Использование лексико-грамматических баз данных в русской диалектной лексикографии[1]

The use of lexico-grammatical databases in the Russian dialectal lexicography

Тер-Аванесова А. В. (teravan@mail-ru)
Институт русского языка им. В. В. Виноградова РАН

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

С помощью СУБД STARLING обогащена построенная ранее лексико-грамматическая база данных (ЛГБД) по русским народным говорам с различением двух фонем «типа *о*». К созданной ранее базе данных по среднерусскому говору с. Пустоша Шатурского р-на Московской обл. добавлена ЛГБД по вологодскому слободскому говору, включающая ок. 30 тыс. словоформ, представляющих ок. 4500 лексем. Ядерный диалектный корпус (ЯДК) содержит тексты с частичной лексико-грамматической разметкой. В сентенциальной базе единицами являются предложения ЯДК в фонологической транскрипции, пронумерованные в порядке вхождения в ЯДК. На ее основе создан прямой алфавитный лексико-грамматический конкорданс и обратный алфавитный лексико-грамматический указатель словоформ. ЛГБД содержит информацию об условной фонологической транскрипции данной единицы, о словоизменительных и акцентных типах лексем, смысловые пометы о лексических значениях семантических диалектизмов, а также метаязыковые социолингвистические пометы о возрастных и территориальных особенностях употребления словоформы.

1. Предмет исследования и материал: русские народные говоры с различением двух фонем «типа *о*»

В рамках проекта РГНФ в 2006 г. было продолжено создание ЛГБД по русским народным говорам с различением двух фонем «типа *о*». К созданной ранее базе данных среднерусского (владимирско-поволжского) говора с. Пустоша Шатурского р-на Московской обл. добавилась построенная формате STARLING лексико-грамматическая база данных по севернорусскому слободскому говору деревень Арзубиха, Захариха и Злобиха Харовского р-на Вологодской области.

Предметом исследования являются русские говоры, в системе вокализма которых представлены две фонемы «типа *о*», распределенные в соответствии с правилом Л.Л. Васильева — А.А. Шахматова: «*о* закрытое» (фонема /уо/) выступает на месте **о* под праславянским «восходящим» ударением, «*о* открытое» (фонема /о/) — на месте **о* под «нисходящим» ударением, на месте **ъ, *е, *ь*. В говорах с различением двух фонем «типа *о*» обычно также различаются две фонемы «типа *е*», наряду с фонемами /а/, /у/, /и/, в связи с чем их системы вокализма получили название семифонемных. В настоящее время такие говоры достаточно редки, не образуют сплошных ареалов, сохраняются главным образом в восточной части Европейской территории России и лишь отдельными вкраплениями — к югу и юго-западу от Москвы. Данные русских говоров с семифонемным вокализмом имеют особое значение для истории русского языка и славянской акцентологии, поскольку тембр ударного *о* < **о* является (по крайней мере, в части случаев) отражением праславянских слоговых тонов.

Некоторые косвенные данные свидетельствуют о том, что в прошлом системы вокализма расматриваемого типа были распространены в русских говорах гораздо шире. Ареалы таких систем должны были быть не меньше современных ареалов нескольких типов диссимилятивного яканья, пред-

[1] Данная работа выполнена при финансовой поддержке Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» (проекты «Восточнославянский диалектный корпус: праславянское наследие и лингвогеография» и «Генезис балто-славянской языковой общности: акцентологический аспект»), а также гранта РГНФ № 08-04-12132в.

полагающих семифонемный и шестифонемный ударный вокализм (обоянский, задонский, дмитриевский, новосёлковские, ореховские типы яканья). Карты и материалы Диалектологического атласа русского языка показывают, что небольшие кружевные ареалы юго-восточных семифонемных систем вокализма «вписаны» в несравненно более обширные ареалы перечисленных типов диссимилятивного яканья. Следовательно, эти типы яканья являются косвенным свидетельством наличия в прошлом различения под ударением двух фонем «типа *о*» на гораздо большей территории, чем сегодня. Локализация памятников письменности XIV–XVII вв., графико-орфографические системы которых отражают противопоставление двух фонем «типа *о»,* показывает, что говоры с различением двух *о* в старорусский период были распространены почти на всей территории русского языка.

