Министерство образования Республики Беларусь

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«ГРОДНЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ ЯНКИ КУПАЛЫ»

СЛОВО И СЛОВАРЬ
VOCABULUM ET VOCABULARIUM

СБОРНИК НАУЧНЫХ ТРУДОВ ПО ЛЕКСИКОГРАФИИ

Гродно
ГрГУ им. Я. Купалы
2009

эта работа ведется в Институте литовского языка. В современных хранилищах имеются личные архивы видного литовского лексикографа Йонаса Паулаускаса (1923 – 2003), известного диалектолога Эляны Гринавецкенс (1928 – 1999), одного из авторов академической «Грамматики литовского языка» и редактора «Словаря литовского языка» Казиса Ульвидаса (1910 – 1996), известного специалиста в области терминологии и лексикографии, редактора «Словаря литовского языка» Йонаса Круопаса (1908 – 1975) и других ученых. Недавно Институт литовского языка получил часть архива выдающегося ученого, филолога, балтиста и индоевропеиста, а также специалиста в области мифологии и истории культуры Владимира Топорова (1928 – 2005).

Все архивы описываются, их материалы тщательно каталогизируются и вводятся в научный оборот. Это научная деятельность *Очага литуанистики*.

Открытие *Очага литуанистики* способствовало поиску новых форм популяризации научных знаний. В Музее языка проводятся открытые уроки для учащихся, экскурсии и др. мероприятия. Благодаря творческому потенциалу и изобретательности создателей Музей стал особым местом притяжения, где через презентацию научной продукции возможна обратная связь, т.е. контакт и диалог с обществом. А это, как оказывается, весьма существенно влияет на политиков, от решений которых в конечном счете зависит финансирование литуанистических исследований.

Список литературы

1. Zabarskaitė, J. Lituanistikos židinys. Kalbos muziejus / J. Zabarskaitė. – Vilnius, 2008. ISBN 978-9955-704-63-8 (по-литовски); ISBN 978-9955-704-62-1 (по-французски); ISBN 978-9955-704-61-4 (по-английски).

2. Zabarskaitė, J. Lietuvių kalbos centras: tikslai ir perspektyvos / J. Zabarskaitė // Best Practices of Learning Less widely-used Languages in Multicultural and Multinational Europe. Materials of the conference, 23–25 september, 2004. – Vilnius: Lietuvių kalbos instituto leidykla, 2004.

Забарскайте Йоланта, директор Института литовского языка, старший научный сотрудник, доцент, доктор (Вильнюс, Литва).

Морозова Надежда, старший научный сотрудник Института литовского языка, доцент, доктор (Вильнюс, Литва).

The Hearth of Lithuanian Studies was founded in 2006. It is the only centre of this type in Lithuania.

Aims: to disseminate the latest news in linguistics, to raise the society awareness of the Lithuanian language as a key factor for preserving national identity and citizenship, to promote ideas of European plurilingualism and inter-cultural dialogue. The Museum provides access to the documents from the Archive, interesting items from everyday life of linguists, other unique materials; electronic file directory offering valuable information in various fields of linguistics: language history, dialectology, lexicography, language contacts research, history of writing, terminology, contemporary language usage, history of linguistics, sign language, etc; computer games, dialect karaoke, interesting computer books; language toys; unique exhibits, which can help «touch the language». The most unexpected section of the Museum for the visitors is the section of Language Toys. The idea is both complex and simple. Visitors are invited to touch language with their hands and thus learn something new about it.

There are some interactive tools for investigation of Lithuanian lexis and semantics, such as: Pyramid of meanings; House of word origin; Words as shadows; Words as fossils, etc. Language toys invite to look differently at lexis and semantics, to see the language variety and power, its past. Language toys help the visitors to understand that Language is not only the instrument of communication.

