УДК 81’27
DOI: 10.17223/22274200/11/7

З.И. Резанова

ПОДКОРПУС УСТНОЙ РЕЧИ РУССКО-ТЮРКСКИХ БИЛИНГВОВ ЮЖНОЙ СИБИРИ: ТИПОЛОГИЧЕСКИ РЕЛЕВАНТНЫЕ ПРИЗНАКИ[1]

В статье определяются типологические релевантные признаки подкорпуса устной речи русско-тюркских билингвов (русско-татарский, русско-шорский, русско-хакасский билингвизм), создаваемого в рамках проекта исследования языков Южной Сибири. Подкорпус рассматривается как эмпирическая база лингвистического, психолингвистического и лингвокогнитивного изучения билингвальной (мультилингвальной) интерференции. Характеризуются основной принцип отбора текстов в подкорпус, принципы метаразметки корпуса, обусловленные исследовательской направленностью проекта.
Ключевые слова: *корпус билингвов, русско-тюркский билингвизм, метаразметка, устный подкорпус, языки Южной Сибири.*

Корпусная лингвистика, идеи которой были высказаны еще в 60-е гг. XX в., начала интенсивно развиваться с 80-х гг. с опорой на активно совершенствующиеся информационные технологии. В настоящее время лингвистические корпуса текстов (собрания текстов, отобранные и размеченные на основании теоретически обоснованных лингвистически релевантных принципов) играют все более значимую роль в качестве источников лингвистического исследования, конкурируя с традиционным источником – словарем. Так как лингвистический корпус, как правило, включает морфологическую, лексико-семантическую и стилистическую квалификацию слова, он в качестве источника лингвистического исследования соотносим с так называемыми аспектными словарями, включающими единицы и их квалификацию определенного типа. Вместе с тем основной особенностью корпуса как лингвистического источника является то, что он дает возможность извлечь информацию о широчайшем спектре контекстной реализации языковых единиц, их форм, вариантов. Принципиальное противопоставление принципов отбора единиц словаря и корпуса – словоцентричность и текстоцентричность – делают их взаимодополняемыми источниками, совместное

[1] Исследование выполнено при поддержке гранта Министерства образования и науки РФ, договор №14Y26.31.00.14.

использование которых способствует повышению эффективности лингвистического исследования.

В современной интенсивно развиваемой корпусной лингвистике все большее развитие получают корпуса текстов, собранные в качестве эмпирической основы разнообразных лингвистических проектов, что мотивирует разработку особых принципов разметки и метаразметки текстового материала, которые дополняют выработанные ранее стандарты его обработки.

В совместной статье с Г.Ю. Весниной мы уже писали о подкорпусе русской речи билингвов, рассматривая его как часть долгосрочного проекта создания лингвистического корпуса «Томский региональный текст», целью которого являлось собирание текстов регионального варианта русского языка во всей полноте его функциональной дифференциации [1]. В совокупности функционально обусловленных речевых вариантов реализации русского языка значительное место занимает билингвальная речь, формы которой обусловлены многими факторами, в том числе типом вступающих во взаимодействие языков и характером их контактирования, что, в свою очередь, зависит от действия значительного количества социолингвистических параметров коммуникации. Будучи частью крупного проекта, подкорпус речи билингвов, по замыслу авторов, наследовал основные принципы отбора текстов основного корпуса, его разметки и метаразметки, структура которых определялась стремлением отразить полноту функционального варьирования территориального варианта национального языка. К таким принципам мы относили: при отборе текстов – отражение устной и письменной речи всех функциональных стилей, стремление к жанровому и тематическому разнообразию, при планировании разметки – введение дискурсивно значимых маркеров[1]. Подкорпус речи билингвов закономерно дополнялся параметрами метаразметки, введением базовых показателей контактирующих языков в билингвальной речи, в аннотировании корпуса – введением системы тегов отклонений от речевого стандарта [1].

Концепция данного подкорпуса получила свое развитие в рамках крупного проекта по изучению языкового многообразия Южной Сибири, поддержанного грантом Правительства Российской Федерации. Значимую часть программы исследований составляют сбор,

[1] Принципы формирования данного корпуса, реализуемого в настоящее время, охарактеризованы в ряде работ авторов проекта, см.: [2–5].

