

преобразился, и, заулыбавшись, сказал: – Мои дела как сажа бела' [XΘ: 45].

Таким образом, в ходе анализа материала было выяснено, что в произведениях Л. Инджиева в основном распространены двухкомпонентные сравнения, наиболее частотны сравнения, построенные по модели существительное + послелог. Среди сравнительных послелогов выделяется послелог *мет*. Семантические группы сравнений носят антропоцентрический характер (человек сравнивает те или иные действия, предметы, признаки с тем, что его окружает).

**Литература и источники**

Алимарова Д. М. Сравнения и метафоры в тургеневских текстах // Проблемы общего и дагестанского языкознания. Вып. 6. Махачкала, 2010. С. 64–66.  
 Биткеев Н. Ц. Сравнения и метафоры в «Джангаре» Ээлян Овла и его последователей: к проблеме эпической традиции // Вестник КНИЯЛИ. Вып. 14. Элиста, 1976. С. 132–140.  
 Борджанова Т. Г. [Басангова Т. Г.] О поэтике ойратского героического эпоса: сравнение, гипербола, элементы фантастики // Вестник КНИЯЛИ. Вып. 14. Элиста, 1976. С. 141–162.  
 Бурыкин А. А., Басангова Т. Г. [Борджанова Т. Г.] Устойчивые сравнения в «Джангаре» и жанр примет-шинж (К проблеме интертекстуальности героического эпоса и генезиса жанра примет-шинж в калмыцком фольклоре) // Ойраты и

калмыки в истории России, Монголии и Китая: мат-лы Междунар. науч. конф. (Элиста, 9–14 мая 2007 г.). В 3-х ч. Ч. 2. Элиста: КИГИ РАН, 2008. С. 5–10.  
 Джимгиров М. Э. Писатели Советской Калмыкии (библиографический справочник). Элиста: Калм. кн. изд-во, 1966. С. 93–100.  
 Монраев М. У. Послелого // Грамматика калмыцкого языка. Фонетика и морфология. Элиста: Калм. кн. изд-во, 1983. С. 276–283.  
 Монраев М. У. О некоторых языковых особенностях романа А. Балакаева «Золотая Бумба» («Алтн Бумб») // Современный литературный процесс и литературная критика в Калмыкии. Элиста: Калм. кн. изд-во, 1977. С. 147–153.  
 Очирова Н. Ч. Лексико-семантическая характеристика сравнительных конструкций в произведениях К. Эрендженова // Молодежь в науке: проблемы, поиски, перспективы. Мат-лы II-й Межрегион. науч. конф. Вып. II. Элиста, 2005. С. 237–240.  
 Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. Пособие для учителей. Изд. 2-ое, испр. и доп. М.: Просвещение, 1976. 543 с.  
 Словарь литературоведческих терминов / ред.-сост.: Л. И. Тимофеев и С. В. Тураев. М.: Просвещение, 1974. 509 с.  
 Инжсин Л. Большевикүд. Түүк. Элст: Хальмг дегтр харһач, 1980. 204 х.  
 Инжсин Л. Харалта өдрмүд. Документальн түүк // Мартгдшго нерд. Түүк болн келвүрмүд. Элст: Хальмг дегтр харһач, 1990. 192 х.

УДК 81'33  
 ББК 81.2 3

**АРХИТЕКТУРА МЕТАОПИСАНИЯ  
 В НАЦИОНАЛЬНОМ КОРПУСЕ КАЛМЫЦКОГО ЯЗЫКА**

*В. В. Куканова*

Сегодня корпусная лингвистика является одним из популярных направлений в области языкознания. Это не особая лингвистическая дисциплина, как, например, социолингвистика или психолингвистика, хотя некоторые языковеды рассматривают корпусную лингвистику именно как самостоятельный раздел языкознания (см., например: [Корпусная лингвистика 2011; Захаров 2005: 3]). Вместе с тем зарубежные исследователи, занимающиеся проблемами создания и разработки корпусов, придерживаются несколько иного мнения, но при этом признают тот факт, что это уже более чем просто способ сбора материала (ср.: «it becomes quite evident that corpus linguistics is more a way of doing linguistics, „a methodological basis for pursuing linguistic

research“, than a separate paradigm within linguistics» [Meyer 2004: xi]). Таким образом, корпус — это способ работы лингвиста с текстами, более удобный, более быстрый, нежели карточки, которые использовались исследователями языка совсем еще недавно (ср.: «понятие корпуса является продолжением традиционных карточек, с которыми работали лингвисты. В XX в. эти карточки стали компьютерными и общедоступными» [Герд, Захаров 2004: 124]).

