного вовлечения устной коммуникации, обыденной письменной коммуникации в структуру основного корпуса трудно переоценить. Преодолевая литературоцентризм, корпусы письменных текстов так или иначе остаются в рамках ограничений другого рода, фиксируя прежде всего тексты разных жанров институциональной, письменной коммуникации, преуменьшая долю обыденного личностного общения, текстов, порождаемых в непосредственных неформальных дискурсах. Если мы исключаем устное обыденное общение из структуры корпуса, он не может считаться представительным и сбалансированным при решении задачи представления текстов, отражающих **структуру общения** на данном языке. Проведенные в конце прошлого века исследования русской разговорной речи интерпретировали РР (разговорную речь) как самостоятельную подсистему, оппозиционирующую КЛЯ (кодифицированный литературный язык) на всех уровнях языковой системы ([27; 28] и др.). Напомним, что исследования велись на ограниченном количестве текстов, и создание сбалансированного корпуса устных текстов всех жанров и типов коммуникации стало бы основой нового уровня интерпретации соотношения системы устной и письменной коммуникации, институционального и личностного общения. Если мы говорим об отражении в структуре корпуса структуры *коммуникации,* то доля устного обыденного общения должна быть представлена в соответствии с его ролью в коммуникативном существовании современного человека. В данном случае большое значение имеет проект, выполняемый исследовательской группой Санкт-Петербургского университета [29].

При решении задачи пропорционального включения текстов всех форм коммуникации в состав корпуса его составители должны дополнительно решить проблему сбора материала обыденного устного неинституционального общения, которой не существует при обращении к письменным текстам, так как в корпус обычно включаются тексты, уже существующие на бумажных и электронных носителях. Обширных собраний транскрибированных текстов устного общения литературных языков в объемах, хоть сколь-нибудь соотносимых с письменными, не существует. Вследствие этого перед авторами корпуса, ориентированного на сбалансированность представления устного и письменного общения, в качестве самостоятельной ставится задача создания базы устных текстов.

При этом авторы должны решить и проблему второго уровня сбалансированности – устные тексты должны быть не просто представлены в значительном объеме, но представлять разные жанры обыденного общения, преодолевая преобладание жанра интервью в составе лингвистически релевантных баз данных и корпусов. С этой проблемой знакомы прежде всего диалектологи, в материалах полевых исследований которых на первых этапах развития диалектологии преобладали тексты бесед диалектологов с носителями данной формы национального языка. Так собирались материалы для диалектных словарей, и если для лексикоцентрического, лексикографического этапа развития диалектологии такой тип собирания материала был оптимальным, то для решения корпусных задач, которые, в свою очередь, ориентированы на

репрезентативность в текстоцентрических и дискурсивных исследованиях, одного такого способа сбора материала явно недостаточно.

В проекте предполагается достижение параметра сбалансированности не только за счет включения транскриптов устных текстов, но и за счет принципиального увеличения доли разных жанров обыденной коммуникации, в том числе тех, что находятся в зоне пересечения институционального и личностного общения (письма-обращения в официальные инстанции, протоколы собраний и т.д.). Сбалансированность должна достигаться также и более широким включением в структуру корпуса текстов новых современных типов коммуникации, опосредствованных различными техническими средствами: смс, самые разные жанры интернет-коммуникации. Весьма значительная степень различия моделей языка, получаемых на основе анализа корпуса текстов письменного институционального общения и корпуса, представляющего тексты Интернет-коммуникации, была весьма убедительно представлена в работе [30].

Отметим еще одно отличие в интерпретации сбалансированности авторов проекта Лингвистический корпус «Томский региональный текст» от принятого в НКРЯ, который сбалансирован относительно письменных текстов **литературного** языка. Авторы представляемого проекта интерпретируют сбалансированность в пределах литературного языка, но ставят целью отразить структуру общения в регионе как неоднородное в отношении нормированности и кодифицированности образование. Язык региона включает нормированный литературный язык, сельское и городское просторечие, диалекты, социолекты, жаргоны. В корпусе предполагается сбалансированное представление разных форм общения в аспекте их контактирования с литературной формой общения.