инвентаризация и исследование форм контактирования языков сквозь призму речевых структур русскоязычных билингвов с интерферентными проявлениями тюркских языков: татарского, хакасского, шорского.

В проекте предполагается лингвистическое, психолингвистическое и лингвокогнитивное исследование форм проявления билингвальной (мультилингвальной) интерференции. При решении комплекса данных задач формирование корпуса русскоязычной билингвальной речи имеет, во-первых, самостоятельную научную ценность, так как его результатом станет фиксация типов речевых отклонений на всех уровнях языковой системы в соотнесении с типами языкового контактирования. Во-вторых, мы рассматриваем сбор билингвальных текстов и их лингвистическую разметку в качестве основного источника при формировании материалов психолингвистических экспериментальных исследований механизмов контактирования языков в когнитивных структурах билингвов.

Данная фокусировка исследовательских задач обусловила необходимость корректировки архитектуры формируемого подкорпуса, расширения системы метаразметки и увеличения количества тегов его аннотирования, а также необходимость дальнейшего углубления структуры самого подкорпуса. В структуре основного подкорпуса русской речи билингвов мы выделяем: 1) на основании различий контактирующих языков – подкорпуса речи русско-татарских, русско-хакасских, русско-шорских билингвов; 2) в каждом из подкорпусов противопоставляются далее более частные подкорпуса по противопоставлению основных форм (модусов) речи – устной и письменной.

Большинство из существующих в настоящее время наиболее представительных корпусов национальных языков являются собранием прежде всего письменных текстов, к которым присоединяются подкорпуса устные или мультимедийные. В концепции теоретической направленности представляемого корпуса русской речи билингвов именно *подкорпус устной речи* является ядерным, так как только при формировании его текстовой основы можно максимально полно зафиксировать психолингвистически и социолингвистически значимые особенности билингва, что является необходимым при исследовании когнитивных эффектов билингвального взаимодействия. Весьма значимой является также возможность привлечь одних и тех же информантов в качестве как авторов текстов корпуса, так и

испытуемых при проведении поведенческих психолингвистических исследований билингвизма.

В данной статье характеризуются основные признаки создаваемых устных подкорпусов русско-татарских, русско-хакасских, русско-шорских билингвов, подходы к отбору текстового материала и общие принципы метаразметки, определяющие их место в типологии лингвистических корпусов.

Как справедливо отмечает М. Копотев, при отборе текстов корпуса «единственный критерий – задача, для которой собран корпус» [6. С. 33]. Далее мы охарактеризуем принципы отбора материала в формируемый корпус, сопоставляя его с другими вариантами селекции текстов. Отметим, что мы ориентировались на принятую в мировой практике корпусного проектирования типологию, разработанную в проекте EAGLES [7][1], однако характеризуем корпус только по параметрам, значимым для спецификации представляемого подкорпуса устной речи билингвов.

Лингвистические корпуса противопоставляются:

– по языкам коллекций текстов: одноязычные vs многоязычные;

– формам речи: устные vs письменные vs мультимодальные vs смешанные;

– формам национального языка: литературный язык vs диалектный язык vs. недифференцированные по данному признаку;

– дискурсивным и, реже, жанровым формам текстов, обычно современные глобальные проекты национальных корпусов стремятся к сбалансированности материалов, т.е. к отражению объемов жанровых и дискурсивных форм текстов пропорционально соотношению типов коммуникации на данном языке (см. обоснование таких принципов компоновки текстового массива в НКРЯ [8], а также примеры таких корпусов в работах [9–13]);

– временной отнесенности текстов, по этому основанию противопоставляются корпуса современных текстов (временные границы при этом определяются в соответствии с выработанными в теоретической лингвистике положениями о темпах динамики национальных языков) и являющиеся собраниями текстов, созданных в другие эпохи (пример такого корпуса – лингвистически размеченное собрание агиографических текстов XV–XVII вв. [14]);

[1] Отметим, что на данный стандарт ориентируются все обсуждаемые в статье корпуса текстов.

