Диалектная лексикография: электронная картотека «Архангельского областного словаря»

Dialect lexicography: an electronic corpus of «The Arkhangelsk region dialect dictionary»

Качинская И. Б. (kacza@rambler.ru)
МГУ имени М. В. Ломоносова, Москва

Крылов С. А. (krylov-58@mail.ru)
Институт востоковедения РАН, Москва

«Архангельский областной словарь» (АОС) — крупнейший диалектный словарь одного региона, его словник насчитывает ок. 180 тыс. слов. Основой словаря является «бумажная» картотека — в ней более 5 млн карточек. В Электронной картотеке АОС содержится уже более 1 млн «карточек». Важной проблемой Электронной картотеки является совершенствование программ по лемматизации: по переходу от фонетической словормы (первоначальные полевые записи осуществляются в фонетической транскрипции) к грамматической и далее к начальной форме слова. Доклад сопровождается презентацией с демонстрацией автоматизированной обработки материала: (а) перевода текста полевой тетради (rtf) в базу данных (dbf) с заполненными полями; (б) грамматическими словоформами, расставленными в алфавитном порядке; и (в) гипотетическими начальными формами, созданными автоматическим анализатором в среде StarLing.

1. «Архангельский областной словарь» (АОС) под редакцией О. Г. Гецовой — крупнейший диалектный словарь одного региона. Его архив содержит более 2 тыс. полевых тетрадей и ежегодно пополняется на 50–100 тетрадей, записанных студентами русского отделения филологического факультета МГУ в рамках полевой диалектологической практики, аспирантами и руководителями практики[1]. Количество «бумажных» карточек составляет ок. 5 млн (ежегодное пополнение — 20–40 тыс. карточек). Словник АОС включает ок. 180 тыс. слов[2]. В 12-м выпуске АОС закончился материал на букву Д[3]. 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»[4], в базе уже более 1 млн «карточек» (ок. 10 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) ча-

[1] Картотека «Архангельского областного словаря» и архив тетрадей хранятся в кабинете диалектологии кафедры русского языка филологического факультета МГУ.

[2] Обратный словарь архангельских говоров / Под ред. О. Г. Гецовой. М.: «Наука», 2006.

[3] Первые 9 выпусков АОС выставлены на сайте http://www.philol.msu.ru/~dialectology/dictionary/ в формате .pdf (филологический факультет МГУ, каф. русского языка, каб. диалектологии, Словарь).

[4] Работа была поддержана грантом РГНФ, проект № 02-04-12020 в «Электронная картотека “Архангельского областного словаря”». Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов, С. Г. Болотов.

стично 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; постепенно вводятся расписанные тетради из архива; начинается ввод «старой» картотеки, уже использованной для создания первых выпусков АОС (в ней теперь видится много полезного материала, не учтенного в ранних выпусках); предполагается ввод нового материала (буквы А-Ж), поступившего в картотеку после выхода соответствующих томов[5].

Полевые тетради набираются студентами-русистами в рамках камеральной студенческой диалектологической практики. По каждой тетради создается своя база данных, после чего они объединяются в общую, сводную базу. Далее производится сортировка по словоформе размеченного диалектного слова или по ключевому диалектному слову, данному в орфографической записи, — хотя сортировка может осуществляться и по любому другому полю (например, по району записи, населенному пункту). На сегодняшний день обработано более 600 тетрадей, записанных в 72 нас. пунктах в большинстве районов Архангельской области. Основной приток нового материала для работы над АОС сейчас идет именно через Электронную базу данных (БД АОС).

Электронная картотека АОС создана на основе СУБД StarLing (автор — Сергей Анатольевич Старостин). Эта база делает возможной работу с фонетической транскрипцией любого уровня, т. к. позволяет включать произвольные шрифты, знаки и диакритики; позволяет сортировать материал в заданном алфавитном порядке: пользователь может произвольно включать любые знаки, заранее объявляя их последовательность или, напротив, приравнивая их друг к другу, например: е = ё или А = Á-ударное прописное = а = á-ударное строчное[6].