Корпусная лингвистика в основном сфокусирована на создании и разработке общих принципов построения и использования лингвистических корпусов для европейских языков, о чем свидетельствует существенное количество не просто коллекций текстов, а уже аннотированных текстовых мас-

символ<sup>1</sup>. К сожалению, проекты по созданию электронной базы данных по малым языкам, находящимся на грани исчезновения, вообще отсутствуют или находятся только в самом начале своей реализации, когда решаются общие вопросы корпусной лингвистики применительно к конкретному языку. Отсутствует и отработанная схема действий, не говоря уже о необходимых компьютерных программах (лемматизаторах, сегментаторах, тэггерах, парсерах). Предпринимаются попытки создания корпуса дагестанских языков [Кибрик и др. 2007; Мутулов 2007; 2009] и бурятского языка, структурно и типологически близкого калмыцкому языку [Бадмаева 2008; Бадмаева, Бадагаров 2008; Бадмаева и др. 2008; Ринчинов 2009].

К миноритарным языкам относится и калмыцкий язык монгольской группы алтайской семьи языков. Название проекта «Национальный корпус калмыцкого языка» (НККЯ) аналогично названию корпуса русского языка и призвано подчеркнуть преемственность традиций в общей и конкретно русской и калмыцкой лингвистиках. Обоснованию необходимости создания подобного корпуса по калмыцкому языку посвящена работа Э. У. Омакаевой и В. В. Кукановой [2011]. Здесь авторы рассматривают проблемы, связанные с составом текстов и их репрезентативностью, которые планируются представить в НККЯ, а также перспективы использования последнего не только в калмыцком языкознании, но и в алтаистике.

В настоящей статье обсуждаются проблемы метаописания текстового массива, работы по сбору которого активно ведутся в Калмыцком институте гуманитарных исследований РАН. Не вызывает сомнения и тот факт, что тексты на любом языке должны быть аннотированы по определенной системе, которую необходимо выработать.

Метаразметка, или метаописание (метаданные), — это система помет экстралингвистического характера, «структурированные данные», касающиеся формального сегментирования и «внешнего» аннотирования текста, а также фиксации технической

и технологической работы с электронным файлом. Следовательно, выделяют несколько видов метаописания:

1) внешняя, или «интеллектуальная», разметка:

- библиографические характеристики;
- типологические характеристики;
- тематические характеристики;
- социологические характеристики;

2) формальная, или структурная, разметка: поскольку текст состоит из частей разного уровня, то его делят на разделы, главы, части, абзацы, предложения;

3) технико-технологическая разметка: кодировка, даты работы, исполнители, источник электронной версии [Захаров 2011].

Подобная система помет может включать несколько критериев-признаков. Это, прежде всего, хронологический, жанровый, стилевые аспекты. Разметка помогает исследователю-лингвисту ориентироваться во множестве текстов, имеющих разные экстралингвистические и лингвистические характеристики. Весь массив текстов должен быть разбит на некоторые подмножества по различным признакам-критериям для того, чтобы облегчить поиск необходимого для анализа материала.

Следующей целью описания метаданных является выявление взаимосвязи языковой системы и условий его функционирования [Захаров 2011]. Как известно, экстралингвистические факторы напрямую влияют на язык. Существует настоятельная необходимость создания унифицированной структуры метаописания для НККЯ.

НККЯ состоит из двух модулей: текстов в электронном формате и базы данных, в которой дана метаразметка материала. База данных метаописания была сконструирована в программе *Systemic Coder*. Сначала эта программа была создана для анализа электронного дискурса, потом впоследствии, когда исследователи столкнулись с проблемой описания экстралингвистических параметров текста, функции программы были расширены. Постоянно выходят ее обновления, целью которых является улучшение программного продукта и упрощение работы лингвистов.

