С. А. Крылов, А. В. Тер-Аванесова

Электронные базы данных по русским народным говорам

0. Описание структуры базы данных:

Лингвистическая информация в базе организована по многоступенчатому принципу. Выделяется 9 уровней членения письменного текста; на каждом из них выделяется своя основная (базовая) единица членения (характеристику уровней членения текста и единиц разных уровней см., например, в: С.А. Крылов. О частотном словаре фонетических слов // Фонетика и нефонетика. К 70-летию С.В. Кодзасова. М., 2008. С. 387-399).

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

2. Уровень сверхфразового единства. У сверхфразового единства есть некоторая единая общая смысловая тема.

3. Уровень предложения.

4. Уровень предикации. Границы предикаций помечались так: предложение состоит из предикаций, а между предикациями внутри предложения стоит один из клаузальных делимитаторов (";", ":" и "-"). Предикации часто соответствуют простым предложениям и отдельным предикациям (частям) в составе сложных предложений.

5. Уровень синтагмы. Границы синтагм внутри клаузы помечены пунктуационным синтагматическим делимитатором ("запятая"). Содержательно и интонационно синтагмы примерно соответствуют словосочетаниям.

6. Уровень макротакта. Уровень речевого макротакта примерно соответствует фонетическим словам, членам предложения, "синтаксическим молекулам". Важнейшее фонетическое свойство макротакта: внутри него невозможна (или по меньшей мере нетипична) пауза.

7. Уровень мезотакта. Важнейшее фонетическое свойство мезотакта: внутри него лишь один из слогов обладает полноценным (неослабленным) ударением. Мезотакт обязательно включает одно полноударное слово, но может включать также одно или несколько слабоударных слов (клитикоидов).

8. Уровень микротакта. Важнейшее фонетическое свойство микротакта: внутри него лишь один из слогов обладает главным словесным ударением. Микротакт обязательно включает одно акцентное слово, но может включать также одно или несколько безударных слов (клитик).

9. Уровень словоформы. Каждый такт состоит из одной или нескольких словоформ. Словоформы, входящие в состав одного такта, обладают признаком потенциальной подвижности в предложении.

Для обозначения границ словоформ при разметке был использован специальный набор нескольких метаязыковых делимитаторов - "служебных пробелов". Выделены служебные пробелы шести типов: "{" между проклитикой и ее правой опорой; "}" между энклитикой и ее 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: “{}” – между проклитикой и энклиномеом.

Внутри словоформы (так же как внутри такта) невозможна пауза. Фактически наиболее близкий аналог словоформ в письменном тексте, записанном по правилам русской орфографии - это графические слова.

Предложенная многоуровневая схема позволяет при необходимости вывести на обозрение список отрезков текста, обладающих некоторым общим свойством. STARLING

позволяет пользователю базы по выбору вывести (на экран, на принтер или в файл) отрезок не только одного формата, но разных форматов - словоформу, минимальный контекст этой словоформы (например, предложно-падежную форму, сочетание клитики с акцентно автономной словоформой и т. п. - такт), синтагму, предикацию, предложение, сверхфразовое единство.

Комментарии к заголовкам полей

[LOC] = Район.
Указание района (для этой базы данных - это
(1) с. Пустоша Шатурского р-на Московской обл. (записи 1993-2006 гг.),
(2) д. Якушевичи Шатурского р-на Московской обл. (запись 2011 г.),
(3) с. Новоселки Рыбновского р-на Рязанской обл. (запись 2006 г.).

[INFORMANT] = Фамилия, имя, отчество информанта, год его рождения.
Иногда указывается девичья фамилия и родственные связи с другими информантами.
[SOURCE] = Звуковой файл.
Записи хранятся в фонотеке при отделе фонетики Института русского языка РАН, где каждому звуковому файлу присвоен индекс. Здесь указано наименование звукового файла.
[ADRES] = Адрес абзаца
[YEAR] = Год записи
[TEXT] = Текст
[SENTENCE] = Предложение
[CLAUSE] = Предикация
[PHRASE] = Синтагма
[MICROTACT] = Речевой микротакт
[MEZOTACT] = Речевой микротакт
[MICROTACT] = Речевой микротакт
[GLOSSIZOL] = Словоформа без пунктуации
[GLOSS] = Словоформа с пунктуацией
[GLOSSADR] = Номер словоформы
[WORD] = Ближайший нормативный аналог словоформы. Он дается в орфографической записи.
[LEXEM0] = Репрезентация словоформы в словаре: "начальная форма" слова. Она тоже даётся в орфографической записи.
[GRAMM] = словоизменительные грамматические характеристики словоформы (здесь используется аббревиатурная нотация StarLing'а).
[SEMNOTE] = Семантическое пояснение собирателя материала (или реплика собирателя материала, вставляемая между репликами информанта).

