УДК 81’27
DOI: 10.17223/22274200/15/8

З.И. Резанова

КОРПУС УСТНОЙ РЕЧИ РУССКО-ТЮРКСКИХ БИЛИНГВОВ ЮЖНОЙ СИБИРИ: РАЗМЕТКА ОТКЛОНЕНИЙ ОТ РЕЧЕВОГО СТАНДАРТА

Характеризуются подходы к разметке отклонений от речевого стандарта (ОРС, error annotation) в создаваемом корпусе устной речи русско-тюркских билингвов Южной Сибири и обосновывается система отмечаемых в корпусе отклонений. Приводятся примеры тэгов, фиксирующих отклонения от речевого стандарта на фонетическом, грамматическом, лексическом, дискурсивном уровнях, и тэгов, маркирующих источники отклонений (внутриязыковые, межъязыковые), а также примеры разметки (аннотирования) текстов (на материале записей устной речи русско-шорских билингвов).
Ключевые слова: *билингвизм, корпус текстов, аннотирование, отклонения от речевого стандарта, интерференция, тюркские языки, русский язык, сибирские говоры.*

В статье характеризуется система разметки (аннотирования) отклонений от речевого стандарта в Корпусе устной речи русско-тюркских билингвов Южной Сибири, создаваемого в рамках проекта «Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур». Общие подходы к созданию корпуса, его структурные особенности и принципы метаразметки были охарактеризованы в [1, 2]. В данной работе мы представляем систему аннотирования элементов текстов корпуса.

Основной материал корпуса – записи устной спонтанной речи русско-тюркских билингвов, вследствие этого корпус является бимодальным, звучание речи синхронизируется с транскрипцией, что реализуется с использованием программы «ELAN».

Транскрипция проводится на основе использования принципов русской орфографии, дополнительно в записях фиксируются только значительные отклонения от произносительных норм также без применения знаков транскрибирования, например *щас* (сейчас), *грит* (говорит), *мушшина* (мужчина) и под.).

Морфологическая разметка проводится автоматически на основе использования консольной программы компании «Яндекс» «Mystem»

с дальнейшей ручной коррекцией. Морфологическое аннотирование, таким образом, соответствует системе, заложенной в данной программе, которая основывается на принципах, реализованных в «Грамматическом словаре русского языка» А.А. Зализняка [3].

Так как целью корпуса является репрезентативное представление устной речи билингвов, содержащих, в соответствии с гипотезой авторов проекта, проявление интерференции материнских тюркских языков, наряду с традиционным морфологическим аннотированием в корпусе содержится разметка так называемых ошибок (error annotation). Такой тип разметки осуществляется в практике мировой корпусной лингвистики при создании корпусов «неносителей языка», включающих, как отмечает М. Копотев, многоуровневую разметку, тэги которой фиксируют ошибки в употреблении единиц разных языковых уровней, а также «источник ошибки: внутриязыковое, межъязыковое влияние» [4. С. 106].

При разработке системы аннотирования данного типа мы используем другой термин – «отклонения от речевого стандарта» (ОРС), так как в корпусе представлена речь носителей разных форм русского национального языка: литературной, диалектной просторечной, речевые практики которых соответствуют нормам соответствующих подсистем, не являясь «ошибками». В качестве стандарта приняты нормы русской письменной литературной речи, на основе которой формировались принципы грамматического описания языка в «Грамматическом словаре русского языка» А.А. Зализняка. На основе этих принципов как отмечалось, действует система автоматической морфологической разметки «Mystem». Корпус представляет собой записи устной речи, вследствие этого ряд отклонений от стандарта письменной речи обусловливается устным, спонтанным характером коммуникации. Фиксация таких отклонений, на наш взгляд, расширяет информативные возможности корпуса за счет аннотирования маркеров разговорности коммуникации.

При создании представляемого корпуса русскоязычной билингвальной речи применяются два вида маркирования «ошибок»: в соответствии с уровнями языковой системы и источниками ОРС.

При разработке системы маркирования отклонений от речевого стандарта мы ориентировались на существующую систему разметки в близких по направленности корпусах, среди которых назовем прежде всего Русский учебный корпус (RLC) [5], разрабатываемый членами

Лаборатории по корпусным исследованиям НИУ ВШЭ под руководством Е. Рахилиной.

