УДК 81’33
ББК 81.23

В.В. Куканова, Е.В. Бембеев,
Н.Н. Убушаев, Б.Б. Манджиева

УСТНЫЕ ТЕКСТЫ НА КАЛМЫЦКОМ ЯЗЫКЕ: ЗАПИСЬ И РАСШИФРОВКА

Статья подготовлена при поддержке проекта «Национальный корпус калмыцкого языка» подпрограммы фундаментальных исследований Президиума РАН «Создание и развитие корпусных ресурсов по языкам народов России» программы «Корпусная лингвистика» (2012-2014) и проекта РГНФ «Национальный корпус калмыцкого языка: создание и разработка» (12-04-12047/в).

***Аннотация**. В статье посвящена описанию работы по записи и расшифровке устных текстов на калмыцком языке как составной части Национального корпуса калмыцкого языка. Подкорпус устных текстов состоит из трех модулей: звуковых файлов, базы данных и расшифровок.*

***Ключевые слова**: корпусная лингвистика, Национальный корпус калмыцкого языка, база данных, устная речь, запись, расшифровка.*

V.V. Kukanova, E.V. Bembeev,
N.N. Ubushaev, B.B. Mandzhieva

ORAL TEXTS IN THE KALMYK LANGUAGE: RECORDING AND DECODING

Article is prepared with support of the “National Case of the Kalmyk Language” draft of the subprogramme of basic researches of Presidium of the Russian Academy of Sciences “Creation and development of case resources on languages of the people of Russia” the Case Linguistics programs (2012–2014) and the RGNF project “The national case of the Kalmyk language: creation and development” (12-04-12047/in).

***Summary**. The article is devoted to the description of the recording and decoding of oral texts in the Kalmyk language as a part of the National corpora of the Kalmyk language. The subcorpus of oral texts is composed of three modules: sound files, databases and decoding files.*

***Key words**: corpora liguistics, National corpora of Kalmyk Language, database, recording, decoding*

На современном этапе языковая ситуация в Калмыкии сложилась весьма печально: язык титульной нации республики находится в условиях постепенной утраты, неслучайно, что ЮНЕСКО включил его в список определенно исчезающих языков [10]. Калмыцкий язык уступает свои позиции доминирующему языку во всех сферах, начиная с официальной и заканчивая бытовой, глобализация усиливают эти процессы. Диалекты калмыцкого языка находятся еще в большой опасности по сравнению с литературным языком, который более или менее удерживает свои позиции, поскольку территориальные разновидности языка испытывают давление как со стороны кодифицированного языка, так и со стороны русского языка. Остались ли сейчас какие-то диалектные особенности – трудно сказать, они, быть может, уже в большей степени не носят системного характера или совсем утрачены.

В Калмыцком институте гуманитарных исследований Российской академии наук создается корпус калмыцкого языка на основе письменных источников, преимущественно художественных произведений и газетных текстов, которых, надо сказать, не так много. В рамках данного проекта мы пытаемся записать живую речь калмыков, преимущественно старшего поколения, поскольку представители этой возрастной группы владеют языком свободно. Такое расширение задач проекта позволяет сфокусировать свои исследования не только на письменных источниках, в которых в основном не проявляются диалектные черты автора текста в силу разных причин [см. подробнее: 3].

Живая речь носителей языка может сильно отличаться от письменной, и здесь речь идет не сколько о характерных особенностях устной речи (повторы, паузы хезитации, обрывы и др.), а столько о том, что в их языке могут иметь место различные элементы диалектной системы, которые описаны или не описаны в калмыцкой диалектологии. Если на уровне письменных текстов уже разработана система помет того или иного диалекта, которая, по сути, ничего не дает исследователю-диалектологу и является скорее дескриптивной, то на материале текстов, записанных от носителей калмыцкого языка мы получим новый материал, в котором можно обнаружить нечто новое, являющееся проявлением функционирования реликтовых форм в языке.

Для реализации этой цели был проанализирован опыт составления диалектных корпусов в отечественной диалектологии [2; 4; 5; 7; 11 и др.]. В результате решено, что наилучшим способом презентации материала текстов является фонетическая транскрипция звучащей речи, наряду с орфографической записью. Именно она позволит сохранить все особенности диалекта, хотя это, конечно, не отменяет необходимости записывать нетекстовый материал на основе вопросников для сбора диалектологического материала.