– типам соотношения языков, используемых автором текстов при их порождении.

Основным дифференциальным параметром, определяющим направленность отбора текстов для подкорпуса русско-тюркских билингвов Южной Сибири, является последний признак. По этому признаку корпуса делятся на тексты носителей языка, к которым относится абсолютное большинство создаваемых корпусов, и так называемые корпуса второго языка, которые являются собранием «текстов не носителей языка». К последней группе относят учебные корпуса, корпуса билингвов и корпуса лингва франка [6. С. 103–107]. Определение «тексты не носителей языка» представляется нам не совсем удачным, более верное определение, на наш взгляд, – «тексты носителей нескольких языков», т.е. это тексты, которые порождаются билингвами с разным соотношением материнского (L1) и изучаемого или освоенного языка (L2). Так, наиболее представительный англоязычный The Cambridge Learner Corpus (CLC) включает к настоящему времени 40 млн текстоформ, записи речи более 200 тыс. студентов их 217 стран, говорящих на 148 родных языках [6. С. 105].

При этом язык, тексты на котором объединяет соответствующий корпус, может занимать функционально различное положение в коммуникации билингва. Например, в составе русскоязычных корпусов этого типа русский язык может быть материнским, испытывающим влияние других (другого) языка при эмиграции, это так называемые корпуса текстов эритажных (херитажных) носителей языка (использование данных такого корпуса см. в работе [15]). Целевой язык может быть и вторым, изучаемым языком, как, например, в учебных корпусах (русскоязычный учебный корпус RLC) [16].

В корпусе билингвов, являющемся подкорпусом национального корпуса болгарского языка, собраны тексты, написанные на болгарском языке, подъязыке русско-болгарских билингвов разного типа, проживающих в Болгарии, для которых русский язык является родным, материнским, а болгарский – осваиваемым (см. проект данного корпуса в статье К. Петровой [17]).

В характеризуемом в статье подкорпусе собраны русскоязычные тексты, создаваемые людьми, для которых русский язык не является материнским, материнские языки контактирования – языки тюркской семьи: шорский, татарский, хакасский.

При отмеченном выше типологическом сходстве учебных корпусов и корпусов речи билингвов, данные корпуса имеют значительные отличия в целевой направленности отбора текстов, как следствие – в типе преобладающих текстов. Ядро учебных корпусов составляют, как правило, письменные тексты – студенческие работы крупных учебных центров, во вторую очередь – транскрибированные устные тексты. В подкорпусе русско-тюркских билингвов Южной Сибири ядро корпуса составляют устные тексты.

Несомненна прикладная направленность создания учебных корпусов на выявление типичных отклонений в использовании языка носителями разноструктурных языков как основа коррекции методик преподавания языка. При этом данные корпусов широко используются и в типологических исследованиях.

Своеобразие представляемого в статье собрания текстов состоит в том, что русский язык, являясь вторым, не материнским, активно используется авторами текстов корпуса во многих сферах, прежде всего в институциональной коммуникации. Вследствие этого, мы полагаем, отклонения от норм использования русского языка могут носить более глубокий, неявный характер, нежели в ученических работах, и выявляются на основе концентрации значительного объема текстов. Корпус планируется как собрание текстов, являющееся основой типологических и психолингвистических исследований, однако его данные также могут быть использованы в практике преподавания русского языка в школах.

Итак, подкорпус русско-тюркских билингвов Южной Сибири по названным выше признакам противопоставления может быть определен как собрание современных устных текстов, отбор которых проводится с ограничением по локальному принципу: авторы текстов являются носителями локального варианта русского языка, как литературной, так и нелитературной (диалектной, просторечной) форм. Подчеркнем фиксацию регионального характера собираемых текстов, так как полагаем, что тип языкового контактирования может отражать аспекты диалектного взаимовлияния, например сибирского варианта татарского языка и среднеобских говоров русского языка.