В 2006 г. были проведены три диалектологические экспедиции для сбора материала по говорам изучаемого типа: в с. Новосёлки Рыбновского р-на Рязанской обл., с. Пустоша Шатурского р-на Московской обл. и в дд. Арзубиха, Захариха и Злобиха Харовского р-на Вологодской обл. Расшифровки магнитофонных записей речи уроженцев названных вологодских деревень стали материалом для ЛГБД.

Несколько говоров рассматриваемого типа были предварительно сопоставлены, в том числе с помощью построенных ЛГБД, как в отношении акцентных систем и распределения двух фонем «типа *о*», так и в отношении их лексического состава: западно-вологодские слободской и тотемский (последний — по описанию О. Брока); владимирско-поволжский говор с. Пустоша; восточный среднерусский акающий говор с. Лека Шатурского р-на Московской обл. (по описанию А.А. Шахматова); задонский говор (по материалам В. Тростянского); рязанский говора с. Новосёлки.

Говоры с противопоставлением двух *о* обнаруживают сильные различия по общим наборам признаков (они относятся к разным наречиям и группам говоров). Одновременно, относясь к восточной диалектной зоне, все говоры с двумя *о* имеют ряд важных общих черт: «моновариантное» склонение типа рус. лит., маргинальную подвижность ударения в прош. времени глаголов с корнями на нешумные, в целом схожее распределение непроизводных существительных по акцентным типам и ряд других сходств. Так, все семифонемные говоры обнаруживают нетривиальное сходство: сохранение у небольшого числа существительных муж. рода (**u-*, **i-* и консонантные основы а. п. *d*) рефлекса смешанной акцентной парадигмы (с формой-энклиноменом в И.ед. и окситонезой прочих форм). Обнаружены признаки, противопоставляющие друг другу отдельные группы говоров с различением двух фонем *о,* например, 1) /уо/ из **о* в формах мн. числа слов ж. и ср. рода а. п. *b* (*вдуóвы, вдуóвами; долуóты, долуóтами*) в средне- и южнорусских говорах рассматриваемого типа; в севернорусских в тех же случаях — /о/; 2) накоренное ударение в наст. времени и пов. наклонении *i*-глаголов а. п. *b* при насуффиксальном — в инфинитиве и прош. времени, характерное для говоров Рязанской группы и «рязанского ареала» говоров с различением двух *о* (Пустоша *хóжу, хуóдишь,* Новосёлки *хуóжу, хуóдишь*). Списки глаголов с указанной инновацией, однако, сильно различаются в говорах Пустошей и Новосёлок: если в Пустошах этот список ограничен итеративами а. п. *b1* (*ходить, носить, возить, водить, молотить, просить* и т. д.), то в Новосёлках в него входят каузативы и деноминативы а. п. *b2* и даже *c.* Последнее различие должно указывать на гетерогенный характер говоров с различением двух фонем «типа *о*» в «рязанском ареале».

Построение лексико-грамматической базы данных слободского говора (Харовский р-н Вологодской обл.). Аудиозаписи речи носителей говора старшего поколения были расшифрованы и записаны в аллофонемной транскрипции. На основе получившихся текстов с помощью интегрированной информационной среды STARLING (автор — чл.-корр. РАН С. А. Старостин) построена лексико-грамматическая база данных говора — так называемый ядерный диалектный корпус (ЯДК). ЯДК представляет собой исчерпывающее описание говора в рамках определенного корпуса текстов и охватывает тексты общей длиной около 30 тыс. речевых словоформ. Они репрезентируют 7047 языковых словоформ без учёта пунктуации, 9591 пунктуационно-грамматическую словоформу (пунктуационный вариант языковой словоформы).

2. Структура базы данных слободского говора

Структура базы данных слободского говора идентична структуре созданной ранее базы данных говора с. Пустоша Шатурского р-на Московской обл. В качестве исходной базы данных выступает ЯДК. Лингвистическая информация в ЯДК организована по многоступенчатому принципу. Выделяется 7 уровней членения письменного текста; на каждом из них выделяется своя основная (базовая) единица членения; каждой единице членения каждого уровня в ЯДК приписан уникальный номер, способный служить адресом отсылки к этой единице.

1. Уровень целого текста. На этом уровне вводятся параметры, характеризующие личность информанта: фамилия, имя, отчество, год и место рождения, образование и т.п.

2. Уровень абзаца (сверхфразового единства). Сверхфразовое единство — это отрезок текста, пун-

ктуационно выделенный особым абзацным делимитатором («красной строкой», «отступом»). У сверхфразового единства есть некоторая единая общая смысловая тема.

