УДК 81`322.2 (519.682.5)

СОЗДАНИЕ НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Желтов П.В.

ФГБОУ ВПО «ЧГУ им. И.Н.Ульянова», Чебоксары, Россия (428015, Россия, г. Чебоксары, Московский проспект, 15), e-mail:chnk@mail.ru

В статье рассматривается задача создания Национального корпуса чувашского языка и связанные с ней проблемы и перспективы. Национальные языковые корпуса включают в себя большие массивы электронных текстов разных жанров и стилей, что дает возможность всесторонне и полно исследовать различные языковые явления. В отсутствии необходимого финансирования предлагается не добиваться создания полной текстовой базы чувашских текстов, а сделать репрезентативную выборку. Составлен минимальный список компьютерных программ, необходимых для работы с этой текстовой базой данных, рассмотрены вопросы разработки разметки для корпуса, а также обеспечения многопользовательского доступа через Интернет. Также рассмотрены вопросы безопасности. Отмечено, что наиболее безопасным будет использование отдельного сервера.

Ключевые слова: лингвистический корпус, машинный фонд, экстралингвистическая и лингвистическая разметка, многопользовательский доступ.

CREATING NATIONAL CORPORA OF THE CHUVASH LANGUAGE: PROBLEMS AND PERSPECTIVES

Zheltov P.V.

Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail:chnk@mail.ru

In the paper is analyzed the problem of creating the National corpora of Chuvash language and the problems and perspectives linked with it. The national linguistic corporas include large arrays of electronic text of different genres and styles, which gives the possibility to investigate comprehensively and fully different language phenomenas. While lacking necessary financement is proposed not to seek the creation of a full database of Chuvash texts but to make a representative selection. Was composed a shortlist of computer software, necessary for the work with this textual database, were considered questions of elaboration of a tagging system of the corpora, as well as the provision of multiuser access through the Internet. Were also considered question of security. Was noted that the best strategy would be the use of separate server.

Keywords: linguistic corpora, machine foundation, extra linguistic and linguistic tagging, multiuser access.

Одним из наиболее приоритетных направлений в современной компьютерной лингвистике является создание национальных языковых корпусов. Национальный языковый корпус представляет собой проиндексированную базу языковых данных, снабженных филологической разметкой и набором программных средств, осуществляющих лингвистический и лингвостатистический анализ в базе данных [5], [4].

Национальные языковые корпуса включают в себя большой массив электронных текстов разных жанров и стилей: художественные произведения, публицистику, фольклор, диалектный материал, научную, учебную и религиозную литературу, лексикографический материал.

Это дает возможность всесторонне и полно исследовать различные языковые явления и развитие языка. Например, подсчитать сколько раз то или иное слово встречается в

произведениях какого-либо жанра и автора, автоматически составить частотный словарь по произведениям какого-либо автора и т.п.

Используя подобную статистику по различным годам, можно определить изменения, произошедшие в языке на предмет использования тех или иных слов и выражений и т.п.

Филологическая разметка корпуса включает экстралингвистическую и лингвистическую разметки.

Экстралингвистическая разметка включает в себя информацию об авторе (фамилия, имя, отчество, псевдоним, годы жизни, пол) и о произведении (название, год создания, жанр, тематика, тип носителя – печатный или электронный, источник, год издания, издательство, стиль и т.п.).

Лингвистическая разметка (внутрилингвистическая) включает в себя, как правило, морфологическую и синтаксическую разметку (для устойчивых словосочетаний и предложных и послеложных конструкций в ряде языков). При этом морфологическая разметка дополняется, как правило, лексико-семантической информацией, на основе семантической классификации слов по тематическим классам.

В настоящее время существует большое количество национальных лингвистических (языковых) корпусов по различным языкам. В России свои языковые корпуса имеют русский, башкирский, татарский, бурятский осетинский, коми, марийский, эрзянский, мокшанский, удмуртский, вепсский, языки народов Дагестана, ведутся разработки по созданию корпусов для якутского и языков народов Сибири.

Поэтому актуальной задачей является создание Национального корпуса чувашского языка. Эта задача состоит из нескольких подзадач:

1) создание базы данных электронных текстов на чувашском языке;

2) создание обслуживающего ее программного обеспечения;

3) разметка указанной базы данных;

4) обеспечение доступа в сети Интернет.

Создание базы данных электронных текстов на чувашском языке

Реализация этой задачи представляется наиболее длительной и трудоемкой. В идеале подобная база данных должна содержать все тексты на чувашском языке, включая и рукописные фонды. Выполнение этого означало бы создание полного машинного фонда чувашского языка. Для большинства языков, указанных выше, созданию национальных лингвистических (языковых) корпусов предшествовало создание машинных фондов.

Для языков России наиболее полным в настоящее время является Машинный фонд русского языка, доступный в интернете по адресу www.cfrl.ru

Машинный фонд русского языка содержит текстовые и словарные источники общим объемом более 100 миллионов слов (словоупотреблений), занимающие дисковый объем более 1 гигабайта.