И.Б. Качинская

КОРПУСНАЯ ЛИНГВИСТИКА В ДИАЛЕКТОЛОГИИ И ЭЛЕКТРОННАЯ КАРТОТЕКА «АРХАНГЕЛЬСКОГО ОБЛАСТНОГО СЛОВАРЯ»[1]

В рамках темы «Компьютерные технологии в лексикографии» рассказывается о корпусе «Электронная картотека "Архангельского областного словаря"». Так как полевые записи для Словаря осуществлены в стандартной транскрипции, актуальной проблемой для Корпуса является переход от фонетической словоформы к грамматической и далее к начальной (лемматизация). На презентации будет продемонстрирована автоматизированная обработка материала – перевод цельного текста полевой тетради в базу данных с заполненными полями и словоформами, расставленными в алфавитном порядке. Упоминается о крупнейшем корпусе, находящемся в открытом доступе, – Национальном корпусе русского языка – и его Диалектном подкорпусе.

0. Корпусная лингвистика развивается в настоящее время мощными темпами. Крупнейшим проектом является общедоступный Национальный корпус русского языка (НКРЯ, **www.ruscorpora.ru**). Он делится на подкорпуса: Основной, Обучающий, Синтаксический, Параллельный (содержащий переводы с иностранных языков на русский и наоборот), Акцентологический (как правило, содержащий поэтические тексты), Поэтический, Устной речи, Диалектный. Сейчас Корпус диалектных текстов находится в стадии обновления: пересматриваются принципы первоначальной подачи материала (предполагается подача текстов в транскрипции и с ударениями) и принципы метаразметки (грамматической и семантической)[2]. В Диалектном подкорпусе возможен поиск по многим грамматическим (в том числе специфически диалектным) параметрам, по некоторым фонетическим, по тематике, семантике и проч.

Пользуясь случаем, обращаюсь ко всем диалектологам: материалы, записанные в любых **русских** говорах на любой территории (исконного проживания, раннего и позднего заселения, а также мигра-

[1] Работа поддержана грантом РГНФ № 08-04-12132в («Грамматический словарь северных говоров. Электронная версия (на базе Словника "Архангельского областного словаря"»).

[2] Работа поддержана грантом РГНФ № 09-04-12159в («Корпус диалектных текстов Национального корпуса русского языка: грамматическая, фонетическая и метаязыковая разметка. Новый стандарт подачи»).

ции) и десятилетиями хранящиеся на кафедрах институтов, университетов и в частных собраниях, могут наконец быть опубликованы на сайте НКРЯ, дойти до читателя и исследователя. Их можно присылать по адресу: **kacza@yandex.ru.**

1. Темой моего доклада будет несколько более узкая область, но тоже связанная с корпусной лингвистикой и одновременно с диалектной лексикографией. На кафедре русского языка филологического факультета Московского государственного университета имени М.В. Ломоносова с 1956 г. ведется работа по составлению «Архангельского областного словаря» (под редакцией О.Г. Гецовой. Издано 13 выпусков, издание продолжается).

«Архангельский областной словарь» (АОС) – крупнейший диалектный словарь одного региона. Его архив содержит более 2-х тыс. полевых тетрадей и ежегодно пополняется на 50 – 100 тетрадей, записанных в рамках студенческой диалектологической практики. Количество «бумажных» карточек составляет около 5 млн. (ежегодное пополнение: 20 – 40 тыс. карточек). Словник АОС включает около 180 тыс. слов (в 2006 г. был опубликован «Обратный словарь архангельских говоров»). В 12-м выпуске закончился материал на букву Д. Том на Е-Ж пришлось разделить на два выпуска, слова на букву Ж закончатся в 14-м вып. Предполагаемый общий объем издания – не менее 60 выпусков.

2. Традиционно обработка полевых тетрадей сводилась к созданию «бумажной» картотеки и включала следующие этапы: 1) расписывание полевых тетрадей (или расшифровок аудиозаписей) на карточки; 2) карточки расставлялись в алфавитном порядке; 3) выявлялись новые слова по Словнику АОС.