Создаваемый подкорпус устной речи, естественно, не может охватить всю палитру дискурсивных и жанровых форм коммуникации, в нем преимущественно будут собраны тексты устного бытового и, реже, публицистического общения. Дискурсивное, жанровое и тема-

тическое ограничение текстов определяется преимущественным типом сбора материала – в практике интервьюирования, бесед собирателей текстов с информантами, самозаписи информантом разных форм обыденной коммуникации.

Направленность на изучение языковой интерференции в речи билингва, на исследование отражения когнитивных процессов контактирования языков в сознании билингва определяет тип разметки и метаразметки подкорпуса. Проблема разметки (аннотирования) корпуса требует отдельного рассмотрения, в данной статье мы охарактеризуем только принципы метаразметки корпуса, определяемые типом создаваемого подкорпуса.

Целевая направленность создаваемого корпуса потребовала коррекции системы метаразметки по отношению как к метаразметке НКРЯ, так и к учебным корпусам и корпусам ошибок. Метаразметка определяет структуру корпуса, помогает контролировать его наполняемость (репрезентативность и сбалансированность). При формировании параметров метаразметки мы также следовали принятым в корпусной практике принципам соответствия цели и принципу полноты, определяемому относительно цели.

По отношению к НКРЯ метаразметка корпуса устной речи билингвов отличается, с одной стороны, существенной редукцией, с другой стороны, введением дополнительных параметров. Как известно, метаразметка в НКРЯ включает 25 параметров, которые распределяются по трем группам: информация об авторе текста, информация о тексте и служебная информация (см. полное описание принципов метаразметки в НКРЯ в статье С.О. Савчук [18]). Значительное количество параметров метаразметки в НКРЯ мотивировано стремлением авторов корпуса соответствовать принципу полноты дискурсивных и жанровых форм репрезентации текстов в корпусе, что требует введения дополнительных признаков при параметризации устных и письменных текстов разной дискурсивной отнесенности и жанровой природы.

В представляемом устном подкорпусе речи билингвов ограничение жанров определяет и ограничение признаков текстов, включаемых в метаразметку; необходимость же отражения явлений интерферентных проявлений межъязыкового взаимодействия в корпусе требует маркирования социолингвистических и психолингвистических факторов, их определяющих. Вследствие этого в метаразметку включается не только обычная для большинства корпусов информа-

ция об авторах текстов (дата рождения, пол, образование, социальное положение), но также информация о языках, которыми владеет автор текста, и об их функциональном соотношении.

При метаразметке подкорпуса мы используем данные двух анкет, которые заполняет информант: социолингвистической анкеты, разработанной в Институте языкознания РАН, основанной на анкетах О.А. Казакевич и используемой при исследовании языков малых народов Российской Федерации [19].

Социолингвистическая анкета включает 41 вопрос с более подробной детализацией информации об авторе текста: о времени и месте рождения, проживания, обучения, профессиональной деятельности, сведений о родственниках по разным типам родства, о способе приобретения и использования языков.

Языковая анкета билингва, разработанная на основе анкеты языкового опыта и уровня владения языком Marian V., Blumenfeld H.K., Kaushanskaya M. [20], включает 14 блоков параметризации характера и типа владения билингвом взаимодействующими языками: языки ранжируются по мере активности их использования, порядку усвоения, количеству времени пользования языками во время интервьюирования, по предпочтению выбора языков при чтении и при коммуникации с другим человеком; также фиксируется информация об истории пользования языком – о времени изучения или вхождения в язык, о времени пребывания информанта в среде языка, о самооценке информантом уровня владения языком и факторов, стимулирующих изучение каждого из языков, которым владеет информант, о предпочитаемых темах и сферах коммуникации для говорения на каждом из языков.

Как видим, социолингвистическая и психолингвистическая информация двух анкет, заполняемых информантами корпуса, имеет пересекающиеся, хотя и нетождественные параметры, однако вследствие того, что эти анкеты имеют внутреннюю, присущую им системность, определяемую соотнесенными, но разными исследовательскими парадигмами, мы включаем в материалы корпуса обе анкеты. При этом в метаразметку корпуса вносятся только основные параметры, определяющие существенные аспекты взаимодействия языков в когнитивной и коммуникативной системе билингва. Данный фрагмент метаразметки мы структурируем относительно владения русским языком, на котором говорит информант, определяя его статус: является ли русский язык по самооценке информанта род-

ным (материнским) или неродным; используется ли в период записи текста в разных формах коммуникации (активный vs пассивный), порядок усвоения языка (первый vs второй); сфера преимущественного использования (письменная vs устная; бытовая vs официальная vs эстетическая vs другие. По тем же параметрам оцениваются другие языки речевых практик билингва (полилингва).