Целью данной статьи является описание методики записи устного материала и его расшифровки, так как этот методический аспект считается наименее разработанной частью полевой лингвистики в калмыцком языкознании. Следует составить программу действий, учесть все до мельчайших деталей, что позволит избежать многих ошибок в проведении эксперимента по записи диалектного материала. Эта программа состоит из следующих структурных элементов[1]: 1) условия записи; 2) выбор информантов; 3) составление анкет, направленных на характеристик у информанта; 4) отбор тем для беседы; 5) лингвистические опросники; 6) ход записи диалектного материала; 6) расшифровка текстового материала.

1. Условия записи устной речи. Запись в основном должна проводиться в помещениях, где нет сильного зашумления помехами (например, голосами других людей, шума машин, работающих электроприборов и т. д.). Однако интересно записывать речь не только в помещениях, но и в потенциально шумных местах – в больницах, школах и т. д., – поскольку все они отражают разные коммуникативные ситуации, что необходимо для создания сбалансированной коллекции текстов различных речевых сценариев, или фреймов. Необходимо использовать диктофон с выносным микрофоном (лучше всего с встроенной системой подавления шумов с постоянной частотой). Оптимально диктофон должен быть расположен недалеко от говорящего. Признаком хорошей записи считается расшифровка с первого раза, без многократного прослушивания.

Если же запись направлена на сбор материала «одного речевого дня»[2], то диктофон находится у говорящего постоянно, но записывающийся носит его таким образом, чтобы он не создавал, с одной стороны, дополнительных трудностей самому говорящему,

как например, постоянное его перемещение, и, с другой стороны, чтобы получаемая запись была бы, как можно, чище в фонетическом плане, например, чтобы сам диктофон при ходьбе не перемещался и не бился, создавая тем самым звуковые помехи. Оптимально повесить его на шею.

2. Выбор информантов определяется задачами лингвистического исследования. В идеале целью нашей работы является собирание представительного и сбалансированного звукового материала по калмыцкому языку. Информанты в целом должны отражать генеральную совокупность, т. е. это люди разного возраста и пола, социального статуса, профессий, диалектной принадлежности. Однако в сложившейся языковой ситуации представители молодого и среднего поколения не владеют языком вообще или ограниченно.

Тем не менее есть четкое требование в процессе отбора информантов: все должны владеть достаточно свободно калмыцким языком и иметь высокую языковую компетенцию, и желательно, чтобы первичным языком был именно калмыцкий. Дальнейший анализ социальных групп проводится после накопления первичных записей.

3. Составление анкет. Каждый информант должен заполнить анкеты, необходимые для последующего лингвистического анализа записанных текстов.

Пример анкеты.

АНКЕТА ИНФОРМАНТА №____

1. Ваша фамилия, имя, отчество (по желанию): _____
_____
2. Пол: мужской, женский (нужное подчеркнуть).
3. Возраст (год рождения и год по калмыцкому календарю) _____
4. Национальность _____
5. Субэтническая группа: дербеты, торгуты, бузавы, хошуды (нужное подчеркнуть)
6. Этнические маркеры и род внутри группы *арван*, *ясн*, *төрл*, *өлгц*, *туг*, *уран*_____
7. Знаете ли Вы свою родословную? _____
8. Где Вы родились?_____
9. Где провели детство?_____
10. Где провели юность?_____
11. Место проживания в настоящее время (укажите район, село, город) _____
_____
12. Образование: среднее, среднее специальное, высшее (нужное подчеркнуть)
13. Где и кем работаете (работали)? _____
14. Район проживания до депортации_____
15. Район депортации (если были депортированы)_____
16. Район возращения после депортации_____
17. Участник ВОВ: да, нет (нужное подчеркнуть, укажите, в каких частях служили) _____
_____
18. Расскажите о супруге (место рождения, год рождения, год рождения по калмыцкому календарю, род, арван, ясн, төрл, өлгц, туг, уран)
19. Первичный язык (т. е. тот язык, который первым усвоили) _____
20. Какими еще языками владеете? _____
21. Время изучения других языков _____

22. В какой мере вы владеете калмыцким языком? (нужное подчеркнуть)

• Свободно разговариваю, читаю и пишу, думаю на родном языке;

• Неплохо разговариваю, часто думаю на родном языке, относительно хорошо читаю и пишу;

• Разговариваю и думаю на родном языке, но не умею читать и писать;

• Понимаю, но говорю с трудом;

• Говорю с трудом и плохо понимаю.

