ЛЕКСИКО-ГРАММАТИЧЕСКИЕ БАЗЫ ДАННЫХ КАК ИНСТРУМЕНТ ДИАЛЕКТОЛОГИЧЕСКОГО ОПИСАНИЯ

LEXICO-GRAMMATICAL DATABASES AS A TOOL OF DIALECTOLOGICAL DESCRIPTION

А. В. Тер-Аванесова *(teravan@mail.ru)*

Институт русского языка РАН

С. А. Крылов *(krylov-58@mail.ru)*

Институт востоковедения РАН

В среде STARLING создана лексико-грамматическая база данных (ЛГБД) по говору с. Пустоша Шатурского р-на Московской обл., длиной 30 000 словоформ. Ядерный диалектный корпус (ЯДК) со сплошной лексико-грамматической разметкой (лемматизацией) стал основой для создания ряда производных БД (указателей).

0. Использование интегрированной информационной среды STARLing для решения проблем диалектологии русского языка.

В 2003–2005 гг. в рамках работы по проекту РФФИ авторами подготовлено описание восточного среднерусского говора села Пустоша Шатурского р-на Московской области. Монография содержит очерк фонетико-фонологической, морфологической и акцентуационной систем говора, заметки о его синтаксисе и лексике, словарь говора, включающий около 3000 лексем с грамматическими пометами и примерами контекстного употребления, и тексты – образцы речи носителей говора. Описание говора представлено в также форме лексико-грамматической базы данных, построенной в формате STARLing.

1. Говор с. Пустоша: его лингвистические характеристики, внешние и внутренние

1.1. Говор с. Пустоша в контексте проблем славянской диалектологии

Село Пустоша[1] (в XIX в. – Гридино) расположено в 180 км к востоку от Москвы, в 5 км от ж/д станции Черусти. Говор Пустошей имеет ярко выраженные черты, свойственные владимирско-поволжской диалектной группе и особенно муромским говорам. Наряду с этим говор обладает совокупностью особенностей, общих у него с так называемыми “восточными среднерусскими акающими говорами”, небольшой ареал которых расположен примерно в 40 км к югу от Пустошей. Часть этих особенностей встречается и южнее, в отдельных юго-восточных русских говорах (рязанских, липецких, белгородских).

В историческом и типологическом плане наибольший интерес представляют система семифонемного ударного вокализма говора и отсутствие смягчения согласных перед гласными переднего ряда *е* и *и* (вследствие их отвердения, подобного украинскому). В говоре Пустошей под ударением в области среднего подъема попарно противопоставлены фонемы /е/ (“открытое *е*”, из этимологических **е, *ь*) и /ие/ (“закрытое *е*”, из *ять*), /о/ (“открытое *о*”) и /уо/ (“закрытое *о*”). Распределение двух фонем “типа *о*” в говоре Пустошей подчиняется правилу, установленному для великорусских говоров Л.Л. Васильевым и А.А. Шахматовым, а именно: фонема /о/ представлена на месте **о* под праславянским “нисходящим” ударением и на месте **ъ*, фонема /уо/ – на месте **о* под праславянским “восходящим” ударением. С синхронной точки зрения огласовка ударного слога непроизводных основ и флексий, содержащих *о,* является традиционной.

В описании говора Пустошей распределение двух фонем “типа *о*” рассматривается в связи с акцентной системой говора на большом материале, собранном при помощи специальных вопросников и фактически целиком включающем “праславянский фонд” непроизводных именных и глагольных основ говора. Материал говора Пустошей в целом показывает распределение двух фонем “типа *о*” в соответствии с правилом Л.Л. Васильева – А.А. Шахматова; однако имеются отклонения от этого правила. Проблему представляют закономерности появления фонем “типа *о*” в словоформах, ударение которых не восходит к праславянскому, а также в заимствованиях.