1. В настоящее время в базе данных представлены 2562 сверхфразовых единства (абзаца), 12918 предложений, 13057 пунктуационных клауз, 29149 пунктуационных синтагм, 69749 макротактов, 70356 мезотактов, 78263 микротакта, 99630 графических слов.

Проведена морфологическая разметка базы данных: лексико-грамматический разбор с частичным снятием грамматической и лексико-грамматической омонимии. По предварительной оценке, словарь лексем содержит около 7000 вокабул. На основе морфологического анализатора, базирующегося на подкорпусе пустошенских текстов, созданном в 2005 г., разработана новая улучшенная версия морфологического анализатора, кардинально отличающаяся от версии 2005 г. своей опорой на корпус 2012 г.

Материалом для создания базы данных по русским говорам 2012 г. послужили расшифровки аудиозаписей речи носителей следующих говоров:

(1) с. Пустоша Шатурского р-на Московской обл., в прошлом – Судогодского у.

Владимирской губ. (записи 1993-2006 гг.),

(2) д. Якушевичи Шатурского р-на Московской обл., в прошлом – Егорьевского у. Рязанской губ. (запись 2011 г.),

(3) с. Новоселки Рыбновского р-на Рязанской обл., в прошлом – Рязанского у. Рязанской губ. (запись 2006 г.).

Информанты – представители старшего поколения (1910 -1938 гг. р.). Аудиозаписи проводились с начала XXI в., и многие информанты, чья речь была записана за эти годы, связаны родственными отношениями. Теперешние младшее и среднее поколение имеет тенденцию утрачивать местный говор; это видно по аудиозаписям речи родителей, детей и внуков.

Оцифровка аудиозаписей проведена М. Н. Толстой (Ин-т славяноведения РАН). В расшифровке аудиозаписей принимали участие: Ю. Казенова (студентка, Пермский ун-т), А. Киреева (студентка, Санкт-Петербургский ун-т), Е. Корпечкова (аспирантка, ИРЯ РАН), Н. Михова (доцент, Череповецкий ун-т), Ю. Романова (студентка, Череповецкий ун-т), А. Тер-Аванесова (с.н.с., ИРЯ РАН), В. Шапоренко (студентка, Православный Свято-Тихоновский ун-т). Корректура расшифровок выполнена А. Тер-Аванесовой.

Работа по преобразованию текстовых файлов в базы данных выполнена С.А. Крыловым (в.н.с., ИВ РАН), последующее редактирование этих баз данных выполнены С.А. Крыловым при участии А.В. Тер-Аванесовой.

Снабжение базы данных интерфейсом, принятым на сайте «Вавилонская башня», осуществлено Ф. С. Крыловым.

В базу 2012 г. в настоящее время не включены ранее созданные базы данных по говору дд. Арзубиха, Захариха и Злобиха Харовского р-на Вологодской обл. (2008 г.) и по говору с. Пустоша (2005 г.).

2. Говоры, отраженные в базе данных, относятся к разным синхронным диалектным объединениям: говор с. Пустоша – среднерусский «окающий», говор, д. Якушевичи – среднерусский «акающий», говор с. Новоселки – южнорусский с так наз. «новоселковским» типом диссимилятивного предударного вокализма.

Их важной общей характеристикой является противопоставление четырех ступеней подъема гласных и, соответственно, фонем /уо/, /ие/ (верхне-среднего подъема) и /о/, /е/ (среднего подъема). К тому же во всех этих говорах, в противоположность подавляющему большинству современных говоров русского языка, имеющему более простой состав вокализма в области среднего подъема, согласные не смягчаются перед /е/ из *е, *ь (видимо, исконно), а в значительной части перечисленных говоров «общерусские палатализованные» согласные отвердевают и в других позициях. Как правило, отвердение не выражается в полной депалатализации согласных, как, например, согласных перед *е, *i в украинском, а лишь в частичной, и не приводит к утрате оппозиции «твердых» и «мягких» согласных фонем. Полная депалатализация согласных и утрата названной оппозиции имеет место только в вологодском харовском говоре (но и в нем эффект депалатализации снимается на морфонологическом уровне).