Соотношение дискурсов и жанров в представляемом корпусе также не может быть таким же, как в НКРЯ, как вследствие другого целеполагания, так и вследствие своеобразия описываемой языковой онтологии. Как отмечалось, сбалансированность корпусов национальных языков обычно достигается, как правило, уменьшением доли художественных текстов за счет увеличения представительства других форм институционального общения; стремление включить в корпус разные жанры обыденной коммуникации приведет к значительному сокращению доли художественных текстов, созданных томскими авторами, в структуре Томского регионального корпуса.

Отметим значимость еще одного аспекта сбалансированности текстов в составе корпуса. Представление в корпусе функционирования литературного языка в регионе как динамического взаимодействия с разными формами национального языка должно сочетаться с отражением явлений контаминации с контактирующими языками. Регион мононационален, при этом абсолютное доминирование русских (по данным переписи 2002 г., 90,84%) не исключает этноязыкового разнообразия (10% населения составляют представители более двух десятков национальностей: татары – 1,93%, украинцы – 1,60%, немцы – 1,29%, остальные относы составляют от 0,56 до 0,06%, это (в порядке убывания численности), чуваши, белорусы, азербайджанцы, армяне, башкиры, мордва, селькупы, узбеки, удмурты, молдаване, поляки, казахи, корейцы, ханты, марийцы, евреи, эстонцы, латыши, чеченцы, грузины [31]. Как видим,

этнически, социально, в языковом отношении это очень неоднородный состав, включающий и сибирские коренные этносы (например, селькупы, ханты), переселенцев XVIII–XX вв. (украинцы, белорусы, татары, немцы, поляки, литовцы, чуваши, мордва и др.), в том числе и представителей диаспор, возникших в результате современных миграционных процессов, трудовой и учебной миграции. В результате этнических взаимодействий сформировались вариативные би- и полилингвальные языковые ситуации. При этом, в то время как язык коренных народов активно изучается (см., например, данные на сайте: [32]), равным образом как диалектные формы русского языка в регионе [33], варианты русского литературного языка, функционирующего в условиях двуязычия и диглоссии, практически не изучены. Особенности языковой ситуации в регионе свидетельствуют о возможности наличия разных вариантов билингвизма, как следствие – разных вариантов проявлений интерференций. Таким образом, концепция сбалансированности авторов Томского регионального корпуса предполагает текстовое представление коммуникативных практик русскоговорящих билингвов с разными вариантами двуязычия. Этот параметр сбалансированности обеспечивается представлением жанров разных форм национального языка, в пределах последних – репрезентативным представлением текстов монолингвов и билингвов. Естественно, при решении такой задачи актуализируется проблема определения состава коммуникативных практик, которые могут дать достоверные данные о языковой специфике такого типа текстов: устное бытовое общение, записанное для корпуса, бытовые письма в архивах с достоверной фиксацией этнической самооценки, нередактируемые тексты массовой коммуникации с более или менее достоверными показателями этнической принадлежности авторов – сайты, страницы социальных сетей, прежде всего страницы землячеств, сайты национально-культурных автономий (исследования данного типа текстов представлены в работах [34; 35; 36]).

Соответствие корпуса обозначенным в статье принципам репрезентативности связано не только с решением проблем сбалансированности как соотношения конкретных текстов, представляющих формы национального языка, дискурсы, жанры, а также варианты языкового, дискурсивного, жанрового смешения, не только с дополнительной проблемой сбора и предварительной обработкой материала, но и с определением специфических параметров разметки и метаразметки текстов корпуса, а также с решением вопроса о возможности получения пользователями корпуса полнотекстовых материалов (значимость возможности пользователя корпуса обратиться к полнотекстовым материалам в диалектном корпусе отмечается в работах О.Ю. Крючковой и В.Е. Гольдина [24]).

При определении репрезентативности и сбалансированности текстов корпуса важным является также параметр временного ограничения состава текстов корпуса. Как правило, если авторы корпуса не определяют этот аспект отбора материала, предполагается, что это корпус современных текстов, и далее границы понятия «современный» определяются в соответствии с принятыми в данной научной традиции границами, что мотивируется в значительной степени интенсивностью изменчивости языка на разных этапах его существования. Авторы представляемого корпуса вследствие стремления к

отражению структуры общения, включающего максимально полно формы обыденной коммуникации, предполагают значительно сузить временную принадлежность текстов, ядро которых составят тексты XXI в., XX в. будет представлен текстами письменных форм коммуникации, в том числе институциональной и обыденной, и смешанной – институционально-личностной.