С 1996 г. началась работа по созданию Корпуса «Электронная картотека АОС», в базе уже более 1 млн. «карточек» (7 – 8 млн. словоупотреблений). Создание Корпуса АОС позволило: 1) обеспечить лучшую сохранность материалов АОС (бесценная, десятилетиями собираемая картотека – это бумажные карточки в деревянных каталожных ящиках; уже давно стояла проблема сохранности архива); 2) частично решить проблему постоянного дефицита места хранения новых карточек; 3) вносить добавления нового материала в готовящиеся к изданию выпуски АОС почти в «готовом» виде; 4) использовать материалы АОС в самых различных научных целях.

В память компьютера вводятся полевые тетради, в первую очередь нерасписанные, из экспедиций последних лет. Расписанные тетради из архива тоже постепенно вводятся. Полевые тетради набираются студентами-филологами, русистами, в рамках камеральной студенческой диалектологической практики. Чтобы наборщик мог работать в удобной для него программе, для фонетической транскрипции используются заранее оговоренные символы. Потом тексты обрабатываются в специально для этого созданной программе перекодировки (расставляются ударения), выверяются по первоначальной записи и размечаются (производится разметка диалектизмов). По каждой тетради создается своя база данных, после чего они объединяются в общую, сводную базу. Далее производится сортировка по словоформе диалектного слова или по ключевому диалектному слову, данному в орфографии, хотя сортировка может осуществляться и по любому другому полю (например, по району записи, по пункту). На сегодняшний день обработано более 600 тетрадей, записанных в 72 пунктах в большинстве районов Архангельской области.

Электронная картотека АОС создана на основе СУБД StarLing (автор – покойный Сергей Анатольевич Старостин, чл.-корр. РАН). Эта база делает возможной работу с фонетической транскрипцией любого уровня, так как позволяет включать произвольные шрифты, знаки и диакритики, сортировать материал в заданном пользователем алфавитном порядке: пользователь может произвольно включать любые знаки, заранее объявляя их последовательность или, напротив, приравнивая их друг к другу, например: е = ё или А = А-ударное прописное = а = а-ударное строчное. СУБД StarLing, на основе которой создана Информационная система «Электронная картотека "Архангельского областного словаря"», находится в открытом доступе в Интернете на сайте С.А. Старостина: **http.starling.rinet.ru**.

3. Очень важной проблемой в Корпусе, ориентированном на Словарь, является **лемматизация** (приведение словоформы к начальной форме). В 2003 – 2006 гг. авторам для работы над буквами Е, Ж было передано из электронной базы дополнение в 16,5 тыс. словоупотреблений. Почти все эти 16,5 тыс. «карточек» пришлось обрабатывать вручную, заменяя фонетическую словоформу начальной непосредственно в Базе.

Проблема лемматизации, во многом решенная для литературного языка, для диалектного языка еще не решена. Лемматизация основывается на приведении заранее заданных словоформ *письменного языка* к начальной форме слова, заранее заданной в словарях литературного языка. Тогда как основной текст в базе данных – это фонетическая транскрипция, часто разного уровня, иногда с разными способами графической передачи близких фонетических явлений (*есть = есьть = йес'т' = йес' = jec'* и т.д.). Кроме того, в распоряжении диалектологов регулярно оказываются записи, произведенные самими диалектоносителями, не всегда грамотными; эти записи также вводятся в память компьютера, необходима и их адекватная обработка; грамматические характеристики слова в говоре часто отличаются от таковых в литературном языке.