Создание Машинного фонда русского языка (МФРЯ) непосредственно занимался отдел Машинного фонда русского языка при Институте русского языка имени В.В. Виноградова, созданный в 1985 г. по инициативе академика А.П. Ершова после состоявшейся в 1983 г. «Всесоюзной конференции по созданию Машинного фонда русского языка» (первая всесоюзная конференция по данной тематике) [1].

В создании МФРЯ с 1986–1990 гг. также приняли участие более 40 организаций-соисполнителей, среди которых Московский, Санкт-Петербургский, Харьковский, Гродненский, Сыктывкарский, Саратовский университеты.

Для выполнения данного проекта была разработана и принята «Комплексная программа научных исследований и прикладных разработок по созданию Машинного фонда русского языка и информатизации исследований в Институте русского языка имени В.В. Виноградова», рассчитанная на период с 1996 по 2000 г., было проведено (до 1990 г.) три конференции по данной тематике [1–3].

Очевидно, что работы подобного масштаба не под силу современного бюджету Чувашской Республики или бюджету какого-либо одного республиканского учреждения. Основные затруднения вызывает сканирование, распознавание и проверка печатного материала, не представленного в электронном виде, которая представляет собой также чрезвычайную рутинную и утомительную для человека работу. Средняя продолжительность времени, необходимая для выполнения указанных действий на высоком по качеству уровне для произведения объемом 100 страниц, составляет 1 рабочую неделю (5 дней). Средняя оплата такого труда составляет 2000 рублей за 1 рабочую неделю. По подсчетам специалистов в настоящее время литературный фонд чувашского языка, не представленный в электронном виде, может насчитывать до нескольких миллионов страниц. Соответственно, подобного уровня финансовые затраты являются на данный момент неподъемными и нереальными в краткосрочной перспективе, а в долгосрочной перспективе соизмеримая сумма, с учетом приемлемых на проекты подобного рода ежегодных затрат в рамках республиканского бюджета и федеральных дотаций, могла бы быть выделена только в течение двух-трех десятков лет. Поэтому, в сложившейся ситуации, при отсутствии необходимого финансирования невозможно создание полного машинного фонда чувашского языка.

Необходимо ограничиться имеющимися в электронном формате текстами, составив из них репрезентативную с точки зрения жанров, стилей и исторического развития языка

выборку, и создать на основании её национальный корпус чувашского языка, предусмотрев его открытость и пополняемость.

В настоящее время достаточно большой электронной базой чувашских текстов располагают: Чувашское книжное издательство, Издательский дом «Хыпар», Национальная библиотека, Издательство Чувашского государственного университета имени И.Н. Ульянова, Чувашский государственный институт гуманитарных наук и частная Интернет-библиотека Николая Плотникова (доступная на сайте www.chuvash.org).

Большое количество произведений на чувашском языке печатается также через типографию «Новое время».

В настоящее время работы по накоплению текстов в электронном формате ведутся на базе Чувашского государственного института гуманитарных наук, имеющего к тому же достаточно большой собственный рукописный научный архив (в составе которого ценнейшие для чувашского языкознания и чувашеведения в целом архивы Н.И. Ашмарина и Н.В. Никольского).

Однако целесообразно расширить круг организаций и привлечь сюда вышеуказанные, а также Союз чувашских писателей, что позволит централизованно заключить соглашение с чувашскими писателями и ввести в электронную базу данных чувашских текстов целый ряд произведений, включению которых до этого в существующие электронные библиотеки препятствовало отсутствие требуемых законом об авторском праве разрешения и соглашений.

С учетом вышесказанного, необходимо разработать и принять концепцию Машинного фонда чувашского языка на уровне соответствующих республиканских министерств и ведомств. С учётом опыта создания подобных фондов целесообразно реализовывать Машинный фонд чувашского языка как открытую сетевую систему с доступом через Интернет.

Отдельным разделом подобного фонда должны быть материалы по чувашскому языку и культуре, написанные на других языках (русском, немецком, венгерском и т.д.). Значимость подобных материалов велика, а их количество достаточно большое. При этом многие из них в Чувашской Республике недоступны вообще, т.к. находятся за её пределами.

А так как многие из них отражают историческое развитие чувашского языка, то без учета их материала невозможно полноценное исследование чувашского языка в историческом аспекте. В дальнейшем целесообразным является их перевод на современный чувашский язык и включение в состав фонда в соответствующие его разделы.

Общая структура Машинного фонда чувашского языка представляется следующей (рис. 1):

Рис. 1. Структура Машинного фонда чувашского языка (МФЧЯ)

Для администрирования и моделирования МФЧЯ необходимо создать рабочую группу, работающую на общественных началах.

В её задачи должно входить:

- проверка загружаемого пользователями контента;
- соблюдение при загрузке нового контента авторских прав;
- создание при необходимости новых разделов фонда;
- загрузка нового контента.

В связи с указанными функциями и обязанностями и распределённым характером базы данных МФЧЯ, который в сложившихся условиях будет складываться из электронных ресурсов целого ряда организаций, представляется целесообразным равное представление в рабочей группе всех участвующих в этом процессе организаций. Общее же руководство рабочей группой и процессом создания МФЧЯ представляется целесообразным принять на себя Чувашскому государственному институту гуманитарных наук.