Таким образом, представленная в статье интерпретация сбалансированности Томского регионального корпуса коррелирует с рядом следующих типологически важных параметров: *тип данных* – смешанные, т.е. включает письменные и транскрибированные устные тексты; *язык текстов* – русский, *параллельность* – непараллельные (одноязычные), *«литературность»* – представляет тексты как собственно литературные, так и просторечные, разговорные, а также специальные (тексты научной, производственно-технической, деловой коммуникации); *жанровая структура* – стремление к репрезентативности жанрового разнообразия разных функционально-стилевых регистров речи; *хронологический аспект* – синхронический, *динамичность* – элементы синхронной динамики, *разметка* – размеченный, *характер разметки* – морфологическая, синтаксическая, семантическая, *объем текстов* – полнотекстовый; *назначение* – исследовательский; *доступность* – доступный. Параметры разметки и метаразметки корпуса, определяемые целевой направленностью корпуса, интерпретацией характера сбалансированности и репрезентативностью будут представлены в следующих публикациях авторов проекта.

Литература

1. *Мишанкина Н.А.* Лингвистический корпус «Томский региональный текст»: теоретико-методологическое обоснование проекта // Вестн. Том. гос. ун-та. 2014. № 389. С. 28–37.

2. *Sologub O., Rezanova Z.I., Temnikova I.G.* The Concept of the Tomsk Regional Corpus: Balance and Representativeness // The XXV Annual International Academic Conference, Language and Culture, 20–22 October 2014. Procedia-Social and Behavioral Sciences 154 (2014), p. 175–178.

3. *Захаров В.П.* Корпусная лингвистика: учеб.-метод. пособие. СПб., 2005. 48 с.

4. *Британский* национальный корпус British National Corpus. URL: http://www.natcorp.ox.ac.uk

5. *Международный* корпус английского языка=International Corpus of English // URL: http:// ice-corpora.net

6. *Банк* английского языка Bank of English. URL: http://www.collins.co. uk/ Corpus/ CorpusSearch.aspx)

7. *Национальный* корпус русского языка. URL: http://ruscorpora.ru

8. *Национальный* корпус польского языка. URL: http://nkjp.pl

9. *Словацкий* национальный корпус. URL: http://korpus.juls.savba.sk

10. *Чешский* национальный корпус. URL: http://ucnk.ff.cuni.cz

11. *Polish* and English Language Corpora for Research and Applications. URL:http:// korpus.ia.uni.lodz.pl

12. *St. Petersburg* Corpora of hagiographic Texts XV–XVII centuries // URL: http:// project. phil. pu. ru/skat)

13. *Chemnitz* German-English Translation Coropus. URL: http://www.tu-chemnitz.de/phil/ Internet-Grammar

14. *Резникова Т.И.* Корпуса славянских языков в интернете: обзор ресурсов // Die Welt der Slaven, 2008. LIII. С. 10–38.

15. *Резникова Т.И.* Славянская корпусная лингвистика: современное состояние ресурсов // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009, 402–461.

16. *Резникова Т.И., Копотев М.В.* Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005. М., 2005. С. 31–61.

17. *Шаров С.А*. Представительный корпус русского языка в контексте мирового опыта // НТИ, Сер. 2. 2003. № 6. С. 9–17.

18. *Беляева Л.Н.* Лексикографический потенциал параллельного корпуса текстов // corpora.iling.spb.ru/Docs/Belyaeva_corpora.pu

19. *Перцов Н.В*. О роли корпусов в лингвистических исследованиях // Тр. Междунар. конф. «Корпусная лингвистика-2006». СПб., 2006. С. 318–331.

20. *Перцов Н.В*. К суждениям о фактах русского языка в свете корпусных данных // Рус. яз. в науч. освещении. 2006. № 1(11). С. 227–245.

21. *Плунгян В.А*. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 16 (2). С. 7–20.

22. *Летучий А.Б*. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005, 215–232. URL: .http://ruscorpora.ru/ sbornik2005/ 13letuchy.pdf

23. *Летучий А.Б.* Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 114–128.

24. *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. Вып. 10 (17). М., 2011. С. 359–367.

25. *Сичинава Д.В., Качинская И.Б.* Корпус диалектных текстов в национальном Корпусе русского языка: сегодняшнее состояние и перспективы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог», Бекасово, 4–8 июня 2014 г. Вып. 13 (20). М., 2014. С. 620–629.