1.2. Источники данных и материал исследования

Материал по говору Пустошей собран в 2001–2005 гг. А.В. Тер-Аванесовой. В описании говора использованы расшифровки магнитофонных записей речи носителей говора и записи из полевых дневников. Информантами являются 70–90-летние женщины; более молодые коренные жители села сохраняют традиционный говор несравненно хуже. Однако даже у лиц старшей возрастной группы наблюдаются языковые различия, которые можно квалифицировать как архаичную "старшую" (у 80–90-летних) и "младшую" (у 70-летних) разновидности говора. Записи представляют (1) связную речь – в основном монологи, реже диалоги; по содержанию монологи представляют собой рассказы на бытовые темы, воспоминания о прошлом, разъяснения об ушедших в прошлое явлениях традиционной материальной и духовной культуры, размышления о настоящем и будущем села; (2) ответы на вопросники, нацеленные на выявление особенностей фонетики, морфологии, акцентуации и лексики говора. Это могут быть отдельные словоформы (если контекст задан собирателем), словоформы в минимальных контекстах или отдельные фразы.

1.3. Фонологическая система говора с. Пустоша и её отражение в транскрипции текстов

Материал говора Пустошей приводится в условной записи, отражающей фонемный состав говора. В отдельных случаях в записи делается уступка фонетическому принципу написания: обозначение аллофона *ы,* позиционного смягчения согласных (inf. *несьть*), смягчения задненебных после мягких согласных фонем (nom. sg. *Ваᵢнькиа,* voc. *Ванькь*), смягчения задненебных после *и* (*стариᵢкь*).

Под ударением в говоре различается 7 гласных фонем: /а/, /е/, /ие/, /о/, /уо/, /у/, /и/. Дифтонги *иа, ио, иу* и трифтонг *иуо* являются бифонемными сочетаниями (*там-сиам, день, диет, нос, ниос, нуож, тиуоᵢплый, тут, тиук, тих, ты*). Безударный вокализм говора характеризуется так называемым неполным оканьем: (1) в 1-м предударном слоге различается 5 гласных фонем после твердых и после мягких согласных фонем: /а/, /е/, /о/, /и/ /у/ (*самаᵢ, пиатаᵢк, водаᵢ, виодуᵢ, несиᵢ, сыроᵢй, сидиᵢт, сухоᵢй, сиудаᵢ*), (2) в прочих безударных слогах, кроме конечных открытых слогов, различается 3 гласных фонемы: /и/, /у/, /ъ/ (*тишынаᵢ, питачоᵢк* 'пятачок'*, веᵢсилъ* 'весело', *хърошуоᵢ, пуоᵢмниу,* 1 sg. *ноᵢшу*).

Для системы консонантизма характерно наличие твердых, "полумягких" и мягких согласных. Фонетически мягкими являются только *й* и "долгие" *ш'ш', ж'ж'*; в записи они передаются как *й, щ, ж'ж'*. "Полумягкие" согласные представляют мягкие согласные фонемы, парные твердым, либо аллофоны парных твердых фонем в позициях смягчения. Фонетически имеется не менее двух степеней "полумягкости": "слабая" – перед гласным *е* и "сильная" – перед *и,* перед согласным и на конце слова. "Слабая" полумягкость у большинства согласных очень близка к артикуляции твердых согласных, которые в говоре являются невеляризованными; особое качество в позиции перед *е* имеет /л/, который является т. наз. "средним" (*Леᵢна, леᵢч*), ср. твердый *л* в *лаᵢпа, луᵢк, лоᵢжыт, луоᵢвит* и "сильный" полумягкий в *лиᵢпа, лиаᵢгу, муоᵢль, скуоᵢлько.* "Слабая полумягкость" в записях никак не обозначается, за исключением "слабой полумягкости" аллофона /л/. "Сильная полумягкость" обозначается буквой "ь" на конце слова и перед согласными и никак не обозначается перед гласным *и.* Твердость согласных специальным знаком не обозначается.