3. Предваряя онлайновую версию проекта «Вавилонская башня», С. А. Старостин писал: «Помимо этимологии и сравнительно-исторического языкознания, многолетний предмет моих штудий — автоматическая морфология русского языка. На этих страницах[7]вы имеете возможность ознакомиться с компьютерными базами данных по словарям Ожегова, Зализняка и Мюллера, а также проанализировать любое русское слово и получить его полную акцентуированную парадигму»[8].

Проблема лемматизации, во многом решенная для автоматической обработки текстов русского литературного языка, для диалектных текстов еще не решена. При традиционном подходе лемматизация обычно основывается на приведении заранее заданных словоформ *письменного языка* к начальной форме слова, заранее заданной в словарях литературного языка. Между тем основной текст в БД АОС — это текст в фонетической транскрипции, часто разного уровня, иногда с разными способами графической передачи близких фонетических явлений (*есть = éсьть = йéс′т′ = йес′ = jéс′*, для *есть-2* еще и *juc′* в разных графических вариантах — 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 — восстановлению начальной формы (леммы) в ее орфографическом варианте из фонетической словоформы. Это необходимо (а) для пополнения Словника АОС; (б) для передачи авторам нового материала из Корпуса АОС при написании словарных статей в следующие выпуски и для добавления нового материала в уже написанные словарные статьи; (в) для корректного поиска лексем при работе с Корпусом «Электронная картотека АОС». 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 тысячи «карточек» пришлось обрабатывать вручную, заменяя непосредственно в Базе словоформу в фонетической транскрипции начальной формой в орфографической записи.

4. В первый период работы мы шли от словоформы (в фонетической транскрипции) — к начальной форме (в орфографической записи), т. е. от материалов полевых экспедиционных тетрадей, включенных в Электронный Корпус, к Словнику[9]. Для этого пришлось проделать значительную по объему предварительную работу. Было создано два «детранскриптора». Детранскриптор-1 переводит запись

---

[5] Задумано создание он-лайновой версии АОС с регулярным пополнением.

[6] СУБД StarLing, на основе которой создан Корпус «Электронная картотека “Архангельского областного словаря”», находится в открытом доступе в Интернете на сайте С. А. Старостина: http.starling.rinet.ru.

[7] Имеется в виду сайт проекта «Эволюция языка» — «Вавилонская башня».

[8] http://starling.rinet.ru/morpho.php?lan=ru

[9] Работа была поддержана грантом РГНФ, проект № 05-04-04274а «Создание грамматического словаря северных говоров (на базе транскрипционной записи устной диалектной речи)». Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов.

в традиционной фонетической транскрипции, содержащуюся в полевых тетрадях, в «облегченную» фонетическую транскрипцию, принятую в изданиях АОС, и тем самым во многом унифицирует способ подачи записи. Детранскриптор-2 переводит фонетическую словоформу (в любом варианте фонетической транскрипции) в подобие грамматической орфографической формы (поле lex). В этом поле запускается грамматический анализатор Старостина–Зализняка[10]. Сначала анализатор опознал около 20 % словоформ, потом опознавал даже до 70 %, но в результате выяснилось, что большинство из них были опознаны неверно.

Пришлось учесть особую орфографию АОС, отличающуюся от стандартной орфографии. Это отсутствие *Ь* у существительных 3 скл. (*ноч, доч, мыш*); написание корней *раст-* как *рост-* (*ростъ*), *раб-* как *роб-* (*робутать*); приставок *раз-/рас-* как *роз-/рос-*; объединение приставок *при-* и *пре-* в одну приставку *при-* (*прикра́сной, прикрати́ться*); написание *е/ё* после мягких шипящих (*яицё, пальтецё*) и *о* после твердых шипящих (*жо́нка, жо́рнов*); написание *е* в суффиксе вместо *-иц-* в ЛЯ (*здоро́вьеце, отделе́ньеце*); наличие у субстантивов и слов со смешанным склонением окончаний *-ой/-ей* (вместо *-ый/-ий*) (*ле́шой, зимной, зимней, егорей, жабей, жо́лвей*) и проч. Программа учла эти отличия и приравняла их к формам в стандартной орфографии, имеющимся в Словаре Зализняка (*доч = дочь, василей = василий, яйцё = яйцо*).