26. *Юрина Е.А*. Томский диалектный корпус: в начале пути // Вестн. Том. гос. ун-та. Филология. 2011. № 2 (14). С. 58–64.

27. *Земская Е.А*. Русская разговорная речь: лингвистический анализ и проблемы обучения . М.: Рус. яз., 1979. 240 с.

28. *Земская Е А., Китайгородская М.В., Ширяев Е.Н.* Русская разговорная речь: Общие вопросы. Словообразование. Синтаксис. М., 1981.

29. *Богданова Н.В., Бродт И.С., Куканова В.В., Павлова О.В., Сапунова Е.М., Филиппова Н.С*. О «корпусе» живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам ежегодной Междунар. конф. «Диалог» (2008). 2008. С. 57–61.

30. *Беликов В.И., Копылов Н.Ю., Пиперски А.Ч., Селегей В.П., Шаров С.А*. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьтерная лингвистика и интеллектуальные технологии. М., 2013. С. 84–96.

31. *Национальный* состав населения Томской области. URL: http://worldgeo.ru/ russia/ lists /?id=33&code=70

32. *Языки* народов Сибири, находящиеся под угрозой исчезновения. URL: http:// lingsib.iea.ras.ru/ru/languages/ket.shtml

33. *Томская* диалектологическая школа: историограф. очерк. Томск. Изд-во Том. ун-та, 2006.

34. *Резанова З.И.* Дискурсивные стратегии презентации национально-культурной идентичности // Вестн. Том. гос. ун-та. Культурология и искусствоведение. 2012. № 4 (8). С. 40–54.

35. *Резанова З.И.* Институциональная и личностная презентация национально-культурной идентичности в интернет-коммуникации: жанровые формы и дискурсивные стратегии // Вестн. Том. гос. ун-та. 2013. № 375. С. 33–41.

36. *Костяшина Е.А., Ермоленкина Л.И.* Коммуникативно-языковые механизмы формирования этнокультурной идентичности в дискурсивном пространстве Интернета // Вестн. Том. гос. ун-та. Культурология и искусствознание. 2013. №3 (11). С. 5–16.

"TOMSK REGIONAL CORPUS": TYPOLOGICALLY RELEVANT PARAMETERS OF BALANCE AND REPRESENTATIVENESS.

Tomsk State University Journal of Philology, 2015, 1(33), 38–50. DOI 10.17223/19986645/33/4

Rezanova Zoya I., Tomsk State University, Tomsk Polytechnic University (Tomsk, Russian Federation). E-mail: resso@rambler.ru / resso@mail.tsu.ru / rezanovazi@mail.ru

Keywords: corpus linguistics, balance of corpus, representativeness of corpus, regional corpus, regional linguistics.

The article presents the typologically relevant features of Tomsk Regional corpus produced currently in Tomsk State University. The features are determined by the nature of the interpretation of the balance of the texts of the corpus which reflect the structure of communication in the region.

This interpretation assumes an increase of transcripts of oral communication in the text part of the corpus, first of all, of different genres of everyday oral non-institutional communication.

The project is expected to achieve the balance parameter by a considerable increase of the share of different genres of everyday communication, including those at the intersection of institutional and personal communication (letter of appeal to official institutions, minutes of meetings, etc.). The balance should also be achieved by a wider inclusion in the structure of a corpus of new modern types of communication mediated by various means: text messages, a variety of genres of Internet communication.

The authors of the project interpret balance, on the one hand, within the literary language, but aim to reflect the structure of communication in the region as a formation which is non-uniform in terms of normalization and codification.

The corpus represents the functioning of the literary language in the region as a dynamic interaction with various forms of the national language, which should be combined with the reflection of the phenomena of contamination with contact languages.

The interpretation of balance of Tomsk Regional Corpus correlates with a number of typologically important parameters: the type of data: mixed, that is it includes written and transcribed spoken texts: text language: Russian, parallelism: non-parallel (monolingual), "literariness": represents proper literary, colloquial, conversational, sometimes special and other texts; genres: aim to represent the genre variety of different functional and stylistic registers of speech; chronological aspect: synchronic, with elements of micro-diachrony; dynamism: elements of synchronous dynamics: markup: with marks; markup character: morphological, syntactic, semantic; volume of texts: full text; purpose: research; availability: available.