Подавляющее большинство неодносложных словоформ иметт знак ударения. Не имеют его (1) словоформы, ударение которых перенесено на предлог или частицу *не* (*иᵢс полиа, неᵢ пил*); (2) сравнительно редкие случаи безударности неодносложных словоформ, обусловленной фразовой интонацией. Односложные словоформы имеют знак ударения, если на них размещен фразовый акцент (словесное ударение односложных словоформ не обозначается).

В записи употребляются стандартные пунктуационные знаки. Имеется деление текстов на абзацы.

Образец текста:

Швецова Клавдия Васильевна (1922 г. р.)

Пръ коруоᵢф еᵢтъ ничоᵢ йа не знаᵢйу. О, поᵢрьтиут, да, не знаᵢйу каᵢк. Испоᵢрьцили, пъгодиᵢ, тоᵢ ли они йиеᵢзьдили, то ли коруоᵢву проᵢдали. Там, ф туоᵢм концыэᵢ хтоᵢ-тъ ръскаᵢзъвъл. Еᵢтъ быᵢла. Ну, йиеᵢзьдиут их воруоᵢжут. А вот эᵢта Клаᵢник. Онаᵢ вить мълодаᵢйа ище. И еᵢтъ наᵢда! Поᵢсли уш он мниеᵢ гъвориᵢл не раᵢс, а йаᵢ зъбываᵢйу, с неᵢй ръзговаᵢриву. Гъвориᵢт, ни ръзговаᵢривъй. Иᵢли, гъворит, фиᵢк кажыᵢ, в рукаᵢф. И йа фсиогдаᵢ хожуᵢ и пьрекщуᵢсь, фсегдаᵢ. Йаᵢ уш цепеᵢрь привыᵢкла. А тоᵢ забуᵢдиш, и фсиоᵢ. Онаᵢ фсиоᵢ вреᵢмиа с неᵢй дружыᵢла, а таᵢ колдуᵢньйа згриоᵢбна.

1.4. Грамматика говора с. Пустоша

Грамматика говора в целом очень близка русской литературной. В системе имени наряду с И., Р., Р1 (партитивом, последовательно выраженным флексией *-у* у баритонированных существительных м. рода с «вещественным» значением), Д., В., Т., М., П. имеются следующие падежные формы: (1) "родительный падеж с предлогом *у*" (Р2), выраженный окончанием -{ие} у существительных *а*-склонения, а у имен других классов совпадающий с Р. (*у вдовиеᵢ, у шкоᵢли; у доᵢма, у пеᵢчы*); (2) вокатив с нулевым окончанием и усеченной основой – только у имен собственных и названий старших родственников: *Вась, Клав, Оль, Гришк, ма, па, ба, лё*

(от *лёлик* ‘крестная; тетя’); (3) собирательная форма – главным образом у названий животных и одушевленных существительных с пейотративным значением: nom. sg. *лось, звиеиръ, бараин, воирон, пьянь, хулиган, вор,* nom. pl. *лоиси, бараины, воироны, пьяини, хулиганы, воиры*, coll. nom. sg. *лосьё, зверьё, вороньё, пьяньё, хулиганьё, ворьё,* coll. nom. pl. *лосьяи, зверьяи, бараньяи, вороньяи.* Формы coll. pl. могут употребляться как формы мн. числа.

К числу особенностей именного словоизменения говора относятся: (1) распространение окончаний твердой разновидности склонения после исконно мягких основ. В результате последовательно проведенного выравнивания мягкой разновидности склонения по твердой в окончаниях instr. sg. *а*-основ и gen. pl. *о*-основ после шипящих, *ц* и *j* появляется дифтонг *уо,* а после рефлексов **l’, *r’, *n’* - *иуо: душуоий, землиуоий, оццуоиф, молиуоиф, сыновйуоиф.* Как и в лит. языке, безударные окончания имен с мягкими и твердыми основами одинаковы (отсюда только в именных окончаниях возможно произношение [ъ] после мягких согласных и различение безударных рефлексов **а, *о, *е, *ять*): nom. sg. *баибъ, вуоил’ъ,* instr. sg. *баибъй, баин’ъй, ваилъм, воипл’ъм,* dat. pl. *стиеинъми, стаивн’ъми.*