Набор параметров, обозначенных в узлах классификации, соответствует практике кодирования текстов таких известных корпусов, как Брауновский корпус, Британский национальный корпус. Все эти параметры, с другой стороны, следуют рекомендациям

<sup>1</sup> См., например: Брауновский Корпус [BC]; Хельсинкский аннотированный корпус [ХАНКО]; Национальный корпус русского языка [НКРЯ]; Банк английского языка [BEL]; Британский национальный корпус [BNC]; Корпус современного китайского языка [LIVAC]; Корпус современного итальянского языка [CORIS/CODIS]; Мангеймский корпус немецкого языка [CCDB] и др.

TEI [Sperberg-McQueen, Burnard 2001] и EAGLES [Sinclair 1996]. Однако приходится признавать, что классификация Дж. Синклера не может учитывать национальную специфику функционирования калмыцких текстов, калмыцкой художественной литературы. Дж. Синклер выделяет два класса факторов, которые, предположительно, могут оказывать влияние на свойства текстов: внешние (E), внеязыковые факторы, которые могут обусловить структуру построения или содержание текста, и внутренние (I), отражающие свойства языка, используемого в тексте:

- E1 (origin) – факторы, относящиеся к созданию текста автором;
- E2 (state) – факторы, относящиеся к внешним признакам текста (включая устную или письменную речь);
- E3 (aims) – факторы, относящиеся к причинам создания текста и его влиянию на аудиторию.
- I1 (topic) – предметная область текста;
- I2 (style) — стилистические особенности (частично пересекающиеся с E-факторами).

В ходе анализа научных трудов, посвященных возможным решениям метаописания текстового массива в различных корпусах, было выявлено, что в практике корпусной лингвистики существуют два вида метаданных: общее метаописание корпуса [Волков и др. 2004; Гарабик 2004; Герд, Захаров 2004; Плунгян, Сичинава 2004; Шаров, Савчук 2004; Савчук 2005] и специализированное метаописание, направленное на структуризацию метаданных в каждом конкретном подкорпусе, который создавался под определенные исследовательские задачи. Так, например, метатарметка в поэтическом корпусе [Гришина и др. 2009] или метаописание древнерусских агиографических текстов [Алексеева и др. 2004].

Ниже дана попытка свести все стандарты и классификации текстов в один формат с учетом специфики калмыцких текстов и их функционирования с опорой на метаописание, предложенное С. А. Шаровым [2011]. При необходимости вводятся новые тэги. Метаинформация, как правило, хранится в отдельном файле для того, чтобы было легче ориентироваться во множестве текстов.

**Заголовок** <teiHeader> имеет следующие элементы:

1) <id> — уникальный номер, позволяющий идентифицировать документ в корпусе

(как правило, это название текста);

2) <target> — имя файла, в котором находится документ;

3) <type> — тип описания. В нашем случае можно описывать не только один текст (“text”), а группу текстов (“texts”). Например, пословицы и поговорки, являющиеся небольшими по объему текстами, как правило, состоящими из одного или двух предложений;

4) <lang> — язык, на котором написан документ, в нашем случае “xal”, в TEI используется указание языка по стандарту ISO 260.

**Библиографические данные о тексте** документа <fileDesc> состоят из следующих помет.

1. <titleStmt> — библиографическая информация о тексте:

- <title> — название текста (возможно отдельно полное и краткое);
- <author> — автор.

В описании этого параметра имеется несколько атрибутов: type=“sole” (стоит по умолчанию); type=“corporate”; type=“multi-authored”. При обозначении первого типа может указываться и настоящая фамилия автора (если она известна), и его псевдоним, поскольку именно под этим ним этот текст известен широкому кругу читателей. Если текст был создан группой авторов, как, например, инструкции или конституции, то используется в описании второй тип type=“corporate”. Если же автором текста является весь народ, как в случае с фольклорными произведениями, то приписывается последняя помета type=“multi-authored”. Во втором и третьем случае авторство установить невозможно;

• <extent> — размер документа в некоторых условных единицах.