5. Признав путь от словоформы к начальной форме слова на начальном этапе несколько преждевременным, мы пошли в обратную сторону: от начальной формы слова, зафиксированной в Словнике, к его грамматической словоформе, чтобы впоследствии уже новая программа «узнанную» грамматическую форму (а) возводила к начальной, зафиксированной в Словнике; (б) предлагала несколько начальных форм из Словника на выбор в случаях омонимии; (в) предлагала новую (гипотетическую) начальную форму, отсутствующую в Словнике[11].

На основе электронной версии Словника АОС был создан вариант словника с грамматической разметкой, т. е. с указанием частеречной принадлежности *всех* слов — в «Обратном словаре архангельских говоров» указана частеречная принадлежность всех служебных слов, местоимений и наречий; отметка о части речи у прилагательных, существительных и глаголов присутствует лишь в особых случаях. Снова был запущен грамматический анализатор Старостина-Зализняка, каждое слово получило свой словоизменительный индекс. Таким образом, была проведена индексация всех случаев, где Словник АОС совпадал с Грамматическим словарем А. А. Зализняка (а именно, повторены индексы у общерусских слов[12]). Велась активная работа по созданию программ, расширяющих возможности грамматического анализатора Старостина–Зализняка. Лексемы, не опознанные анализатором, размечались принудительно: были повторены индексы у всех префиксальных образований; назначены соответствующие индексы у собственно диалектных слов, имеющих определенные финали, с привлечением некоторых алгоритмов, которые удалось записать специальными строками в Программу. Работа велась главным образом по существительным (более 70 тыс. лексем), глаголам (более 63 тыс. лексем) и прилагательным (более 16 тыс. лексем).

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 *любе́зной, божа́той* (имеющие адъективное склонение). В случаях, когда слово имелось в Грамматическом словаре Зализняка, оно, конечно, опознавалось правильно (*зной*). Но таких случаев оказалось немного. Грамматическая информация, которая представлялась «лишней» для пользователя-лингвиста из-за своей «очевидности», для парсера такой очевидностью не обладает.

Словообразование в говорах происходит гораздо интенсивнее, чем в литературном языке, в т. ч. префиксальное и постфиксальное. Поэтому в программе произведены специальные записи, приравнивающие формы с различными префиксами к формам (с иными префиксами или без них), имеющимся в Словаре Зализняка (*добеспоко́иться = беспоко́иться, дозаставля́ть = заставля́ть*). Различие в постфиксах в случае их наличия (*гости́ться, дозаусыпля́ться*) оказалось не столь актуальным, т. к. модель Зализняка–Старостина практически везде порождает гипотетические постфиксальные формы, рассматривая их как пассив. В то же время каждый раз в качестве начальной формы она предлагает либо две глагольные формы (с *-ся* и без *-ся*), либо одну (без *-ся*). Этот момент оказалось легко преодолеть, прописав строку, сравнивающую словоформу с конечным *-ся* с начальной формой слова.

---

[10] Находятся в свободном доступе на http://starling.rinet.ru/downl.php?lan=ru#soft

[11] Работа была поддержана грантом РГНФ, проект № 08-04-12132в «Грамматический словарь северных говоров. Электронная версия (на базе Словника "Архангельского областного словаря")» Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов.

[12] В Словник АОС, как и в любой диалектный словник или словарь дифференциального типа, включено большое количество общерусских слов, т. к. эти слова часто отличаются по своей семантике от аналогов в литературном языке.

При унификации индексов по окончаниям могли оказаться ошибочными указания на переходность или вид глагола (совершенный — несовершенный). Но, во-первых, вид и переходность для большей части словоизменительных таблиц не имеет значения; во-вторых, в большом количестве случаев вид диалектного глагола по словнику определить невозможно — необходимо обращаться к контексту.

От учета диалектной грамматики анализатором пока пришлось отказаться, т. к. во многих случаях это бы резко увеличивало омонимию и ничуть не уменьшало количество ручной обработки.