Система разметки ошибок в RLC ориентирована на письменную речь, на анализ речевых практик прежде всего учебных билингвов, поэтому центральным термином является термин «речевая ошибка», так как привлекаются тексты, порождаемые в учебной деятельности и оцениваемые с точки зрения соответствия письменным нормам современного русского языка. В проекте разработана сложно организованная система тэгов, маркирующих ошибки в использовании единиц всех уровней письменной речи: морфологические, например Num – употребление слова в неверной числовой форме (несоответствующей контексту или аномальной для этого слова); синтаксические, например Conj – ошибка в употреблении союза; лексические, например, Par – ошибки в использовании паронимов, а также орфографические и пунктуационные [5].

«Корпус контактно-обусловленной русской речи билингвов – носителей малых языков Севера Сибири и Дальнего Востока», создаваемый в рамках проекта «Динамика языковых контактов в циркумполярном регионе» (Н.М. Стойнова, П.С. Плешак, И.А. Хомченкова), более близок по направленности к разрабатываемому, однако как действующий корпус он еще не представлен, система тэгов также находится в состоянии разработки, приведем некоторые из них: number – нестандартное употребление числовой формы существительных, agr_adj – рассогласование по роду, числу, падежу: адъективы; calque – лексическая калька; pros – нестандартная интонация: потенциальная калька [6].

В Корпусе устной речи русско-тюркских билингвов Южной Сибири также принято поуровневое представление отклонений от речевого стандарта. При выборе системы тэгов мы ориентировались на вариант, представленный в RLC, по мере необходимости добавляя новые маркеры ОРС, действуя в той же логике обозначений.

Основное отличие в составе маркеров от RLC определяется различием фиксируемой формы речи: письменной – в RLC, устной – в корпусе речи билингвов, следствием чего является отсутствие тэгов орфографических и пунктуационных ошибок и введение помет, фиксирующих отклонения в области фонетики.

Отличия от представленного проекта разметки в Корпусе контактно-обусловленной русской речи содержатся в конкретных решениях размечаемых групп отклонений от речевого стандарта.

Далее охарактеризуем тэги, их обоснование, а также фрагменты текстов с соответствующими тэгами.

Как и в RLC, в Корпусе устной речи русско-тюркских билингвов Южной Сибири используем сокращенные варианты англоязычных терминов: Phon – phonetics, фонетика; Synt – syntax, синтаксис, Morph – morphology, морфология, Lex – lexis, лексика; Der – derivation, деривация (словообразование); Disc – discurs, дискурс; Sem – semantics, семантика; Acc – accent, ударение, Iinfl – inflexion, окончание; Aff – affix, аффикс; Decl – declension, склонение, Agr – agreement, согласование; Gov – government, управление; Id – idiom, идиома (устойчивые сочетания разных типов); Prep – preposition, предлог; Gen – Gender, грамматический род, Num – number, грамматическое число, Con – construction, конструкция, нарушения согласования в пределах простого и сложного предложения; Red – reduction, редукция и т.д.

При этом принята система одного или двух уровней квалификации ОРС, в незначительном количестве случаев представлен третий уровень конкретизации, что мотивируется необходимостью отражения интерферентных явлений. На первом уровне маркируется уровень языковой системы, к которому относится частный вариант ОРС: Phon – фонетика, Morph – морфология, Synt – синтаксис, Lex – лексика, Disk – дискурс; на втором уровне – конкретное языковое явление: PhonAcc – фонетика, ударение; MorphInfl – морфология, флексия; SyntGov – синтаксис, управление и т.д. (базовый вариант отклонения дополнительно не маркируется, например, маркер Phon отмечает все варианты отклонений в произношении звуков, а маркер PhonAcc – фонетика, ударение). На третьем уровне отражаются варианты отклонений при использовании единиц конкретного языкового уровня, их форм: SyntAgrGen – отклонения от речевого стандарта в согласовании по роду.

В качестве явлений фонетического уровня маркируются особенности произношения отдельных слов, позиционно обусловленные особенности произношения звуков, характерные для формы национального языка, редукция звуковой оболочки слов – нормы разговорной речи, особенности ударения и др. Морфологическими тэгами маркируются отклонения от норм грамматической категоризации слов и форм слова, отклонения от норм литературного письменного языка в выборе вариантов морфологических аффиксов, синонимическая замена морфологических формантов, варианты образования и использования маркеров определенной грамматической категории и под.