По предложению С. А. Шарова, размер документа может фиксироваться в словах (type=“w”), в предложениях (type=“s”) и условных единицах (type=“u”). Трудности возникнут с текстами, написанными на тодо бичг, поскольку часто в них не ставились знаки, обозначающие, например, конец предложения;

• <respStmt> — информация о людях, внесших вклад в создание электронной версии текста (фамилии исследователей-лингвистов). Данная информация предназначена для внутренней работы, т. е. закрыта для широкого круга пользователей корпуса;

2. <publicationStmt> — библиографическая информация об издании;

- <publisher> — информация об издательстве. Если текст, например стихотворение или рассказ, берется из книги, то приводятся выходные данные с указанием страниц, где опубликован этот текст;

- <pubPlace> — место издания;

- <publDate> — год издания оригинального документа. Указывается точная или приближительная дата появления того или иного текста. Существует ряд текстов, не имеющих точной даты появления. Это в основном касается только ранних текстов, написанных на тодо бичг;

- <graphicSystem> — графическая система, с использованием которой текст был впервые напечатан или написан (тодо бичг, латиница, кириллица). Все тексты для удобства анализа переводятся в кириллицу с сохранением всех особенностей;

- <editor> — редакция текста. Указывается наличие или отсутствие редактуры текста. Если текст был отредактирован одним человеком, указывается помета “sole”, если же принял участие коллектив редакторов, то приписывается параметр “group”;

- <translator> — переводчик текста, что важно для выявления индивидуальных стратегий перевода текста. Если текст был переведен одним человеком, указывается помета “sole”, если же принял участие коллектив переводчиков, то приписывается параметр “group”.

3. <sourceDesc> — информация об источнике, из которого получена электронная версия документа: type=“internet”; type=“scanning”; type=“handTyping”; type=“authorialCopy”. Всего лишь несколько текстов было взято с различных сайтов, но в основном тексты на калмыцком языке сканируются и распознаются, так как электронные версии практически отсутствуют.

**Метаинформация о профиле документа** <profileDesc> включает следующие параметры:

1. <creation> — информация о времени и месте создания текста.

Этот параметр отличается от времени и места публикации, хотя он в некоторых случаях может совпадать (например, у газетных статей). Иногда невозможно указать точную дату и точное место создания текста. Например, историческая песня «Сэм хамрта парнцс» появилась приблизительно в 1810-е гг. во время или после войны с На-

полеоном. Следовательно, указывается период 1810-е гг., поскольку точно датировать появление текста песни невозможно. Данный параметр состоит из следующих тэгов:

- <creationDate> — информация о времени создания текста;

- <creationPlace> — информация о месте создания текста;

1. <textClass> — классификация текста по определенным признакам.

В этом параметре возникают проблемы теоретического характера, которые в принципе характерны для всех типов текстов на любом языке. Существует конечное множество текстов, хотя и «с открытым концом», и описать их все по одной и той же структуре практически невозможно. Многие из них могут находиться на границе двух или трех, допустим, предметных областей.

С. А. Шаров предлагает классифицировать тексты по следующим параметрам: предметная область в зависимости от науки и стиля, признавая, что «полноценной классификации, покрывающей все [курсив автора — С. А.] частотные виды использования русской речи, также не существует». Если здесь, в русском языковедении, налицо проблема охвата текстов существующими классификациями, то на калмыцком языке имеет место другая проблема: количество текстов на калмыцком языке совсем незначительно, в особенности, если сравнивать с числом текстов, созданных на английском или русском языке. Возникают опасения в возможности сбалансированности и репрезентативности объема текстового массива.

Тем не менее формально этот параметр должен присутствовать в метаописании, поскольку только тогда мы можем понять, в какой языковой области имеются лакуны. Следовательно, в таких предметных областях существует проблема неразработанности терминологической базы. По предложенной С. А. Шаровым схеме (по его замечанию поверхностной) тексты следует классифицировать по предметной области <textClassContent> следующим образом:

**natsci (Естественные науки)**

mathematics (Математика)

biology (Биология)

physics (Физика, включая астрономию, оптику и т. п.)

chemistry (Химия)

geo (География, геология, метеорология и т. д.)

...

**appsci (Прикладные науки)**

agriculture (Сельское хозяйство)  
 medicine (Медицина, включая ветеринарию, питание и т. п.)  
 ecology (Экология, окружающая среда)  
 engineering (Техника и технология)  
 computing (Вычислительная техника)  
 military (Военное дело)  
 transport (Транспорт, мореплавание, авиация и т. п.)  
 ...