Отметим, что через систему отсылок пользователь корпуса при необходимости может получить доступ к расширенному составу информации, полному содержанию анкет, за исключением фамилии, имени, отчества авторов, которые открыты только в качестве служебной информации, в пользовательской системе они представлены в закодированном виде.

Информация о тексте включает данные о времени и месте записи текста, его размере в словах.

Далее следуют лингвистически релевантные параметры текста. Теоретическая нейтральность – ведущий принцип аннотирования текстов корпуса. Наиболее общий признак параметризации – форма (модус) текста, с базовой оппозицией письменной и устной форм, которая в настоящее время дополняется третьим видом – тексты мультимедийной коммуникации. Фундаментальность противопоставления данных типов (модусов) коммуникации обоснована и в психолингвистических исследованиях процессов порождения и восприятия речи [21. С. 270], и в дискурсивных исследованиях [22. С. 16–17].

Следующие параметры – тип коммуникации (монолог, диалог, полилог); тип дискурса (личностный – институциональный) и его конкретные виды – параметризируются в соответствии с традицией, сложившейся в зарубежной и российской практике социолингвистических исследований, см., например, работы [23–25].

Решая вопрос о жанровой принадлежности текстов, мы основываемся на типологическом членении, обоснованном Т.В. Шмелевой, выделяющей информативные, оценочные, этикетные, императивные жанры [26]. Конкретные виды жанров диагностируются по жанрообразующим признакам (рассказ, беседа, воспоминание, сообщение, разговор и др.). При разметке по данному признаку участники проекта получают инструкцию с описанием признаков идентификации речевого жанра. Завершается метаразметка указанием на тему текста: свадьба, встреча с друзьями, мои родители, учеба в школе, посещение театра (список открыт).

Раздел метаразметки «Служебная информация» включает условное название подкорпуса по соответствующему варианту билингвизма (русско-татарский, русско-шорский, русско-хакасский), по форме (модусу) речи (устный), а также указываются имена, отчества, фамилии лингвистов, ответственных исполнителей: тех, кто записывал аудиофайл, собирал анкеты, производил письменную расшифровку, метаразметку и автоматическое аннотирование, ручную разметку интерференции, проверку разметки после автоматической обработки при автоматической разметке текста.

В настоящее время проводится тестирование системы разметки и метаразметки текста.

Создаваемый корпус русской речи билингвов станет первым репрезентативным собранием текстов билингвальной речи данного типа: русскоязычной речи, испытывающей интерферентное влияние материнских языков тюркской группы, которые в настоящее время являются средством обыденного общения вследствие особенностей языковой ситуации региона, послужит ценным источником научных исследований в области социолингвистического, психолингвистического и когнитивного аспектов языкового взаимодействия.

Литература

1. *Резанова З.И., Веснина Г.Ю.* Подкорпус русской речи билингвов лингвистического корпуса «Томский региональный текст»: принципы разметки и метаразметки корпуса // Вопр. лексикографии. – 2016. – № 1 (9). – С. 29–39.

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

3. *Резанова З.И.* Лингвистический корпус «Томский региональный текст»: типологически релевантные параметры сбалансированности и репрезентативности // Вестн. Том. гос. ун-та. Филология. – 2015. – № 1 (33). – С. 38–50.

4. *Sologub O., Rezanova Z., Temnikova I.* 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.

5. *Мишанкина Н.А., Филь Ю.В.* Лингвистический корпус «Томский региональный текст»: концепция и структура // Слово: Фольклорно-диалектологический альманах: Материалы научных экспедиций. – Вып. 12. – Благовещенск, 2015. – С. 38–49.