К деривационным явлениям относим отклонения от норм образования конкретных слов, образование слов по синонимичной словообразовательной модели, использование вариантных форм деривационных аффиксов и под.

В качестве синтаксических явлений маркируем отклонение от норм литературной письменной речи в выборе форм согласования, управления, порядка слов, синтаксических связей в составе простого и сложного предложения,

К лексическим явлениям относим использование в речи диалектных, просторечных, других региональных межсистемных синонимов, в том числе заимствований из материнских языков, проявление интерферентных явлений. К лексическому уровню в соответствии со сложившейся традицией относим также особенности употребления фразеологизмов, вариантные формы, идиомы-диалектизмы и пр.

В качестве дискурсивных явлений отмечается использование различного рода маркеров дискурсивной связности, заполнителей пауз, особенностей ритмической организации речи.

В табл. 1 представлены образцы тэгов, маркирующих ОРС на разных языковых уровнях[1].

В речи билингвов при использовании слов и построении высказываний могут быть проявлены отклонения от речевого стандарта на нескольких языковых уровнях, репрезентированных в одной языковой единице, в таком случае используется комбинация тэгов. Примеры сочетания тэгов представлены в табл. 2.

Таблица 1

Образцы тэгов, маркирующих ОРС на разных языковых уровнях

Тип тэга	Пояснение / определение	Пример аннотирования
Phon	Особенности произношения отдельных слов	*…А вы* ***чё*** [Phon] (что)? *Как? Живёте там? На горе были? подымались? Интересно?* ***Ничё*** [Phon](ничего)?
Phon	Отклонения от норм реализации системных фонетических явлений, регу-	*Мне сильно плохо было гыпертония* [Phon] (гипертония); *там сидела я эты* [Phon](эти) дни; *поетому* [PhonS]

[1] В качестве иллюстративных приводятся контексты из записей подкорпуса текстов русско-шорских билингвов, сделанных автором статьи. В приводимом контексте при наличии других отклонений маркируется только то, которое представляется в данной графе.

Тип тэга	Пояснение / определение	Пример аннотирования
	лярные (возможно, не всегда), позиционные мены: смягчение / отвердение, оглушение / озвончение, протеза, эпентеза, беглость гласных и под.	(поэтому) *он пришел*; *от* [Phon] (вот), *тяжело от* [Phon] (вот) *прошлое думать*; *у старого человека нету сона* [Phon] (сна), *оказывается*
PhonAcc	Отклонения от норм ударения	*Зимой на лыжАх* [PhonAcc] (лЫжах) *за продуктами ходили*
PhonRed	Значительная редукция звуковой оболочки слова, характерно для разговорной спонтанной речи	*Перво место, гыт* [PhonRed] (говорит), *заняли*; *а щас-то* (сейчас-то) [PhonRed] *я, здесь-то я в баночках солила всё время*
PhonLess	Слово в разговорной спонтанной речи не договаривается	*Потому что мне уже будет тридца.. семьдесят девять* [PhonLess] *девятого июня*
MorphAff	Отклонения от норм образования грамматических аффиксов, грамматических форм слова, выбор синонимичных формообразовательных аффиксов	*Перво* [Morphaff] (первое) *место, гыт, заняли*; *ой, как там солдатов* [MorphAff] (солдат) *обнимали, как плакали*; *я говорю, у менЕ* [MorphAff] (меня) *там все русские*; *…оне* [MorphAff] (они) *сказали Матор, поселок*; *Аньке своёй* [MorphAff] (своей) *бабушкина квартира*; *я сама-то не трогаю, боюся* [MorphAff] (боюсь)
MorphDecl	Отклонения от нормы нулевого склонения	*У них, у шорцев, было вот такие осенние пОльты* [MorphDecl] (пальто)
MorphNum	Отклонения от норм категоризации по числу	*А вы там делайте, чтоб в руках молочный был, сухая сливка* [MorphNum] (сливки), *да туда разбавляйте, она очень вкусна*
DerAff	Отклонения от норм словообразования, использование синонимичного деривационного аффикса, другого способа словообразования	*Вот это соседна* [DerAff] (соседская), *правнучка Дашечка*; *ну, ты мне, на меня посмотри, я ли колдовка* [DerAff] (колдунья) *или шаманка*; *сегодня ночью маленечко дремнула* [DerAff] (вздремнула); *старость такая, откудова* [DerAff] (откуда) *знать*
SyntGov	Отклонения от норм падежного управления	*Он у нас не работает, около наверно уже 8 лет прошло, как он у нас не работает, после него у нас пять священник* [Morph-Gov] (пять священников) *поменяли*; *кассеты есь. Я все прослушала, пять кассеты* [MorphGov] (пять