Структуру разделов МФЧЯ можно предварительно определить следующим образом:

- картотека МФЧЯ;
- лексикографический и лексикологический подфонд;
- архивный подфонд;
- старопечатные издания;
- электронные издания;
- диалектологический подфонд.

Создание обслуживающего программного обеспечения

К программному обеспечению, обслуживающему национальный корпус чувашского языка, относятся:

- морфологический анализатор чувашского языка;
- синтаксический анализатор чувашского языка;
- семантический анализатор чувашского языка;
- корректор орфографии чувашского языка;
- система разметок (тэгов) для морфологии, синтаксиса и семантики чувашского языка;
- система автоматического составления частотных словарей;
- система автоматического составления тезаурусов;
- система лингвостатистического анализа.

Как видим из этого списка, который является минимальным и далеко не полным, разработка обслуживающего НКЧЯ программного обеспечения требует также либо привлечения большого коллектива программистов с целью её осуществления в короткие сроки (в пределах нескольких лет), что в сложившейся вокруг проекта финансовой ситуации просто нереально, либо долговременной и планомерной работы, рассчитанной на десятилетия.

В настоящее время подобная работа ведется при отделе лексикографии и лексикологии Чувашского государственного гуманитарного института.

Однако ввиду отсутствия необходимости финансирования, работа продвигается крайне медленно и непозволительно затягивается даже с расчетом на долговременную перспективу.

Поэтому необходимо привлечение на общественных началах или с минимальной оплатой большого количества специалистов.

Помимо разработки программного обеспечения для анализа чувашских текстов необходимо также разработать или получить в свободное пользование программное обеспечение, способное выделять русские тексты, абзацы или предложения, которые в ряде жанров (например, периодике и газетных изданиях) часто включены в чувашский текст (статьи федеральных законов, цитаты из интервью и высказываний и т.п.).

Для этого нужно как минимум иметь в составе обслуживающего корпус программного обеспечения – морфологический анализатор русского языка.

Разметка базы данных

Как уже указывалось выше, в национальных лингвистических (языковых) корпусах существует два типа разметок: экстралингвистическая и лингвистическая

(внутрилингвистическая). Последняя делится на морфологическую, синтаксическую и лексико-семантическую.

В настоящее время частично разработана морфологическая разметка для чувашского языка, которая должна быть дополнена и доработана.

В качестве примера научно обоснованной и удачной разметки для тюркских языков можно привести систему разметов в национальном корпусе башкирского языка, которая была опубликована в [5].

В качестве лексико-семантической разметки можно воспользоваться системой лексико-семантических разметок для русского языка [4] или аналогичной системой разметок для башкирского языка [5].

Обеспечение доступа в сети Интернет

В этом отношении существует несколько важных аспектов: возможность быстрого и многопользовательского доступа, а также его безопасность.

Для этого желательно использовать специально выделенный и настроенный сервер с определенной системой защиты, а также систему транзакций, для осуществления резервного копирования обновлений, сделанных авторизированными и имеющими соответствующие права пользователями.

В случае использования специального выделенного сервера возможна организация более защищённой системы, чем в случае использования т.н. удалённого сервера (или виртуального).

Таким образом, задача создания НКЧЯ представляется перспективной и актуальной научной и практической задачей, которая имеет, однако, и определенные трудности, которые можно обойти при наличии необходимого финансирования и применении правильной стратегии его создания.

Публикация подготовлена в рамках поддержанного РГНФ научного проекта № 15-04-00532.

Список литературы

1. Машинный фонд русского языка: идеи и суждения //Материалы I Всесоюзной конференции по созданию МФРЯ. – М.: Наука, 1986. – 234 с.

2. Материалы II Всесоюзной конференции по созданию МФРЯ. – М.: Наука, 1988. – 230 с.

3. Материалы III Всесоюзной конференции по созданию МФРЯ. – М.: Изд-во МГУ имени М.В. Ломоносова, 1990. – 148 с.

4. Плунгян В.А. Национальный корпус русского языка: опыт создания корпуса текстов современного русского языка / В.А. Плунгян, Д.В. Сичинава // Труды международной конференции «Корпусная лингвистика-2004». – СПб: Изд-во Санкт-Петербургского университета, 2004. – С. 216-238.
5. Бускунбаева Л.А. Система разметок в национальном корпусе башкирского языка /Л.А. Бускунбаева, З.А. Сиразитдинов // Материалы международной конференции «Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы». – Йошкар-Ола, 2011. – С. 46-51.

Рецензенты:

Губанов А.Р., д.фил.н., профессор кафедры русского языка как иностранного ФГБОУ ВПО «Чувашский государственный университет имени И.Н. Ульянова», г. Чебоксары;

Охоткин Г.П., д.т.н., профессор, заведующий кафедрой автоматики и управления в технических системах ФГБОУ ВПО «Чувашский государственный университет имени И.Н. Ульянова», г. Чебоксары.