3. Уровень предложения (сентенциальный уровень). Границы предложений помечены сентенциальными делимитаторами. В начале предложения стоит инициальный делимитатор — суперсегментная пунктограмма «заглавности»; в конце предложения стоит финальный делимитатор — пунктограммы «.», «?», «!», «…». Содержательно предложение соответствует законченной мысли, а фонетически — интонационно законченному отрезку.

4. Уровень клаузы // предикации (клаузальный уровень). Границы клауз помечались так: предложение состоит из клауз, а между клаузами внутри предложения стоит один из клаузальных делимитаторов. К ним относятся пунктограммы «;», «:» и «–». Содержательно и интонационно клаузы примерно соответствуют простым предложениям и отдельным предикациям в составе сложных предложений.

5. Уровень синтагмы (синтагматический уровень). Границы синтагм внутри клаузы помечены пунктуационным синтагматическим делимитатором — пунктограммой «запятая». Содержательно и интонационно синтагмы примерно соответствуют словосочетаниям.

6. Уровень такта (тактовый уровень). Границы такта в ходе расшифровки помечались стандартной орфографической пунктограммой «пробел», но после создания ЯДК они были размечены вручную так: был использован пунктуационный тактовый делимитатор — пунктограмма «знаменательный (паузальный) пробел». Такты примерно соответствуют фонетическим словам, членам предложения, «синтаксическим молекулам», формам слова (как аналитическим, так и синтетическим). Важнейшее фонетическое свойство такта: внутри него невозможна (или по меньшей мере нетипична) пауза.

7. Уровень глосса (глоссовый уровень). Границы глоссов в ходе расшифровки помечались либо стандартной орфографической пунктограммой «пробел», либо стандартной орфографической пунктограммой «дефис», но после создания ЯДК их границы были размечены вручную. Каждый такт состоит из одного или нескольких глоссов. Глоссы, входящие в состав одного такта, обладают признаком потенциальной подвижности в предложении. Для обозначения границ глоссов при разметке был использован специальный набор нескольких метаязыковых глоссовых делимитаторов — пунктограммы «служебных пробелов». Выделены служебные пробелы шести типов: «{» между проклитикой и её правой опорой; «}» между энклитикой и её левой опорой; «<» между проклитикоидом и его правой опорой; «>{» между энклитикоидом и его левой опорой; «<>» между членами квази-композита с неустойчивым просодическим центром; «&» между компонентами «фразеологического штампа» с множеством просодических центров. Глоссы примерно соответствуют по длине морфологическим словам (в т. ч. служебным словам, синтетическим формам слов и подвижным компонентам аналитических форм). Внутри глосса (так же как внутри такта) невозможна пауза. Фактически наиболее близкий аналог глоссов в русском письменном тексте, записанном по правилам русской орфографии — это графические слова.

Ценность предложенной многоуровневой схемы ЯДК состоит в том, что при необходимости вывести на обозрение список отрезков текста, обладающих некоторым общим свойством, STARLING позволит пользователю по выбору вывести (на экран, на принтер или в файл) отрезок не только одного формата, но разных форматов — графическую словоформу (глосс), минимальный контекст этой словоформы (аналитическую форму, например, предложно-падежную, сочетание клитики с акцентно автономной словоформой и т. п. — такт), словосочетание (синтагму), предикацию (клаузу), предложение, абзац.

Лингвистическая информация о единицах текста на данном этапе в ЯДК такова: 1. Условная фонологическая транскрипция данной единицы (в сочетании с её пунктуационной разметкой). 2. Словоизменительный и акцентный тип данной единицы. 3. Смысловые пометы (при лексических диалектизмах). 4. Метаязыковые социолингвистические пометы о возрастных и территориальных особенностях употребления словоформы.

3. Приложение. Образцы словарных статей (иллюстрирующих семантическое поле «позвоночник» в харовском говоре)

Лён 1 ‘шейный отдел позвоночника’ <а.т. В>.

Вот <nom. sg.> *лён етот са́мой у ч'ело-вие́ка. Голова́ с позвоно́ч'ником свя́зана, между ни́м* <nom. sg.> *лён. А здие́с го́рло. А ше́йя ето фсё вми́сте ше́йа и йе́с. А у шшу́ките ние́ту* <gen. sg.> *лну́-то е́тово. Како́й у шшу́ки* <nom. sg.> *лён. Ние́ту* <gen. sg.> *лну́ у шшу́ки.* <Егоров Виктор Никол. 1940 г. р. Род. в д. Злобиха Харовск. р-н, Волог. обл. 7 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>.