(2) Широко представленное у непроизводных существительных м. рода с подвижным ударением и твердой основой ударное окончание nom.(acc.) pl. *-а,* являющееся яркой особенностью восточнорусских говоров. В Пустошах находим, помимо “литературных” форм с окончанием *-аи,* также *плодаи, лазаи, гробаи, садаи, бороваи, бродаи, брусаи, ястребаи, ящураи, мостаи, мозгаи, мохаи, полаи, разаи, сокаи, тяжаи, зобаи, родаи, горбаи, носаи, плугаи, струбаи.*

(3) Безударным окончанием существительных ср. рода и некоторых консонантных основ (исконно также ср. рода) является *-ы,* что является чертой средне- и южнорусской: *йаийцы, болуоиты, гуоивны, телиаиты, ребиаиты* – при ударном *-аи: полиаи, мориаи,* nom. pl. *имиаи, племиаи.*

(4) Как и в восточнорусских говорах, в говоре Пустошей несколько предикативных прилагательных сохраняет старую форму nom. pl. (m.): *сыити, раиди, виноваити.* Имеется неизменяемое предикативное прилагательное *похуоиж,* ср. *Онаи фсиаи на Пеитиу похуоиш; Рибятишки на тиаи похуоиш* (зафиксировано только у носителей архаической разновидности говора).

(5) Окончание loc. pl. существительных *-аф (в еитих большыих домаиф)*.

(6) В говоре имеется склоняемая постпозитивная частица *-т-* (всегда безударная), однако ее изменяемость сохраняют лишь 80–90-летние носители говора. Более распространена неизменяемая частица *-тъ*. Ср. nom. sg. m. *доим-ът,* dat. sg. m.-n. *доиму-ту*, acc. sg. f. *шкоилу-ту,* прочие формы – *-тъ*.

К числу особенностей морфологии и словоизменения глагола относятся следующие. (1) Способность частицы *бы (б)* выражать, помимо сослагательного наклонения, значения уступки, желательности, неуверенности, сочетаясь как с *л*-причастием и инфинитивом, так и с формами наст.-буд. времени и повел. наклонения: (а) *Хоруоишый бы паирень, глядиеить хоруоишый, йа туоильки чеирис йавуои жылаи;* (б) *А онииеи туидь жы поимерьли, вот йа гъворюи: выходии п за ниих заимуш да хъронии на своии деиньги!* (в) *Взошлии ф Черьниаитин, а оин не знаил бы, кудаи пойдиить;*

(2) Наличие особой формы сослагательного наклонения, образованной сочетанием неизменяемой связки *был* с формой прошедшего времени и обозначающей действие в прошлом, результат которого ликвидирован к настоящему моменту или в настоящий момент не востребован. *Я уш и книишки был купиила* (но сына не приняли в школу). *Онаи такайа не оичинь, был пъступиила в буфеити робуоитать, хотиеила пръдафцоим – диеилъ ни пошлуои.* В отдельных редких примерах эта форма имеет значение ‘чуть не сделал’, ср. *Мениаи рас тоикъм был убиила в завуоиди.*

(3) В архаическом слое говора употребляются деепричастия на *-я* и на *–(м)шы* и причастия на *-н-/-т-*. Первые употребляются в роли сказуемого, обозначая состояние или свойство субъекта действия (*Он после суиток, не спаимшы. Йаи нъ любуоий рабуоити привыикшы.*), или в роли обстоятельства причины, образа действия (*Он надойиеиист тебе ходиаи* (о бегучей собаке, которую приходится часто разыскивать). *Он сидиеил цыеилый деинь таигжа, не ръзговаиривамшы*). Причастия на *-н-/-т-* также употребляются в роли сказуемого (образуясь как от переходных, так и от непереходных глаголов), обозначая результат действия или состояние. *Там уже фсиеи домаи сожгоины. Им такиийе деиньги даидены! Тепеирь там фсиои згориеита, фсиаи згориеила клиуиква. В войнуи Рошаиль был веись задымлиоиный. Там завуоит, поирах диеилъли, – он уибраный од гоирада.*