6. *Копотев М.* Введение в корпусную лингвистику. – Прага, 2014. – 194 с.

7. *Sinclair J.* EAGLES. Preliminary recommendations on Corpus Typology. EAG--TCWG--CTYP/P. Version of May, 1996. – URL: http:// www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (дата обращения: 05.05.2017).

8. *Национальный* корпус русского языка. – URL: http://www.ruscorpora.ru/ (дата обращения: 05.05.2017).

9. *Līdzsvarots* mūsdienu latviešu valodas tekstu korpuss. – URL: http://www.korpuss.lv/ (дата обращения: 05.05.2017).

10. *Британский* национальный корпус British National Corpus. – URL: http://www.natcorp.ox.ac.uk (дата обращения: 05.05.2017).

11. *Международный* корпус английского языка = International Corpus of English. – URL: http:// icecorpora.net (дата обращения: 05.05.2017).

12. *Словацкий* национальный корпус. – URL: http://korpus.juls.savba.sku (дата обращения: 05.05.2017).

13. *Чешский* национальный корпус. – URL: http://ucnk.ff.cuni.cz (дата обращения: 05.05.2017).

14. *St. Petersburg* Corpora of hagiographic Texts XV–XVII centuries. – URL: http:// project. phil. pu. ru/skat) 13. (дата обращения: 05.05.2017).

15. *Полинская М., Рахилина Е.В., Выренкова А.С.* Грамматика ошибок и грамматика конструкций: «эритажный» («унаследованный») русский язык // Вопр. языкознания. – 2014. – № 3. – С. 3–19.

16. *The Russian* Learner Corpus (RLC). – URL: http://web-corpora.net/RussianLearnerCorpus/search/. http://web-corpora.net/RLC (дата обращения: 05.05.2017).

17. *Петрова К.* Проект о создании корпуса устной речи русско-болгарских билингвов. – URL: http://www.dialog-21.ru/media/2727/petrova.pdf (дата обращения: 05.05.2017).

18. *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.

19. *Социолингвистическая* анкета. – URL: http://iling-ran.ru/main/departments/ural-altaic (дата обращения: 05.05.2017).

20. *Marian V., Blumenfeld H.K., Kaushanskaya M.* Language Experience and Proficiency Questionnaire (LEAP-Q) // Speech Language and Hearing Research, 50 (4). P. 940–967. – URL: http://www.bilingualism.northwestern.edu/leapq/ (дата обращения: 05.05.2017).

21. *Лурия А.Р.* Язык и сознание. – Ростов н/Д: Феникс, 1998. – 319 с.

22. *Кибрик А.А.* Анализ дискурса в когнитивной перспективе: дис. в виде науч. докл. …. д-ра филол. наук. – М., 2003. – 90 с.

23. *Фуко М.* Воля к истине: по ту сторону знания, власти и сексуальности: Работы разных лет. – М.: Касталь, 1996. – 446 с.

24. *Макаров М.Л.* Основы теории дискурса. – М.: ИТДГК «Гнозис», 2003. – 280 с.

25. *Карасик В.И.* О типах дискурса // Языковая личность: институциональный и персональный дискурс. – Волгоград, 2000. – С. 5–20.

26. *Шмелева Т.В.* Модель речевого жанра // Жанры речи. – Саратов, 1997. – Вып. 1. – С. 88–99.

SUBCORPUS OF ORAL SPEECH OF RUSSIAN-TURKIC BILINGUALS OF SOUTHERN SIBERIA: TYPOLOGICALLY RELEVANT SIGNS

Voprosy leksikografii – Russian Journal of Lexicography, 2017, 11, pp. 105–118.
DOI: 10.17223/22274200/11/7

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

Keywords: corpus of bilinguals, Russian-Turkic bilingualism, meta-marking, oral subcorpus, languages of South Siberia.

The article describes the typological relevant parameters of the oral speech subcorpus of Russian-Turkic bilinguals (Russian-Tatar, Russian-Shor, Russian-Khakass bilingualism) which is being created in the framework of the South Siberian languages research project. This corpus will be used as an empirical basis for linguistic, psycholinguistic and linguocognitive studies of the forms of bilingual (multilingual) interference.