**socsci (Общественные науки)**

law (Юридическая тематика)  
 history (История, включая археологию)  
 philosophy (Философия)  
 psychology (Психология)  
 sociology (Социология)  
 anthropology (Антропология)  
 language (Лингвистика, филология)  
 education (Образование)

**religion (Религия)****politics (Политика)**

inner (Внутренняя)  
 world (Внешняя)

**commerce (Экономика)**

finance (Финансы)  
 industry (Промышленность)

**life (Общество)****arts (Искусство)**

drawing (Изобразительное искусство, включая скульптуру)  
 literature (Литература)  
 architecture (Архитектура)  
 performing (Театр, кино, танец)

**leisure (Досуг)**

reading (Чтение)  
 sports (Спорт)  
 travels (Путешествия)  
 fashion (Мода/одежда)

Оговоримся, что этот параметр не отражает полностью содержания текста, лишь дает небольшое приближение к его тематике. Но в целом, если будут охарактеризованы все позиции по схеме, то мы получим более или менее полную картину о тексте. Например, историческое предание о Мазан-Батыре можно отнести к истории (history) в общественных науках (socsci), хотя, как известно, все же это далеко не история. В таком случае другой параметр классификации (признак стиля) отделит этот текст от текстов, написанных историками.

Следующий параметр <textClassStyle> носит также дискуссионный характер, но выделить его необходимо, поскольку от-

ражает совокупность подмножеств текстов по определенному критерию. Ученые предлагают характеристику текстов по нескольким критериям-признакам (неформальный / нейтральный / формальный / академический стиль или почтительно / просторечно / лично / небрежно), однако мы остановимся на двух традиционных классификациях: формальности/неформальности речи (type="formal"; type="informal") и пяти функциональных стилей (научный, публицистический, деловой, художественный и разговорный стили).

Третьей характеристикой текста является его жанровая принадлежность. В зависимости от стиля текста выделяют жанры:

- type="NStyle" — научный стиль: монография, диссертация, статья, тезисы, доклад, рецензия, ответ на рецензию;
- type="PStyle" — публицистический стиль: статья в газете, очерк, интервью, фельетон и др.;
- type="RStyle" — разговорный стиль: монолог/диалог; спонтанная, квазиспонтанная (интервью, рассказ на заранее подготовленную тему, пересказ, описания и т. д.), подготовленная (чтение, публичные выступления, пересказ, чтение стихотворений и т. д.);
- type="DStyle" — деловой стиль: протокол, акт, справка, заявление и др.;
- type="ChStyle" — художественный: роман, рассказ, повесть, лирическое стихотворение и др.; среди фольклорных произведений — эпос, сказка, предание, легенда, миф, песня, малые жанры и т. д.

Все многообразие жанров калмыцкой литературы будет отражено в планируемом корпусе текстов, однако уже и здесь существует проблема сбалансированности материала, думается, что часть стилей и жанров будет представлена в меньшей степени в базе данных, следовательно, существует проблема диспропорции текстового массива.

2. <textDesc> — описание текста в терминах условий его создания (situational parameters), которые могут включать в себя:

- <formSpeech> — форма речи: устная type="oral" и письменная type="writting". Следует выделить также электронную коммуникацию, которая совмещает в себе признаки устной и письменной форм речи (type="electComm"), хотя в Интернете (форумы, сайты, смс-сообщения, электронные письма и т. д.) обнаружено не очень много текстов;

- <kindSpeech> — монолог / диалог / полилог. Данный критерий также релевантен для описания языковых особенностей, проявляющихся в том или ином тексте.
- <preparedness> — признак спонтанности. Вслед за И. Н. Борисовой [2005] и Е. А. Гришиной [2005], мы выделяем три степени спонтанности текстов, при этом данная характеристика может быть релевантным критерием как и для письменной, так и для устной речи: 1) спонтанная, или неподготовленная, речь; 2) частично подготовленная, или квазиспонтанная, речь; 3) подготовленная речь;
- отношение между автором (или говорящими для устной речи) и аудиторией (<interaction>).