Поэтому следующим этапом работы с Электронной картотекой АОС стала работа по лемматизации. Проект «Грамматический словарь северных говоров. Электронная версия (на базе Словника "Архангельского областного словаря")» явился продолжением работы по проекту «Создание грамматического словаря северных говоров (на базе транскрипционной записи устной диалектной речи)»[3]. Тогда движение шло от фонетической словоформы к грамматической и к начальной (от материалов полевых экспедиционных тетрадей, включенных в Корпус, – к словнику). При этом пришлось проделать большую предварительную работу: например, создать «детранскриптор» – перевести фонетическую словоформу в грамматическую, т.е. в подобие орфографической. Далее был запущен грамматический анализатор Старостина-Зализняка, который сначала опознал около 20 %

[3] Работа была поддержана грантом РГНФ № 05-04-04274а.

словоформ, потом опознавал до 70 %, но в результате выяснилось, что многие из них были опознаны неверно. В настоящем Проекте движение шло наоборот: от начальной формы – к грамматической словоформе. Для этого тоже пришлось проделать большую работу: в электронном варианте Словника АОС все слова были разделены по их грамматической принадлежности. Работа велась отдельно по каждой из 5 частей речи, имеющих словоизменительные категории, т.е. с существительными, прилагательными, местоимениями, числительными и глаголами. Снова был запущен грамматический анализатор Старостина-Зализняка, каждое слово получило свой словоизменительный индекс. Была проведена индексация всех случаев, где Словник АОС совпадает с Грамматическим словарем А.А. Зализняка (повторены индексы у общерусских слов[4]; повторены индексы во всех префиксальных образованиях; принудительно назначены соответствующие индексы во всех финалях собственно диалектных слов).

Например, по существительным (около 73 тыс. лексем) первоначально было опознано всего около 15 % лексем. Пришлось учесть особую орфографию АОС, отличающуюся от орфографии, принятой в литературном языке. Это отсутствие *Ь* у существительных 3 скл. (*ноч, доч, мыш*); написание приставок *раз-/рас-* как *роз-/рос-*; написание *е/ё* после мягких шипящих (*яицё, пальтецё*) и *о* после твердых шипящих (*жонка, жорнов*); написание *е* в суффиксе вместо *-иц-* в ЛЯ (*здоровьеце, отделеньеце*); наличие у субстантивов и слов со смешанным склонением окончаний *-ой/-ей* (вместо *-ый/-ий*) (*егорей, жабей, жолвей*) и проч. Программа учла различные написания и приравняла их к формам, уже имеющимся в Словаре Зализняка (*доч = дочь, василей = василий, яйцё = яйцо*). После этой унификации анализатор опознал около 20 % слов. Возникло много сложностей в связи с нечеткой разработанностью помет, сделанных для Обратного словаря. Так, например, в Словнике особо помечены все существительные на *-ой/-ей*, для того чтобы отличать их от прилагательных. Но при этом одинаково помеченными оказались существительные типа *зной, промой* (имеющие стандартное 2 скл. муж. рода) и *любезной, божатой* (имеющие адъективное склонение). В случаях, когда слово имелось в Грамматическом словаре Зализняка, оно, конечно, опознавалось правильно (*зной*). Но таких случаев оказалось немного. Трудности возникли и с определением мн. числа существительных. В АОС помета мн. ч. проставлена только у существительных с И. мн. на *-а* (*долгощёла, козла, пёрла*). У существительных на *-е* такой пометы, к сожалению, не было. Таким образом, слова *море, промёжговне* и *бояре, прихожане* оказались представлены для парсера одинаково.

Правильность постановки индекса при начальной форме слова проверялась 1) при просмотре таблиц: анализатор строит таблицы всех гипотетических словоформ (не диалектных); 2) при попытке получить лемму из поля словоформы в Корпусе.

От учета диалектной грамматики анализатором пришлось отказаться, так как это резко увеличило омонимию и ничуть не уменьшило количество ручной обработки.