Осёл 1 ‘шестой шейный позвонок’ <а.т. А>.

<nom.-acc. sg.> *осе́w,* <gen. sg.> *осе́ла* <Клешнина Нина Васил. 1936 г. р. Род. в д. Арзубиха Харовск. р-н, Волог. обл. 7 кл. Зап. Тер-Аванесова в д. Арзубиха, Харовск. р-н, Волог. обл., 2002 г.>.

<nom. sg.> *Осе́w ето хря́шш о́коло хрепта́, о́коло позвоно́ч'ника, пе́рва-та шы́шка. Ешшо́ до позвоно́ч'ника не дошлẃ, и ше́йа конц'я́ец'ц'е —*

ето <nom. sg.> *осёw.* <nom. sg.> *Осёw болúт, горúт, как ц'úрей рвёт, как пересúлиш себя́, етот* <nom. sg.>*осёw.* <Егорова (Фокина) Зинаида Никол. 1941 г. р. Род. в д. Полутиха Харовск. р-н, Волог. обл. 8 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>.

Хребёт 'хребет, позвоночник без шейного отдела' <а.т. В>.

<nom.-acc. sg.>*хребёт,* <gen. sg.> *хрептá,* <instr. sg.>*хрептóм,* <loc. sg.>*на хрептué,* <nom. pl.> *хрепты́* <Клешнина Нина Васил. …>; <gen. sg.> *óколо хрептá* <Егорова (Фокина) Зинаида Никол. …>.

Хрип 1 'соединение позвоночника и черепа у рыб' <а.т. А sg.>.

Йа сломúw шшýки <acc. sg.> *хрúп, ет тшóлко схрýпало. Фсё, онá уш готшóва. Ф сиétку попадёт, ну кáк йейóзадавúт? гóлову рáс — фсё. В étом миéсте у нейó слáбойо éто миéсто-то. А болшýйу кáк, никáк немóжно, мнué прóшлой гóт попáла, повезлшó — йедвá* <acc. sg.> *хрúп сломúw. Шшýчина болшáйа, ак* <acc. sg.> *хрúп йедвá сломúл. Ниétу лнý у шшýки. Йёсли бы бы́w у шшýки лён, шшýки тáг бы* <acc. sg.> *хрúп не сломúт. Уш* <gen. sg.> *хрú па не сломúт.* <Егоров Виктор Никол. 1940 г. р. Род. в д. Злобиха Харовск. р-н, Волог. обл. 7 кл. Зап. Белова, Тер-Аванесова в д. Злобиха, Харовск. р-н, Волог. обл., 2003 г.>

4. Приложение. Полные синонимы в слободском говоре (на материале существительных)

При помощи ЛГБД легко показать, что говор деревень Арзубиха, Захариха и Злобиха, а также других деревень бывш. Слободского с/с Харовского р-на Вологодской обл., является единым. Наблюдаемые различия отчасти объясняются, по свидетельству информантов, как относящиеся к «младшей» или «старшей» разновидностям говора, причем соответствующие единицы «младшей» разновидности, как правило, заимствованы из литературного языка. Имеются и такие различия внутри говора, которые представляют собой внутрисистемные колебания. Крайне редки различия, которые могут претендовать на принадлежность к разным диалектным системам (например, название шеста, вокруг которого укладывают сено в стог: Арзубиха, Захариха *стожаир* или *стогаир,* Злобиха *островиина;* название помещения над избой, чердака: Арзубиха *иизбиця,* Митиха *потолшика*).

Ниже приводятся пары существительных — полных синонимов, выявленных в словаре слободского говора при помощи ЛГБД. Первый из пары синонимов является элементом традиционного лексического состава говора (иногда даже это — устаревшее слово); второй, как правило, представляет собой заимствование из литературного языка. Этот список выделен из словаря существительных, включающего около 2000 лексем; тем самым, примерно десятая часть словаря существительных говора представляет собой пары полных синонимов.

Пары полных синонимов (в большинстве своих случаев обязанных факту диалектно-литературного двуязычия) выделялись на основе явных показаний информантов («можно так сказать, а можно и так сказать»). Хронологические различия между членами пар («старое» / «новое») отмечались также на основании показаний информантов («сейчас говорят так-то, а раньше говорили так-то»).