К числу особенностей глагольного словоизменения относится совпадение безударных окончаний 3 pl. наст. времени I и II спряжений в *-ут: хуоидиут, плуоитиут, криичут, суоилиут, куоилиут,* но *тащаит, дойаит, пойуит.*

2. Структура лексико-грамматической базы данных говора Пустошей

2.0. Ядерный диалектный корпус (ЯДК)

С помощью интегрированной информационной среды STARLING (автор – чл.-корр. РАН С. А. Старостин) построена лексико-грамматическая база данных говора Пустошей – ядерный диалектный корпус (ЯДК). ЯДК – это исчерпывающее описание говора в рамках определенного корпуса текстов; он мыслится как естественная часть Генерального корпуса русского языка (ГКРЯ), созданного и ведущегося в формате STARLING с помощью СУБД STARLING. ГКРЯ составляет органическую часть Машинного фонда русского языка.

ЯДК охватывает тексты общей длиной около 30 тыс. вхождений словоформ (глоссов). Они репрезентируют около 7000 грамматических аллоглоссем (словоформ без учёта пунктуации), около 9600 пунктуационно-грамматических аллоглоссем (пунктуационных вариантов словоформы) и около 4000 лексем. Проведена полная морфологическая разметка ЯДК.

2.1. Уровни членения текста

Лингвистическая информация в ЯДК организована по многоступенчатому принципу. Выделяется 7 уровней членения письменного текста; у каждого из них -- своя основная (базовая) единица членения.

1.Уровень целого текста. Здесь задаются личные параметры информанта: фамилия, имя, отчество, год рождения.

2. Уровень абзаца (сверхфразового единства, СФЕ). СФЕ – это отрезок текста, выделенный абзацным делимитатором (“красной строкой”, “отступом”).

3. Уровень предложения (сентенциальный). Сентенциальные делимитаторы: инициальный – “заглавность”; финальные – “.” (“точка”), “?” (“вопросительный знак”), “!” (“восклицательный знак”), “…” (“многоточие”).

4. Уровень клаузы //предикации (клаузальный). Предложение состоит из клауз, а между ними стоят клаузальные делимитаторы: “;” (“точка с запятой”), “:” (“двоеточие”) и “–” (“тире”).

5. Уровень синтагмы (синтагматический). Клауза состоит из синтагм. Синтагмы отделяются синтагматическим делимитатором – “запятой”.

6. Уровень такта (тактовый). При расшифровке такты отделялись “пробелами”, но в ЯДК их границы размечены особо. Синтагма состоит из тактов. Для обозначения границ тактов («потенциальных пауз») использован особый тактовый делимитатор – “знаменательный (паузальный) пробел”. Такты – это, грубо говоря, фонетические слова. Внутри такта невозможна (или хотя бы нетипична) пауза.

7. Уровень глосса (глоссовый). Границы глоссов при расшифровке помечались “пробелами” и “дефисами”, но в ЯДК их границы размечены особо. Такт состоит из глоссов. Глоссы обладают признаком потенциальной подвижности в предложении. Для обозначения границ глоссов при разметке использован набор метаязыковых глоссовых делимитаторов – “служебных пробелов”. Их выделено шесть 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 – это, грубо говоря, морфологические слова (в т. ч. служебные слова, синтетические формы слов и подвижные компоненты аналитических форм).

Если надо вывести на обозрение список отрезков текста, обладающих некоторым общим свойством, STARLING позволяет по желанию получить отрезок не одного формата, но разных: графическую словоформу, ее минимальный контекст (аналитическую форму, например, предложно-падежную, сочетание клитики с акцентно автономной словоформой и т. п.), словосочетание, предикацию, предложение, абзац.