**Характеристика автора текста/редактора/составителя** — один из важных критериев, оказывающих значительное влияние на порождение текста и его функционирование: «данные параметры могут быть потенциально релевантны не только для автора текста, но и для редактора, цензора, переводчика и интерпретатора оригинального текста» [Шаров 2011]. Следующие параметры входят в элемент <person>:

- <role> — роль в коммуникации (для письменного текста role="author");
- <sex> — пол: значения "m/f";
- <age> — возраст в момент создания текста. Здесь вслед за С. А. Шаровым указывается примерный возраст: "child" (ребенок), teen (подросток), "mid-aged" (средних лет), "senior" (старше 55 лет). Сразу оговоримся, что в нашей коллекции текстов авторы в основном взрослые люди средних лет. Тексты, созданные детьми, практически отсутствуют на момент написания статьи;
- <birth> — релевантные параметры рождения (дата и место);
- <firstLang> — родной язык. Этот признак — один из самых важных, поскольку осуществляющих всю коммуникации на калмыцком языке уже единицы, причем это люди пожилого возраста;
- <dialect> — описывает диалектную принадлежность автора. В калмыцком языке существует три диалекта (дербетский, бузавский, торгутский) и в соответствии с ними выделяются три пометы: type="torgut"; type="buzav"; type="derbet". Этот признак является релевантным для социологической характеристики авто-

ра текста, поскольку литературная форма языка сформировалась относительно недавно по времени;

- <langKnown> — знание других языков;
- <residence> — место жительства в момент создания текста;
- <education> — уровень образования;
- <occupation> — род занятий;
- <socecStatus> — социально-экономический статус;
- <circumstances> — дополнительное описание обстоятельств, в которых был создан текст.

Предложенная в статье система метаописания необходима, во-первых, для того, чтобы систематизировать материал, который имеется в нашем распоряжении. Во-вторых, подобная разметка текстов поможет ориентироваться лингвисту во множестве текстов, а также выявлять коррелятивную зависимость функционирования языка от экстралингвистических факторов. Характеристика как текстов, так и авторов является важной задачей в создании корпуса текстов на любом языке. В ходе работы над разметкой НККЯ некоторые ее принципы и структура будут уточняться исследователями-лингвистами в соответствии основными нормативными требованиями в описании метаданных корпуса, как, например, требованием унификации описания параметров текста.

#### Литература

- Алексеева Е. Л., Лаврентьев А. М., Азарова И. В., Захарова Л. А. Разметка корпуса древнерусских агнографических текстов // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 16–23.
- Бадмаева Л. Д. Корпус бурятского языка: проект [Электронный ресурс] // URL: <http://www.globecsi.ru/Articles/2008/Badmaeva3.pdf> (20.04.2011).
- Бадмаева Л. Д. Бурятский язык и корпусная лингвистика // Состояние и перспективы развития бурятского языка: мат-лы форума бурят. яз. Улан-Удэ, 2009. С. 83–86.
- Бадмаева Л. Д., Бадагаров Ж. Б. О репрезентативности текстов и элементах программного инструментария для корпуса бурятского языка // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Е1 Manuscript-08. Мат-лы Междунар. науч. конф. (Казань, 26–30 августа 2008 г.). Казань: Изд-во Казан. гос. ун-та, 2008. С. 28–31.
- Бадмаева Л. Д., Бадагаров Ж. Б., Цыдытов Б. З. Общие проблемы формирования корпуса бурятского языка // Труды Международной конференции «Корпусная лингвистика – 2008». СПб., 2008. С. 24–30.