23. На каком языке вы говорите в семье, на работе, во время учебы, в общении с друзьями? (нужное подчеркните)

• только на родном языке (в семье, на работе, во время учебы, в общении с друзьями);

• в основном на родном языке (в семье, на работе, во время учебы, в общении с друзьями);

• в одинаковой степени на русском и калмыцком языках (в семье, на работе, во время учебы, в общении с друзьями);

• в основном на русском языке (в семье, на работе, во время учебы, в общении с друзьями);

• только на русском языке (в семье, на работе, во время учебы, в общении с друзьями).

24. Знаете ли Вы фольклорные произведения на калмыцком языке? Если да, то укажите, пожалуйста, жанры: эпос, мифы, предания, сказки, песни, пословицы, поговорки, загадки и др. _____

25. Знаете ли Вы калмыцкие обычаи и традиции? _____

26. Контактные данные_____

4. Отбор тем для беседы. Поскольку в Калмыкии калмыцкий язык как средство коммуникации и познания утрачивается, то создать сбалансированную и более или менее полную коллекцию речевых сценариев априори уже невозможно, но, тем не менее, мы выработали примерный круг тем для сбора диалектного и устного материала. В качестве исходной посылки выступает идея, которая заключается в том, что одна и та же тема для разговора будет содержать одинаковую лексику, которая будет появляться естественным образом у разных информантов, с одной стороны. С другой – в связном тексте «работает» морфология и синтаксис языка, не говоря уже о фонетике. Тексты, записанные от информанта, в большей степени отражают речь в естественных условиях.

Итак, структура может состоять из нескольких модулей:

а) блок монологической речи по темам[3]:

• о себе (автобиография);

• семья;

• национальный образ жизни (о праздниках, традициях и обычаях);

• предсвадебные подготовки и обычаи;

• свадьба;

• календарные праздники;

• рождение;

• наречение именем;

• первая стрижка;

• воспитание;

• диплом;

• похороны;

• заговоры;

• болезни;
• депортация;
• Великая Отечественная война;
• профессия;
• культура;
• политика;
• спорт;
• игры;
• фольклорные произведения[4];

б) блок диалогической речи (основные коммуникативные ситуации):
• знакомство;
• встреча;
• благодарность;
• прощание;
• выражение согласия/несогласия;
• разговор о погоде;
• поздравления и пожелания;
• в банке;
• в магазине;
• разговор по телефону;
• в кино;
• в больнице;
• в ресторане;
• городской транспорт;
• автомобиль;
• чрезвычайные ситуации;
• средства связи;
• хобби и увлечения;
• семья и родственники;
• карьера и профессии;
• романтическое свидание и отношения.

5. Лингвистические опросники. В рамках записи устной речи разных диалектных групп было решено собрать материал по словнику Сводеша, в который входят 207 лексических единиц и которые, по его мнению, присутствуют в любом языке. Действительно, эта базовая лексика, которая состоит из прилагательных, существительных, глаголов, местоимений и др. Информанту будет предложено перевести список слов с русского на калмыцкий. Такой опросник в большей степени направлен на изучение дифференциации на лексическом и фонетическом уровнях. Однако этот список для перевода нужно предоставлять не всем информантам. Это продиктовано тем, что не все информанты обладают дикторскими способностями, а это важно для изучения фонетических особенностей, а также после оценки исследователем языковой компетенции информанта.

6. Ход записи устной речи. Сбор материала проводится по программе, определенной заранее. Выявление диалектных особенностей производится в ходе направленной беседы. При подготовке к беседе обдумывается план разговора, формулировка вопросов. Задавать вопросы так, как они сформулированы в программе, нецелесообразно. Перед записью информант получает свой индивидуальный номер у ответственного за регистрацию, который фиксирует номер, дату записи, ответственного за запись в таблице.

До записи лингвист инструктирует информанта, который должен заполнить анкеты с помощью исследователя. Важно в процессе заполнения анкеты расположить информанта к беседе, заполнять анкету может и исследователь, и сам информант. Задавать вопросы в процессе заполнения анкеты нужно только на калмыцком языке, тем самым лингвист помогает быстрее переключится информанту с одного кода на другой.