Например, в северных говорах у существительных встречается вариативность флексий *е/и* в ед. ч. Д.–П. I скл. (*к козы́, в Москвы́*), флексий *е/у/и* в ед. ч. П. п. II скл. (*в ле́се, в городу́, на кони́*), расширена сфера употребления флексии *-у* в ед. ч. Р.п. II скл. (*около го́роду, у мужику́)*, 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 пече́)* и т. д. Некоторые словоформы оказалось возможным задать анализатору «списком» (*бра́товья́, бра́ты́, бра́теи* в И. мн.; *бра́тьёв, бра́тове́й, братовьёв, братовье́й, бра́те́й, бра́тье́й, бра́тее́й; сыновье́й, сыновьёв* в Р.-В. мн. и проч.). Списком даны словоформы личных и некоторых других местоимений.

Расширять Программы грамматического анализа, учитывающего омонимичные словоформы, на данном этапе мы отказались, т. к. основной нашей задачей пока является восстановление начальной формы слова по ее грамматическим вариантам. В этом случае варианты в системах, где Р. = Д. = П. (*из Москвы́, к Москвы́, в Москвы́*), дадут нам одну и ту же лемму (*Москва*). К случаям, когда происходит ориентация 3 скл. на 1-е (*на пече́*), когда для всех слов II скл. оказываются возможны Р-2 (*табака́* и *табаку́*) или П-2 (*в ле́се* и *в лесу́*), а также П-3 (*на кони́*), можно будет вернуться позже, учитывая конкретную частотность встретившихся в Базе словоформ. То же касается лексем с широким варьированием грамматических форм по говорам (слов типа *брат, сын*): готовые таблицы их словоизменения созданы и адаптированы уже не к Словнику, а к Электронному Корпусу.

Несмотря на высокую частотность и грамматическую подвижность некоторых лексем, подавляющее большинство слов, составляющих Словник АОС, реально зафиксировано в 1-3 контекстах. В дальнейшей работе над Грамматическим Словарем индексы будут постоянно уточняться — как в работе с контекстами из «бумажной» картотеки, так и из Электронной.

6. Для удобства работы была создана **программа перекодировки серии шрифтов** АОС, созданных специально для нужд Словаря еще в 90-е годы (Times New Roman AOS), в шрифты уникод — современные шрифты, позволяющие работать с диакритиками. Для этого в среде WW были созданы макросы: а) макрос по переброске шрифтов АОС в уникоды; б) макрос, необходимый для промежуточной обработки полевых экспедиционных тетрадей перед проверкой и помещением их в базу (где знак * заменяется на знак стандартного ударения, а также производятся еще некоторые замены).

Был создан **модуль для импорта полевых тетрадей в dbf**. Теперь процедура обработки полевой экспедиционной тетради в 90 страниц, подготовленной для передачи в Корпус, полностью автоматизирована и занимает ок. 1 минуты: цельный текст (файл rtf) разбивается на поля (в т. ч. поля, содержащие сведения: «фонетическая диалектная словоформа», «восстановленная начальная форма», «пример», «адрес записи», «год записи», «автор записи», «информант», «примечания» и нек. другие) и выстраивается по алфавиту словоформ в поле lex («начальная форма слова»).

7. В дальнейшем предполагается создать новый **Автоматический анализатор** — серию программ, которые позволят «отлавливать» нестандартные («окказиональные») грамматические (и фонетические) словоформы и новые слова в постоянно пополняющемся Корпусе АОС. Анализатор будет сравнивать эту словоформу с уже имеющимися — гипотетически построенными — словоформами, созданными на основе Словника АОС. По своим результатам программа должна напоминать систему для проверки правописания типа Spell Checker или «Орфо»: (а) она должна находить слова и словоформы, отсутствующие в Словнике АОС (и в гипотетических таблицах, построенных на основе Словника), (б) исходя из анализа «ошибки», предлагать пользователю возможность выбрать вместо фонетической словоформы орфографически «правильный» вариант написания (*поцанки* > *по́дсанки* или *паца́нки; ця́ица* > *ча́ица* или *ча́яться; ш* > предлог *с* или частицу *ж*); (в) отмечать и предъявлять новую словоформу: давать возможность пользователю пополнять словарь словоформ и слов, предлагая новую таблицу гипотетических словоформ и в дальнейшем учитывая эти новые сведения.