2.2. Лингвистическая информация о текстовых словоформах в ЯДК

1. Условная фонологическая транскрипция.
2. Морфологический разбор, включающий:
2А. Указание глоссемы, репрезентируемой словом.
2Б. Указание грамматемы (грамматической формы) -- сочетания граммем, выражаемого словом.
3. Классифицирующая характеристика глоссемы:
3А. Акцентный тип (АТ).
3А’. Диалектная специфика АТ (его отклонения от АТ литературного аналога).
3Б. Общекатегориальный тип (ОКТ) (часть речи).
3Б’. Диалектная специфика ОКТ (отклонения от ОКТ литературного аналога).
3В. Флексионный тип (ФТ) (т.е. деклинационный или конъюгационный тип).
3В’. Диалектная специфика ФТ (отклонения от ФТ литературного аналога).
3Г. Смысловые пометы (при лексических диалектизмах).
4. Метаречевые социолингвистические пометы о возрастных и территориальных особенностях употребления словоформы.

3. Технологическая цепочка работы над ЯДК

Работа над ЯДК складывается из нескольких звеньев, образующих особую «технологическую цепочку» (ТЦ). Эта ТЦ состоит из следующих звеньев.

0.0. Полевая работа. В полевых условиях устный текст («дискурс») фиксируется магнитофонной записью. Этот этап – «докомпьютерное» («предкомпьютерное») звено ТЦ.

0.1. Собственно компьютеризация. Магнитофонные записи всех текстов расшифровываются и набираются на компьютере (вручную) в среде Winword в формате файлов текстовых документов.

0.2. Компьютерное предредактирование. В среде Winword тексты приводятся к стандартному виду, пригодному для дальнейшего конвертирования. Практически предредактирование состоит в замене (перекодировке) каллиграфически оптимальной кодировки специфических транскрипционных символов (в частности, акцентных диакритических знаков) в кодировку, совместимую с форматом STARLING.

0.3. Экспорт данных из Winword в формат файлов .RTF или .TXT. В этих форматах данные становятся доступными для работы с ними в среде STARLING.

1.0. Импорт данных в среду STARLING.

1.1. Конвертирование данных из формата текстового файла в формат примитивной БД.

1.2. Первичная структурация данных.

1.3. Слияние отдельных ТБД в одну.

1.4. Введение полей для записи информации об абзацах и их заполнение.

1.5. Введение полей для записи информации о предложениях и их заполнение с опорой на сентенциальную делимитацию.

1.6. Введение полей для записи информации о клаузах и их заполнение с опорой на клаузальную делимитацию.

1.7. Введение полей для записи информации о синтагмах и их заполнение с опорой на синтагменную делимитацию.

1.8. Введение полей для записи информации о тактах и их заполнение с опорой на исходный перечень служебных глоссов.

1.9. Введение полей для записи информации о глоссах и их заполнение с опорой на графические пробелы.

2. Введение полей для записи информации о морфологических характеристиках словоформ и их заполнение.

2.1. Введение вспомогательных полей для ближайших орфографических литературных аналогов диалектных словоформ и их заполнение с опорой на систему буквенных соответствий между транскрипцией диалектного текста и орфографической записью литературных словоформ.

2.2. Введение поля для записи морфологических характеристик словоформ и их автоматическое заполнение с помощью автоматического анализатора с опорой на знание морфологии русского литературного языка в объёме грамматического словаря А. А. Зализняка.

2.3. Сплошная ручная правка результатов морфологического разбора, предполагающая сплошной просмотр всего корпуса.

2.4. Полуавтоматическая расстановка акцентологических и семантических помет при именах лексем с опорой на материалы тетрадей.

2.5.А. Полуавтоматическая расстановка грамматических помет при словоформах с опорой на пометы в тетрадях.

2.5.Б. Полуавтоматическая расстановка помет о частях речи и грамматических формах с опорой на результаты работы морфологического анализатора.