The author used the EAGLES project standard as the basis for the typological specification of this corpus. The most important signs of the subcorpus are: language of the text subcorpus: monolingual subcorpus (Russian language texts); form (mode) of speech: transcription of oral speech; form of the represented national language: the local version of the Russian language; discursive forms of texts: texts of oral everyday and (more rarely) journalistic communication; genre forms of texts: story, conversation, reminiscence, etc.; time attribution of texts: modern; types of correlation of languages the author of the texts used: Russian-language texts created by people with Russian as a non-mother tongue, languages of contacting: languages of the Turkic family: Shor, Tatar, Khakass.

The texts of the subcorpus should reflect the sociolinguistic and psycholinguistic factors of interference in interlingual interaction; this influences the characteristics of its meta-marking. Meta-markup includes not only the generally accepted information (date of birth, sex, education, social position of the informant), but also information on the languages that the author of the text speaks, and their functional relationship.

The author uses two informant questionnaires to form this aspect of meta-markup 1) the sociolinguistic questionnaire and 2) the language questionnaire of a bilingual. The first questionnaire includes 41 questions detailing information about the informant: the time and place of birth, residence, education, professional activity, relatives for various types of kinship, the nature of the acquisition and use of languages.The second questionnaire was developed by Marian, Blumenfeld and Kaushanskaya. This questionnaire includes 14 blocks of signs characterizing the acquisition and use of languages by the informant. In the meta-markup of the subcorpus, the author includes only the basic parameters from these questionnaires. These are the signs that determine the essential aspects of the interaction of languages in the cognitive and communicative bilingual system: whether the language is native (parent) or non-native according to the self-evaluation of the informant; is it used actively VS passively during the recording of the text in different forms of communication, the first VS second order of mastering the language; written VS oral; household VS official VS aesthetic VS other spheres of preferential use.

The user of the subcorpus can access the extended information, the full content of the questionnaires via the system of references. Parametrization of the text includes the following features: the form (mode) of the text (oral), the type of communication (monologue vs. dialogue vs. polylogue); type of discourse (personal vs. institutional), types of speech genres (informative vs. evaluative vs. etiquette vs. imperative); specific genres (story vs. conversation vs. reminiscence vs. message vs. conversation, etc.); theme of the text (wedding vs. meeting with friends vs. my parents vs. schooling vs. visiting a theater, etc.).

References

1. Rezanova, Z.I. & Vesnina, G.Yu. (2016) Meta-data and annotation design of the Russian-speaking bilinguals speech subcorpus in the structure of the Tomsk Regional

Corpus. *Voprosy leksikografii – Russian Journal of Lexicography*. 1 (9). pp. 29–39. (In Russian). DOI: 10.17223/22274200/9/3

2. Mishankina, N.A. (2014) Linguistic corpus "Tomsk regional text": theoretical and methodological background of the project. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. 389. pp. 28–37. (In Russian). DOI: 10.17223/15617793/389/4

3. Rezanova, Z.I. (2015) Tomsk Regional Corpus: typologically relevant parameters of balance and representativeness. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology*. 1 (33). pp. 38–50. (In Russian).

4. Sologub, O., Rezanova, Z. & Temnikova, I. (2014) The Concept of the Tomsk Regional Corpus: Balance and Representativeness. *Procedia – Social and Behavioral Sciences*. 154. pp. 175–178.

5. Mishankina, N.A. & Fil', Yu.V. (2015) Lingvisticheskiy korpus "Tomskiy regional'nyy tekst": kontseptsiya i struktura [Linguistic corpus "Tomsk Regional Text": concept and structure]. *Slovo: Fol'klorno-dialektologicheskiy al'manakh. – Materialy nauchnykh ekspeditsiy*. 12. pp. 38–49.

6. Kopotev, M. (2014) *Vvedenie v korpusnuyu lingvistiku* [Introduction to corpus linguistics]. Prague: Animedia.

7. Sinclair, J. (1996) *EAGLES. Preliminary recommendations on Corpus Typology. EAG–TCWG–CTYP/P. Version of May, 1996.* [Online]. Available from: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html. (Accessed: 05.05.2017).