References

1. Mishankina N.A. Linguistic corpus "Tomsk regional text": theoretical and methodological background of the project. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*, 2014, no. 389, pp. 28–37. (In Russian).

2. Sologub O., Rezanova Z.I, Temnikova I.G. The Concept of the Tomsk Regional Corpus: Balance and Representativeness. *Procedia – Social and Behavioral Sciences*, 2014, 154, pp. 175–178. DOI: 10.1016/j.sbspro.2014.10.131

3. Zakharov V.P. *Korpusnaya lingvistika* [Corpus Linguistics]. St. Petersburg: St. Petersburg State University Publ., 2005. 48 p.

4. *British National Corpus*. Available from: http://www.natcorp.ox.ac.uk.

5. *International Corpus of English*. Available from:

6. *Bank of English*. Available from: http://www.collins.co.uk/Corpus/CorpusSearch.aspx.

7. *Natsional'nyy korpus russkogo yazyka* [Russian National Corpus]. Available from: http://ruscorpora.ru.

8. *Natsional'nyy korpus pol'skogo yazyka* [National Corpus of Polish]. Available from: http://nkjp.pl.

9. *Slovatskiy natsional'nyy korpus* [Slovak National Corpus]. Available from: http://korpus.juls.savba.ski.

10. *Cheshskiy natsional'nyy korpus* [Czech National Corpus]. Available from: http://ucnk.ff.cuni.cz

11. *Polish and English Language Corpora for Research and Applications*. Available from: http://korpus.ia.uni.lodz.pl.

12. *Sankt-Peterburgskiy korpus agiograficheskikh tekstov* [St. Petersburg Corpus of Hagiographic Texts of XV-XVII centuries]. Available from: http://project.phil.pu.ru/skat.

13. *Chemnitz German-English Translation Corpus*. http://www.tu-chemnitz.de/ phil/ InternetGrammar.

14. Reznikova T.I. *Korpusa slavyanskikh yazykov v internete: Obzor resursov* [Corpora of Slavic languages on the Internet: An Overview of resources]. *Die Welt der Slaven*, 2008, LIII, pp. 10–38.

15. Reznikova T.I. *Slavyanskaya korpusnaya lingvistika: sovremennoe sostoyanie resursov* [Slavic corpus linguistics: the current state of resources]. In: *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006-2008. New results and prospects]. St. Petersburg: Nestor-Istoriya Publ., 2009, pp. 402–461.

16. Reznikova T.I., Kopotev M.V. *Lingvisticheski annotirovannye korpusa russkogo yazyka (obzor obshchedostupnykh resursov)* [Linguistically annotated corpora of the Russian language (review of public resources)]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus: 2003–2005]. Moscow: Indrik Publ., 2005, pp. 31–61.

17. Sharov S.A. Predstavitel'nyy korpus russkogo yazyka v kontekste mirovogo opyta [The representative corpus of the Russian language in the context of international experience]. *NTI, Seriya 2*, 2003, no. 6, pp. 9–17.

18. Belyaeva L.N. *Leksikograficheskiy potentsial parallel'nogo korpusa tekstov* [Lexicographical potential of parallel corpus of texts]. Available from: corpora.iling.spb.ru/Docs/Belyaeva_corpora.pu.

19. Pertsov N.V. [On the role of corpora in linguistic studies]. *Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2006"* [Proceedings of the International Conference "Corpus Linguistics-2006"]. St. Petersburg, 2006, pp. 318–331. (In Russian).

20. Pertsov N.V. K suzhdeniyam o faktakh russkogo yazyka v svete korpusnykh dannykh [On judgments on the facts of the Russian language in the light of corpus data]. *Russkiy yazyk v nauchnom osveshchenii*, 2006, no. 1(11), pp. 227–245.

21. Plungyan V.A. Korpus kak instrument i kak ideologiya: o nekotorykh urokakh sovremennoy korpusnoy lingvistiki [Corpus as a tool and as an ideology: some lessons of modern corpus linguistics]. *Russkiy yazyk v nauchnom osveshchenii*, 2008, no. 16 (2), pp. 7–20.

22. Letuchiy A.B. *Korpus dialektnykh tekstov: zadachi i problemy* [Corpus of dialect texts: challenges and problems]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005* [Russian National Corpus: 2003–2005]. Moscow: Indrik Publ., 2005, 215–232. Available from: http:// ruscorpora.ru/ sbornik2005/13letuchy.pdf.