Тип тэга	Пояснение / определение	Пример аннотирования
		кассет); *около наверно уже восемь лет прошло* [MorphGov] (около восьми лет) *прошло*; *тяжело мы всякое от прошлое думать* [MorphGov] (о прошлом думать)
SyntPrep	Отклонение от норм использования предлогов	*Берут прямо с фляги* [SyntPrep] *(из фляги)*; *ну, например, я делаю с толкана* [SyntPrep] *(из толкана) кашу*
SyntAgrGen	Отклонения от норм согласования по роду	*Мы говорим, ну ладно, ты там самый этот* [MorphAgrGen] от *шорского байраму, такая всё-всё можешь сделать* (в рассказе о женщине)
SyntAgrGen	Отклонения от норм согласования по числу	*Я, это, дети войны* [SyntAgrGen] (мы дети войны)
SyntWO	Нарушение порядка слов	*Как эти говорят узбеки* [SyntWO] (как говорят эти узбеки)
SyntLessW	Пропуск слова, незаконченные предложения / высказывания	*Рима говорит, давай новый* [SyntLessW] (диван купим). *Она говорит, давай позовём* [SyntLessW] (мастера), *который диван делает*
SyntCon	Нарушение норм семантико-синтаксического сочетания элементов	*И вот сюда* [Lex] (здесь) *вот лежу*
Lex	Использование диалектизмов, просторечных лексем, заимствований из тюркских языков	*Вот у вас Баба Яга, а у нас Чельбень* [Lex]; *там дерево стоит и человек, это Чельбень* [Lex], *это баба Яга; пельмени, вот это на рёбрах, называется сало, по-шорски – каза* [Lex]
LexId	Использование сочетаний разной степени устойчивости, имеющих фразеологическое значение	*Кого там* [LexId] (невозможно) *спать*; *Иногда гости за гостями бывают, а другой раз* [LexId] (иногда) *никого нет*; *Старость такая, откудова* [LexId] *знать* (не могу знать); *Ну, не знаю, я всю дорогу* [LexId] (все время) *по-шорски говорю*
LexSem	Использование общерусского слова в ином значении	*Она, во-первых, не курила, не пила, поэтому, наверно, с бабушками не обращалась* [LexSem] (общалась)
Disc	Использование в качестве средств ритмической организации речи маркеров авторизации	*Да, Вера, грит* [Disc] *(говорит), ты, говорят, гыт* [Disc] *(говорит)*; *предсказываешь. Я говорю, это мне нехорошо, говорю* [Disc]. *Мне в глаза, говорю* [Disc], *человеку умереть говорю* [Disc]

Таблица 2

Примеры сочетания тэгов

Сочетание тэгов	Пояснение / определение	Пример аннотирования
SyntGov; SyntAgrGen	Отклонение от норм грамматического управления, нарушение норм согласования по роду	*Сделала тут такой яма* [SyntGov; SyntAgrGen] (такую яму)
Phon; MorphAff; DerAff	Отклонение от нормы образования флексии, использование синонимичного деривационного суффикса	*Вот это соседна* [Phon; MorphAff; DerAff] (соседская) *правнУчка Дашечка*
MorphGov; MorphAff	Отклонение от нормы образования флексии, от норм падежного управления	*И дорога эта, котора* [MorphGov; MorphAff] (которую) *сделали*

Как было отмечено ранее, аннотирование отклонений от речевого стандарта включает также тэги источника отклонения, в корпусе маркируется внутриязыковое и межъязыковое влияние. Данный тип маркирования обусловлен направленностью создаваемого корпуса на фиксацию речевых практик билингвов, на выявление факторов, определяющих характер и степень проявления интерференции на всех языковых уровнях.

Носители билингвизма, речевые практики которых представлены в корпусе устной речи русско-тюркских билингвов Южной Сибири, принадлежат к разным возрастным и социальным группам, имеют различное образование и являются носителями разных вариантов русского языка: литературного, диалектного, диалектно-просторечного, просторечного.