8. The National Corpus of the Russian Language. [Online]. Available from: http://www.ruscorpora.ru/. (Accessed: 05.05.2017).

9. Korpuss.lv. (c. 2007) *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss* [Balanced textbook of the contemporary Latvian language]. [Online]. Available from: http://www.korpuss.lv/. (Accessed: 05.05.2017).

10. The British National Corpus. [Online]. Available from: http://www.natcorp.ox.ac.uk. (Accessed: 05.05.2017).

11. The International Corpus of English. [Online]. Available from: http://icecorpora.net. (Accessed: 05.05.2017).

12. The Slovak National Corpus. [Online]. Available from: http://korpus.juls.savba.ski. (Accessed: 05.05.2017). (In Slovak).

13. The Czech National Corpus. [Online]. Available from: http://ucnk.ff.cuni.cz. (Accessed: 05.05.2017). (In Czech).

14. St. Petersburg Corpora of hagiographic texts of the XV–XVII centuries. [Online]. Available from: http://project.phil.pu.ru/skat. (Accessed: 05.05.2017).

15. Polinskaya, M., Rakhilina, E.V. & Vyrenkova, A.S. (2014) Grammatika oshibok i grammatika konstruktsiy: "eritazhnyy" ("unasledovannyy") russkiy yazyk [Grammar of errors and grammar of constructions: "erytazhny" ("inherited") Russian language]. *Voprosy yazykoznaniya*. 3. pp. 3–19.

16. The Russian Learner Corpus (RLC). [Online]. Available from: http://web-corpora.net/RussianLearnerCorpus/search/. http://web-corpora.net/RLC. (Accessed: 05.05.2017).

17. Petrova, K. (n.d.) *Proekt o sozdanii korpusa ustnoy rechi russko-bolgarskikh bilingvov* [Project on the creation of the corpus of oral speech of Russian-Bulgarian bilinguals]. [Online]. Available from: http://www.dialog-21.ru/media/2727/petrova.pdf. (Accessed: 05.05.2017).

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

Corpus of the Russian language: basic principles and basic functions]. In: *Natsional'nyy korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [The National Corpus of the Russian language: 2003–2005. Results and prospects]. Moscow: Indrik.

19. Iling-ran.ru. (n.d.) *Sotsiolingvisticheskaya anketa* [Sociolinguistic questionnaire]. [Online]. Available from: http://iling-ran.ru/main/departments/ural-altaic. (Accessed: 05.05.2017).

20. Marian, V., Blumenfeld, H.K. & Kaushanskaya, M. (2007) Language Experience and Proficiency Questionnaire (LEAP-Q). *Speech Language and Hearing Research.* 50 (4). pp. 940–967. [Online]. Available from: http://www.bilingualism.northwestern.edu/leapq/. (Accessed: 05.05.2017).

21. Luriya, A.R. (1998) *Yazyk i soznanie* [Language and consciousness]. Rostov-on-Don: Feniks.

22. Kibrik, A.A. (2003) *Analiz diskursa v kognitivnoy perspektive* [Analysis of discourse in the cognitive perspective]. Philology Dr. Diss. Moscow.

23. Foucault, M. (1996) *Volya k istine: po tu storonu znaniya, vlasti i seksual'nosti. Raboty raznykh let* [The will to truth: beyond knowledge, power and sexuality. Works of different years]. Translated from French. Moscow: Kastal'.

24. Makarov, M.L. (2003) *Osnovy teorii diskursa* [Fundamentals of discourse theory]. Moscow: Gnozis.

25. Karasik, V.I. (200) *Yazykovaya lichnost': institutsional'nyy i personal'nyy diskurs* [Language personality: institutional and personal discourse]. Volgograd: Peremena. pp. 5–20.

26. Shmeleva, T.V. (1997) Model' rechevogo zhanra [Model of the speech genre]. In: Gol'din, V.E. (ed.) *Zhanry rechi* [Genres of speech]. Vol. 1. Saratov: Kolledzh.