Например, в северных говорах у существительных встречается вариативность флексий *е/и* в ед. ч. Д.-П. I скл. (*к козы, в Москвы*), флексий *е/у/и* в ед. ч. П.п. II скл. (*в лесе, в городу, на кони*), расширена сфера употребления флексии *-у* в ед. ч. Р.п. II скл. (*около городу, у мужику*), вариативны окончания мн. ч. в И., Р., Тв.; смешиваются парадигмы склонения «разносклоняемых» существительных (*день, путь, время, мать, дочь*), регулярна ориентация слов III скл. на I (*на пече*) и т.д. Некоторые словоформы оказалось возможным задать анализатору «списком» (*братовья, браты, братеи* – И. мн.; *братьёв, братовей, братовьёв, братовьей, братей, братьей, братеей, сыновьей, сыновьёв* – Р.-В. мн.). Списком даны словоформы личных и некоторых других местоимений. Но от многих случаев пришлось отказаться из-за омонимии словоформ. Так, например, форма *дочи* зафиксирована практически во всех падежах ед. ч. (кроме Тв.) и в И. мн.[5]. С другой стороны, словоформе *дочи* в Р. ед. может быть приписано сразу 3 леммы: *дочи, дочь, доча*, словоформе *дочери* в Р., Д., П. ед. – тоже 3 леммы: *дочь, дочи, дочерь*.

4. Доклад предполагается сопроводить презентацией. На экране проектора будут показаны

1) страницы полевых тетрадей;

2) образцы «бумажных» карточек АОС;

3) памятка студентам, вводящим в компьютер полевые тетради;

4) страницы полевой тетради после первоначального набора;

5) страницы полевой тетради, подготовленные для обработки в dbf;

6) образец Корпуса АОС в dbf с заполненными полями;

7) извлеченные из dbf образцы материала на букву Д, предоставленные авторам для работы над словарными статьями АОС (12 выпуском).

8) Будет продемонстрирована автоматизированная обработка тетради – перевод цельного текста полевой тетради в dbf с заполненными полями. В поле «лексема» с помощью детранскриптора слова из фонетической словоформы превращаются в грамматическую (насколько это возможно сделать с помощью автоматического анализатора) и расставляются по алфавиту. Обработка одной тетради в 90 страниц занимает примерно 1 минуту.

Программы по введению размеченного текста в Базу, программы лемматизации фонетических словоформ могут быть использованы не только для пополнения Словника АОС или для предоставления авторам АОС дополнительного материала из Электронной картотеки для работы над дальнейшими выпусками, но и для помощи в создании Словарей и Слов-

[4] В Словник АОС, как и в любой диалектный словник или словарь дифференциального типа, включено большое количество общерусских слов, так как эти слова часто отличаются своей семантикой от своих аналогов в литературном языке.

[5] Исторически эта омонимичность легко объяснима: В В. ед. вместо архаической формы *дочерь* возникает новая форма *дочи*, ориентированная на И. ед. В то же время архаическая форма И. ед. *дочи* уступает место более новой форме *дочь* (в синхронии обе формы равнозначны и одинаково частотны). На базе И. ед. *дочь* происходит унификация парадигмы и ориентация ее на стандартное 3 скл., в котором Р.=Д.=П.=И. мн.

ников любых диалектных корпусов по имеющимся полевым материалам, введенным в память компьютера. Изданием диалектных словарей сейчас занимаются многие региональные университеты. Московский государственный университет имени М.В. Ломоносова всегда готов оказывать посильную помощь и проводить любые консультации по интересующим коллег проблемам.

Качинская Ирина Борисовна, младший научный сотрудник МГУ имени М.В. Ломоносова (Москва, Россия).

Within the limits of a theme «Computer technologies in a lexicography» in the report it will be told about Corpus «Electronic file of the Arkhangelsk Region Dialect Dictionary». As field records for the Dictionary are carried out in a standard transcription, an actual problem for Corpus is transition from a phonetic to grammar word form and further to initial form. On presentation the automated processing of a material – translation of the integral text of a field writing-book in a database with the filled fields and the word forms placed in alphabetic order will be shown. It is mentioned the largest Corpus which is in open access, – Russian National Corpus – and its Dialectal Corpus.