23. Letuchiy A.B. *Dialektnyy korpus: sostav i osobennosti razmetki* [Dialectal corpus: composition and markup features]. In: *Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006-2008. New results and prospects]. St. Petersburg: Nestor-Istoriya Publ., 2009, pp. 114–128.

24. Kryuchkova O.Yu., Gol'din V.E. [Corpus of Russian dialectal speech: concept and estimation parameters]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 25–29 maya 2011 g.)* [Computational linguistics and intelligent technologies: proc. of the annual International Conference "Dialogue" (Bekasovo, May 25–29, 2011)]. Moscow: Russian State Humanitarian University Publ., 2011, issue 10 (17), pp. 359–367. (In Russian).

25. Sichinava D.V., Kachinskaya I.B. [Corpus of dialect texts in Russian National Corpus: current status and prospects]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 4–8 iyunya 2014 g.)* [Computational linguistics and intelligent technologies: proc. of the annual International Conference "Dialogue" (Bekasovo, June 4–8, 2014)]. Moscow: Russian State Humanitarian University Publ., 2011, issue 13 (20), pp. 620–629. (In Russian).

26. Yurina E.A. Tomsk dialectal corpora: the starting point. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*, 2011, no. 2 (14), pp. 58–64. (In Russian).

27. Zemskaya E.A. *Russkaya razgovornaya rech': lingvisticheskiy analiz i problemy obucheniya* [Russian colloquial speech: linguistic analysis and teaching problems]. Moscow: Russkiy yazyk Publ., 1979. 240 p.

28. Zemskaya E.A., Kitaygorodskaya M.V., Shiryaev E.N. *Russkaya razgovornaya rech': Obshchie voprosy. Slovoobrazovanie. Sintaksis* [Russian colloquial speech: general issues. Word formation. Syntax]. Moscow: Nauka Publ., 1981. 276 p.

29. Bogdanova N.V., Brodt I.S., Kukanova V.V., Pavlova O.V., Sapunova E. M., Filippova N.S. [On the "corpus" of live speech: principles of formation and opportunities of describing]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (2008)* [Computational linguistics and intelligent technologies: proc. of the annual International Conference "Dialogue" (2008)]. Moscow: Russian State Humanitarian University Publ., 2008, issue 7 (14), pp. 57–61. (In Russian).

30. Belikov V.I., Kopylov N.Yu., Piperski A.Ch., Selegey V.P., Sharov S.A. [Corpus as a language: from scalability to differential completeness]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (2008)* [Computational linguistics and intelligent technologies: proc. of the annual International Conference "Dialogue" (2008)]. Moscow: Russian State Humanitarian University Publ., 2013, pp. 84–96. (In Russian).

31. *Natsional'nyy sostav naseleniya Tomskoy oblasti* [Ethnic composition of the population of Tomsk Oblast]. Available from: http://worldgeo.ru/russia/lists/?id=33&code=70.

32. *Yazyki narodov Sibiri, nakhodyashchiesya pod ugrozoy ischeznoveniya* [Endangered languages of the peoples of Siberia]. Available from: http://lingsib.iea.ras.ru/ru/languages/ket.shtml.

33. Blinova O.I. (ed.) *Tomskaya dialektologicheskaya shkola: Istoriograficheskiy ocherk* [Tomsk Dialectological School: historiographical essay]. Tomsk: Tomsk State University Publ., 2006. 392 p.

34. Rezanova Z.I. Discourse strategies of presentation of national cultural identity. *Vestnik Tomskogo gosudarstvennogo universiteta. Kul'turologiya i iskusstvovedenie – Tomsk State University Journal of Cultural Studies and Art History*, 2012, no. 4(8), pp. 40–54. (In Russian).

35. Rezanova Z.I. Institutional and personal presentation of national and cultural identity in Internet communication: genre forms and discursive strategies. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*, 2013, no. 375, pp. 33–41. (In Russian).

36. Kostyashina E.A., Ermolenkina L.I. Communicative and linguistic mechanisms of ethnic and cultural identity in the discourse space of the Internet. *Vestnik Tomskogo gosudarstvennogo universiteta. Kul'turologiya i iskusstvovedenie – Tomsk State University Journal of Cultural Studies and Art History*, 2013, no. 3 (11), pp. 5–16. (In Russian).