Ответственный создает папку под названием Informant-00 (вместо 00 пишется номер информанта, который был ему присвоен при регистрации), обязательно используя латинские буквы. Внутри данной папки создается несколько папок. Исходные звуковые файлы копируются в созданную папку Iskhodnye с теми же названиями, которые были автоматически даны диктофоном или записывающей программой на компьютере. Если даже файлы «пустые», т. е. на них не записана речь информанта и его коммуникантов, то он ни в коем случае не удаляется. Обработанный материал помещается в папку Obrabotannye, а расшифровку звуковых файлов сохраняют в другой папке (Decoding).

Звуковой корпус по калмыцкому языку состоит из трех модулей: базы данных «KalmykSpeech», звуковых файлов и расшифровок в формате .eaf. Общая метаинформация, собранная в процессе заполнения анкеты, фиксируется в специально созданных таблицах в программе Access MS Office. Таблицы Informants, SoundFiles, Epizods связаны друг с другом при помощи общих полей (ниже приведена схема данных).

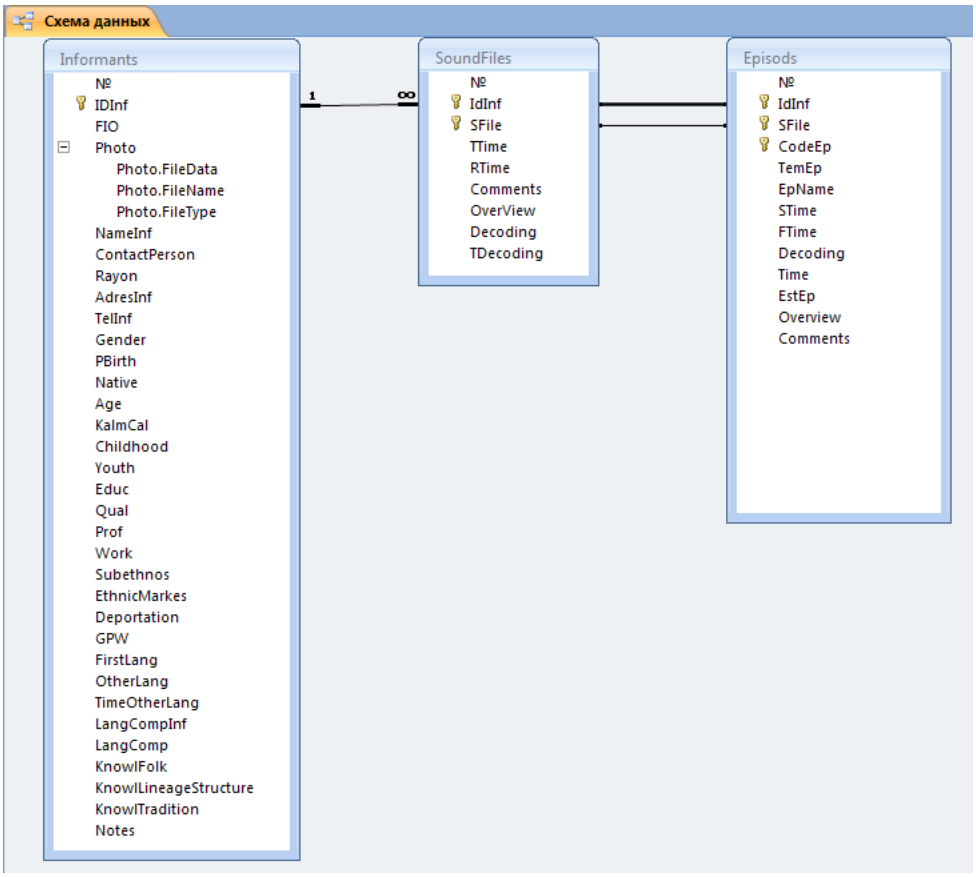

Рис. 1. *Схема данных базы «KalmykSpeech»*

Далее файл передается лингвисту, который предварительно проводит процедуру автоматической шумоочистки (программе после предоставления образца шума статистического характера удаляет данные частоты на протяжении всего звукового файла). Расшифровка (и скоростная, и детальная) проводится в специально разработанной для полевой работы программе Elan 4.0. Далее исследователь производит аннотирование звукового файла, которое на данном этапе состоит из следующих уровней:

1) уровень макроэпизодов;

2) уровень информанта и коммуниканта (коммуникантов);

3) уровень предложений в орфографической расшифровке;

4) уровень предложений в фонетической расшифровке.