- BEL* — Банк английского языка [Электронный ресурс] // URL: <http://www.collins.co.uk/Corpus/CorpusSearch.aspx> (дата обращения: 15.04.2011).
- Борисова И. Н.* Русский разговорный диалог: структура и динамика. Екатеринбург, 2005. С. 132–144.
- BC* — Брауновский Корпус [Электронный ресурс] // <http://www.hd.uib.no/icame/brown/bcm.html> (дата обращения: 15.04.2011)
- BNC* — Британский национальный корпус [Электронный ресурс] // URL: <http://saga.natcorp.ox.ac.uk/> (дата обращения: 15.04.2011).
- Волков С. С., Захаров В. П., Дмитриева Е. А.* Мета-разметка в историческом корпусе XIX века // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 86–98.
- Гарабик Р.* Словацкий национальный корпус // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 99–121.
- Герд А. С., Захаров В. П.* Национальный корпус русского языка в свете проблем современной филологии // Труды международной конференции «Корпусная лингвистика–2004» (11–14 октября 2004 г., г. Санкт-Петербург). СПб.: Изд-во СПбГУ, 2004. С. 122–130.
- Герд А. С., Захаров В. П.* Национальный корпус русского языка в свете проблем современной филологии // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 122–130.
- Гришина Е. А.* Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005. С. 94–110.
- Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В.* Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 71–113.
- Захаров В. П.* Корпусная лингвистика: Учебно-метод. пособие. СПб., 2005. 48 с.
- Захаров В. П.* Экстралингвистическая разметка. Метаданные [Электронный ресурс] // URL: [http://download.yandex.ru/class/zakharov/CL\\_L4.ppt](http://download.yandex.ru/class/zakharov/CL_L4.ppt) (дата обращения 26.03.2011).
- Кибрик А. Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Т., Нахимовский А. Д.* Технологии обработки языковых данных в документировании малых языков // Материалы Международной конференции «ДИАЛОГ 2007» «Компьютерная лингвистика и интеллектуальные технологии». М., 2007; URL: <http://www.dialog-21.ru/dialog2007/materials/html/35.htm> (дата обращения: 23.04.2011).
- CORIS/CODIS* — Корпус современного итальянского языка [Электронный ресурс] // URL: <http://www.cilta.unibo.it/ricerca.htm> (дата обращения: 15.04.2011).
- LIVAC* — Корпус современного китайского языка (LIVAC Synchronous Corpus) [Электронный ресурс] // URL: <http://www.cilta.unibo.it/ricerca.htm> (дата обращения: 15.04.2011).
- Корпусная лингвистика* [Электронный ресурс] // URL: [http://ru.wikipedia.org/wiki/Корпусная\\_лингвистика](http://ru.wikipedia.org/wiki/Корпусная_лингвистика) (дата обращения: 15.03.2011).
- CCDB* — Мангеймский корпус немецкого языка (Institut für Deutsche Sprache, Mannheim, Germany) [Электронный ресурс] // URL: <http://corpora.ids-mannheim.de/~cosmas/> (дата обращения: 15.04.2011).
- Муталов Р. О.* Корпусная лингвистика и перспективы ее развития в Дагестане // Махачкала: Современные проблемы кавказского языкознания, 2007. Вып. 7. С. 160–173.
- Муталов Р. О.* Опыт создания корпусов дагестанских языков [Электронный ресурс] // <http://www.dialog-21.ru/dialog2009/materials/html/50.htm> (дата обращения: 23.04.2011).
- НКРЯ* — Национальный корпус русского языка [Электронный ресурс] // URL: <http://ruscorpora.ru/> (дата обращения: 15.04.2011).
- Омакаева Э. У., Куканова В. В.* Общая структура и перспективы использования Национального корпуса калмыцкого языка в свете проблемы репрезентативности // Монголоведение: сб. науч. тр. Вып. 5 / отв. ред. Н. Г. Очирова. Элиста: КИГИ РАН, 2011. С. 228–243.
- Плунгян В. А., Сичинава Д. В.* Национальный корпус русского языка: опыт создания корпуса текстов современного русского языка // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 216–238.
- Ринчинов О. С.* Корпус бурятского языка и прикладные задачи компьютерной лингвистики // Состояние и перспективы развития бурятского языка. Материалы форума бурятского языка. Улан-Удэ, 2009. С. 88–89.
- 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.
- ХАНКО* — Хельсинкский аннотированный корпус [Электронный ресурс] // URL: <http://www.ling.helsinki.fi/projects/hanko/> (дата обращения: 15.04.2011).
- Шаров С. А.* Параметры описания текстов корпуса [Электронный ресурс] // URL: <http://bokrcorpora.narod.ru/header.html> (25.05.2011).
- Шаров С. А., Савчук С. О.* Типология текстов для представительного корпуса // Труды международной конференции «Корпусная лингвистика – 2004» (г. Санкт-Петербург, 11–14 окт. 2004 г.). СПб.: Изд-во С.-Петербург. ун-та, 2004. С. 352–362.
- Meyer Ch. F.* English Corpus Linguistic. An introduction. Cambridge, 2004. 169 p.
- Sinclair J.* Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P [Электронный ресурс] // <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>. 1996 (дата обращения: 15.04.2011).
- Sperberg-McQueen C. M., Burnard L.* Guidelines for Electronic Text Encoding and Interchang [Электронный ресурс] // <http://www.hcu.ox.ac.uk/TEI/P4X/index.html>. 2001 (дата обращения: 15.04.2011).