Фонемы области среднего подъема в говорах восходят к: /е/ < *е, *ь, /ие/ < *ě, /о/ < *о (под праслав. «нисходящим ударением»), *ъ, а также *е, *ь, /уо/ < *о (под праслав. «восходящим ударением»). Заимствования и исконные словоформы, место ударения которых изменилось в относительно позднее время, образуют особую подсистему (подсистемы), в которой правила выбора огласовки ударного слога (/ие/ или /е/, /уо/ или /о/) иные. Тем самым, база данных включает материал архаичных русских говоров, данные которых имеют большое значение для славянской акцентологии.

В базу данных включены материалы разного рода: по всем говорам, кроме говора Пустошей и вологодского, в базу вошли в основном ответы на вопросники (так называемые «фонетическую», «морфологическую», «акцентологическую» программы сбора материала). Вологодский и пустошенский говоры представлены, помимо этого, текстами. Поэтому в базе

относительно большое место занимают сопоставимые данные, а сама база – достаточно представительна, так как включает целенаправленно собранные непроизводные лексемы праславянского словарного фонда.

База данных включает материал, характеризующий словоизменение, акцентуацию и распределение двух фонем «типа о» в трех территориально и генетически близких говорах. В отношении их акцентных систем обращает на себя внимание: 1) сходство в деталях акцентуации и распределения в корнях двух фонем «типа о» у непроизводных существительных мужского рода; в этом классе слов говоры имеют систему ударения, восходящую к «восточному» позднепраславянскому диалекту (Основы славянской акцентологии, 1990, с. 132-133), и их данные позволяют уточнить акцентную специфику этого диалекта; система ударения существительных мужского рода в трех изученных говорах в целом близка обнаруженной в харовском вологодском, липецком и белгородском говорах, все – с различением двух фонем «типа о»; архаичностью системы выделяется говор Пустошей; 2) ударение существительных а-склонения в трех говорах также в целом сохраняет старое противопоставление а. п. b и с, однако в о-огласовке форм мн. числа есть существенные различия между пустошенским и прочими говорами; все три говора при этом не разделяют инноваций в ударении и огласовке словоформ, свойственных вологодским говорам, с одной стороны, и липецким, белгородским и воронежским – с другой; 3) в том или ином виде в трех говорах сохраняются результаты особого «рязанского» развития а. п. b глаголов: колонного накоренного ударения в презенсе: Пустоша но\шу, нуо\сишь, Якушевичи ска\жу, ска\жешь, Новоселки нуо\шу, нуо\сишь. Акцентологические данные указывают как на специфическую близость трех говоров между собой, так и на включенность их в обширный восточнорусский ареал.

В результате работы над проектом получены сплошные расшифровки большого массива записей устной речи носителей одного говора - с. Пустошей, довольно близкого, по-видимому, к так называемому «московскому просторечию», или «языку московских мещан». Пустошенские тексты дают материал для изучения особенностей организации устной речи, строения предложения, грамматических категорий или форм, значение которых выводится только из контекста, иногда – достаточно обширного (ср. материал по деепричастиям наст. времени: Не работя\ не заплуо\тют ничевуо\; Вуо\т связа\ла ва\режки по па\мяти, не зна\я; Он надойе\ст тебие ходя\ ), прош. времени (Корьзи\ны плесть на\до ведь умие\мшы; И вуо\т ничевуо\ не дие\ламшы - им де\ньги; Оние\ и росчасо\мшы човуо\ у йе\й за\ волосы; И вуо\т она\ жалие\мшы мние спра\вку написа\ла); «антирезультативу» (Меня\ раз то\ком был уби\ло), формообразующих и дискурсивных (преимущественно отглагольных) слов (дай, давай, знать, знай, пой и поди, мол, бы, был, бывало, будь-будь, ведь и др.).

База данных позволяет получить приблизительную оценку степени фонетической вариативности, присущей словоформе.

Жанр текста характеризует запись речи одного информанта, а также более или менее длинный отрывок. Типичные жанры текстов: "рассказ о жизни", "рассказ о работе в колхозе", "рассказ о детях", "рассказ о войне" и т.п.; "ответы на вопросы этнолингвистической программы", иногда с конкретизацией: праздники, свадьба, похороны, родины, нечистая сила, скот, сев и жатва, растения и животные и т. п.; ответы на вопросы акцентологической (морфологической, фонетической программы).