Первоначально исследователь проводит скоростную расшифровку, указывая речевые эпизоды, под которыми понимаются макроэлементы, являющиеся законченными смысловыми отрезками речи. Чаще всего эпизоды соответствуют речевому сценарию. Например, разговор о погоде, рассказ о своем детстве и т. д. Индексом А0 обозначается отрезок звукового файла, который оценивается как «мусор». Например, включение диктофона, проверка и т. п.

Аннотирование на уровне предложений в орфографической записи происходит по следующим правилам:

1) знаком (/) обозначается конец синтагмы;

2) знаком (//) – конец высказывания повествовательного характера;

3) знаками (?) и (!) – конец высказывания вопросительного и восклицательного характеров;

4) знаком (…) – обрыв слова без пробела перед последующим словом;

5) знаком (…) – обрыв высказывания с пробелом перед последующим словом;

6) знаком (//-//) – длительная пауза хезитации;

7) (...) – незаполненная пауза хезитации (например, запинка, которая перцептивно ощущается);

8) (э-э, *а-а*, *м-м* и т.д.) – заполненная пауза хезитации;

9) @ … @ – речь коммуникантов или на фоне разговора информанта и коммуниканта;

10) # ... # – переключение языковых кодов (например, с калмыцкого на русский и наоборот);

11) *С, *К, *П и др. – паралингвистические элементы речи (например, смех);

12) *би-ичә* – протяжка того или иного звука в слове.

Аннотирование на уровне предложений в фонетической записи начинается с копирования уровня предложений в орфографической записи. Нами были рассмотрены несколько систем фонетической транскрипции, выработанных лингвистами для различных задач: Международный фонетический алфавит [6] и Уральская фонетическая система [9]. Было решено остановиться на системе Международного фонетического алфавита в несколько упрощенном и слегка модифицированном виде. Это позволит, с одной стороны, расширить круг исследователей, которые интересуются фонетикой калмыцкого языка, с другой стороны, унифицировать подачу произношения в нормативных словарях, что наиболее актуально для языковой ситуации в Калмыкии. В калмыцком письме не обозначаются так называемые «неясные» гласные, что уже привело к негативным последствиям – изменениям в орфоэпии, исчезновению языку, поскольку дети не могут прочитать скопления согласных, не зная, где произнести сверхкраткие гласные. Многие словари используют для передачи произношения слов систему Международного фонетического алфавита, например Oxford Advanced Learner’s Dictionary и Cambridge Advanced Learner’s Dictionary.

Были выработаны правила применительно к калмыцкому языку и отображают только те качества речи, которые являются дифференцированными (смыслоразличительными) в устной речи. Согласно МФА, речь транскрибируется на латинице с использованием нестандартных символов, поэтому в этих целях был написан специализированный макрос по замене орфографической записи калмыцкой речи кириллическими символами на латинские (см. таблицу символов). Это ускорит расшифровку на фонетическом уровне, хотя это, конечно, незначительно облегчит работу. Было принято решение использовать следующие символы для обозначения звуков калмыцкого языка:

- гласные звуки полного образования: a, ä, o, ö, u, ü, i, e;
- долгие гласные звуки: aː, äː, oː, öː, uː, üː, iː, eː;
- сверхкраткий гласный звук: ə;
- согласные звуки: m, p, b, f, v, n, ŋ, t, d, s, z, r, l, j, k, g, χ, ʁ, ʦ, ʨ, ʤ, ʣ, ʧ, ʃ;
- ^{j} – знак палатализации;
- ‿ – знак объединения орфографических слов в одно фонетическое.

Ниже приведен образец расшифровки в программе ELAN.