бáба — жонá, бáба — жéншына, батуóк — пáwка, берéмё — охáпка, блюдо — мúска, бог — икóна, божáт — хрёсной, божáтка — хрёснамáт, брусníця — брусníка, брюшы́на, брюхо — жывуóт, вар — пиéна, вéред — нары́w, вúця — виéтка, вуóлиха ~ óлха 'ольха', *вówненця — вownýшка, вóтен — лентя́й, грýда — кýчя* (предметов), *губá — подборуóдок, губá — чя́га, губá — погáнка, двойнúк — близнéць, долуóн — ток* (в гумне), *кáтаник — вáленок, колобáшка — лепёшка, колуóда — корыто, ком — ломóт* (хлеба), *косúця — висóк, кошýля — шýба* 'шуба, крытая сукном', *кут — кýхня, куфты́р — жывуóт, избá — куóмната, úзбиця — чердáк, йиéжа — йедá, зáгорода — úзгород, нéпогодь — бýря, лабазá — лесá* (строительные), *лáва — мóстик, лáwка — магазúн, лёжен — лентя́й, лён — шéя, лóшат — куóн, ля́га — топ, ледúна — болуóто, мётлá — вúник, миéw — заквáска* (из пивного сусла), *мизгúр, мызгы́р — паýк, мост — пow, мостúна — половúця, наберёг — заквáска* (из остатков ржаного теста), *назём — навуóз, настáwница* устар. *— учúтелница, оболочúна — тýчя, оболочúнка — óблако, обýтка — óбуw, острежнúк — стрехá, отерёбок — замухры́шка, отченáш — молúтва, плёнка — пузы́р* 'околоплодный пузырь', *поскуóтина, поскуóтка — пáзбишшо, поганéць — погáнка, постéлкя — послиéд, потолуóка — чердáк, простокúша — простоквáша, рúзен — кусóк* (хлеба), *роднúк — колуóдець, ры́ло — нуóсик* (чайника), *слúзен — улúтка, собáка — пёс, стекля́шка, скля́нка — бáнка, буты́шка, солодя́ха — сыройиéшка, соромотá — стыд, сиéра — смолá, стáя — хлеw, тоскá — бол* 'боль', *úлик — úлей, фатá — платóк, хребёт — позвонуóчник* (позвоночник без шейного отдела, ср. *лён*), *хрестóк — хря́шшык, черёд — óчеред, чéрен — рýчка* (лопаты, вил), *чилúк — поберёзоватик, чиря́к — чúрей, шáм — мýсор, сóр, шы́мора — прохиндéй, шубня́к — шýба* 'шуба мехом вверх'.

Литература

1. *Брок 1907* — О. Брок. Описание одного говора из юго-западной части Тотемского уезда // Сборник ОРЯС, 1907. — Т. 83.
2. *Васильев 1929* — Л. Л. Васильев. О значении каморы в некоторых древнерусских памятниках XVI–XVII вв. К вопросу о произношении звука о в великорусском наречии. Л., 1929.
3. *Зализняк 1985* — А. А. Зализняк. От праславянской акцентуации к русской. М., 1985.
4. *Крылов С. А.* Измерение частотности синтаксических молекул (на материале Генерального корпуса русского языка) // Кибрик А. Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам международной конференции «Диалог'2008» (Бекасово, 4–8 июня 2008 г.), М.: РГГУ, 2008а. С. 254–261.
5. *Крылов С. А.* О частотном словаре фонетических слов (на материале Генерального корпуса русского языка) // Архипов А. В. и др. (ред.). Фонетика и нефонетика. К 70-летию Сандро В. Кодзасова. М.: Языки славянских культур, 2008б, с. 387–399.
6. *Крылов С. А., Тер-Аванесова А. В.* Лексико-грамматические базы данных как инструмент диалектологического описания // Труды международной конференции «Диалог 2006». М., 2006. С. 493–497.
7. *Ter-Avanesova A.* Russian dialects with the distinction of two o-phonemes and their contribution to Slavonic accentology (Русские говоры с различением двух о-фонем и их значение для славянской акцентологии) // Second International Workshop on Balto-Slavic Accentology. Copenhagen, 2006. P. 20–24.
8. *Тер-Аванесова А. В.* Акцентуационные особенности русских говоров с различением двух фонем «типа о» // Тезисы докладов Международной конференции «Актуальные проблемы русской диалектологии» 23–25 октября 2006 г. М., 2006. С. 177–180.