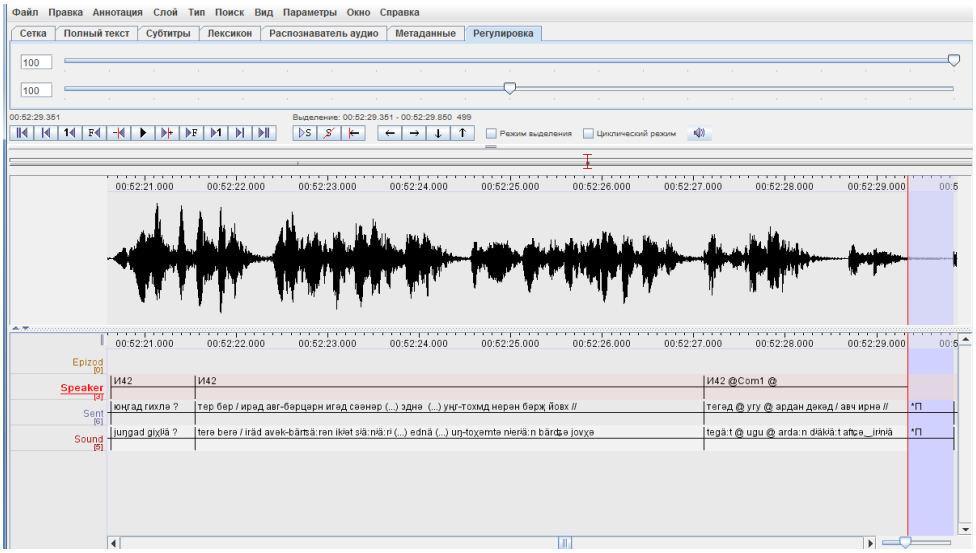

Рис. 2. *Образец расшифровки устной речи в программе ELAN*

В рамках проекта РГНФ «Национальный корпус калмыцкого языка: создание и разработка» (12-04-12047/в) планируется расшифровать 100 минут в орфографической записи. Работа по расшифровке крайне трудна, но тем не менее осуществляется. В реализации данной задачи привлекаются студенты, обучающиеся на калмыцком отделении Калмыцкого государственного университета, что становится достаточно хорошей практикой для них и их преподавателей.

На данный момент записана речь 12 информантов, в основном представителей старшего поколения: это более чем 22 часов записи (некоторые информанты приходят несколько раз). Данные по анкетированию были занесены в базу “KalmykSpeech” и проводится скоростная расшифровка звуковых файлов по мере записи новых информантов. Последующие этапы обработки материала связаны с детальной расшифровкой калмыцкой речи, описанием ее на разных уровнях. Ведется постоянный поиск инфор-

мантов, свободно владеющих родным языком. Мы стараемся записывать по одному информанту каждую неделю. Полная реализация проекта по записи калмыцкой речи принципиально важна, поскольку истинных носителей языка становится с каждым днем все меньше и меньше.

Примечания

[1] Во многом мы ориентировались на опыт работы по проекту «Один речевой день», в котором одному из авторов данной работы посчастливилось принять участие [8].

[2] См. подробнее: [8].

[3] Обычно к каждой теме был разработан перечень вопросов к информанту, направляющих беседу.

[4] Со сбором фольклорных произведений следует быть осторожными, поскольку они могут содержать следы иных говоров. Тексты фольклорных произведений кочуют с места на место вместе со сказителями и имеют свойство впитывать особенности речи наставников. Но собирать такие тексты однозначно нужно: пренебрегать таким материалом непростительно. Использовать такие тексты можно будет использовать в качестве материала для изучения фольклора и устной речи.

Список литературы

1. Американский фонетический алфавит // http://dic.academic.ru/dic.nsf/ruwiki/1361888

2. Крючкова О.Ю., Гольдин В.Е. Корпус русской диалектной речи: концепция и параметры оценки [электронный ресурс] // URL: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf (дата обращения: 30.11.2012).

3. Куканова В.В., Очирова Н.Ч. Общее или индивидуальное, норма или узус в Национальном корпусе калмыцкого языка: к постановке проблемы // Актуальные проблемы диалектологии языков народов России: Мат-лы XII Региональной конференции (Уфа, 27–28 ноября 2012 г.). Уфа, 2012. С. 90-94.

4. Летучий А.Б. Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы / отв. ред. В.А. Плунгян. СПб.: Нестор-История, 2009. С. 114-128.

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

6. Международный фонетический алфавит [электронный ресурс] // URL: http://ru.wikipedia.org/wiki/Международный_фонетический_алфавит (дата обращения: 15.07.2013).

7. Некрасова Г.А. Электронный диалектный корпус как ресурс сохранения и изучения коми диалектов // Языковая палитра. 2010. С. 13-16.

8. Степанова С.Б., Асиновский А.С., Богданова Н.В., Русакова М.В., Шерстинова Т.Ю. Звуковой корпус русского языка повседневного общения «один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008). М., 2008. http://www.dialog-21.ru/digests/dialog2008/materials/html/76.htm (дата обращения: 15.07.2013).

9. Уральский фонетический алфавит [электронный ресурс] // URL: http://ru.wikipedia.org/wiki/Уральский_фонетический_алфавит

10. ЮНЕСКО [электронный ресурс] // URL: http://www.unesco.org/culture/languages-atlas/ (дата обращения: 30.11.2012).

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