При классификации ОРС в данном аспекте мы ориентировались на работы, в которых содержится анализ особенностей русской разговорной речи на всех уровнях языка ([7–9] и др.), русских сибирских говоров и городского просторечия ([10–14] и др.); на работы по теории интерференции и типологии тюркских языков ([15–18] и др.) и работы по языковому контактированию исследуемого региона ([19–21] и др.).

Отклонения от речевого стандарта, являющиеся проявлением регионального варианта современного русского языка, маркируются тэгом [Reg] – regional, региональный, который объединяет все прояв-

ления региональных вариантов – сибирские говоры, сибирский вариант городского просторечия, региональные варианты литературного языка. Как было отмечено ранее, заключение о внутрисистемном источнике отклонений от стандарта литературного языка выносится на основе данных, зафиксированных в работах, посвященных описанию рассматриваемых подсистем русского языка.

Решение не дифференцировать в разметке, является ли отклонение отражением диалектной, диалектно-просторечной, просторечной речи или проявлением регионального варианта русского литературного языка, определяется тем, что в настоящее время границы между первыми тремя формами существования языка размыты, тем, что некоторые черты диалектной речи могут проникать и в речевые практики носителей литературного языка, например так называемое чёканье произношение: чё, чё-то и под.

Отклонения от речевого стандарта, проявления влияния структурных особенностей материнского языка билингва обозначаем тэгом [Int] – interference, интерференция. Однако в том случае, когда характер таких влияний не представляется очевидным, данный тип тэга пропускается.

Отражения общих закономерностей устной спонтанной разговорной речи в корпусе дополнительными тэгами не маркируются.

Примеры аннотирования с использованием двух типов тэгов представлены в табл. 3.

На рис. 1 представлен вариант совмещения фонетического трека, записанного текста, его морфологической разметки и разметки ОРС, реализованный в программу ELAN.

Таблица 3

Примеры сочетания тэгов двух типов в аннотировании

[Phon; Reg]	Регрессивная ассимиляция БМ → ММ, отмечаемая в говорах часто как лексикализованное явление; выпадение [В] в начале слов перед [И] и [О]	*Они мне омманули* [Phon; Reg] (обманули)*; А потом от* [Phon; Reg] (вот) *я одна осталась*
MorphPrep; Reg	Диалектно-просторечный вариант предложного управления	*Лагушка берут прямо с фляги* [MorphPrep; Reg] (из фляги). *Прямо с серёдки* [MorphPrep; Reg] *берут* (из середины)

MorphAff; Reg	Диалектно-просторечный вариант формообразовательного суффикса	*Какой мне позвоночник править, я сама-то не трогаю, боюся* [MorphAff; Reg] (боюсь)
Phon; Int	Позиционно не обусловленные мены*: оглушение в абсолютно сильной позиции перед гласным (Д → Т, Б → П)*	*Куда бабушка лОжила тушинчку* [Phon; Int] (душничку), *но здесь тушыстый* [Phon; Int] (душистый) *перец я только ложу, и соли побольше*
SyntGov; Int SyntAgrGen; Int; SyntGov; Int; SyntAgrGen; Int	Нарушение норм грамматической категоризации существительных как отражение влияния материнского языка отсутствующей категорией рода, нарушение норм управления определяется наличием неопределенного падежа в тюркских языках	*Сделала тут такой яма* [SyntGov; Int SyntAgrGen; Int] (такую яму)
MorphGov; Int; MorphAff; Reg	Нарушение норм управления как отражение межъязыковой интрференции; стяжение флексии прилагательного – региональное явление	*Много, гыт, шорска блюда* [MorphGov; Int; MorphAff; Reg] (шорских блюд) *готовили*

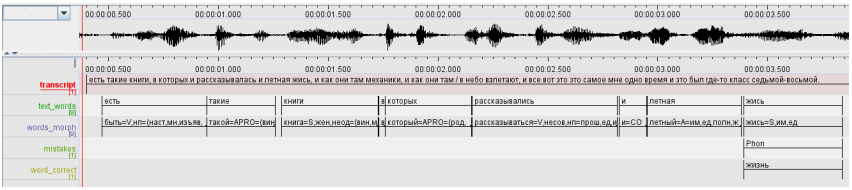

Рис. 1. Фрагмент размеченного корпуса устной речи русско-тюркских билингвов Южной Сибири

Таким образом, морфологическое аннотирование и разметка отклонений от речевого стандарта определяют широкий диапазон поисковой системы корпуса, возможность использования его данных при исследовании единиц разных уровней языковой системы, проявленных в регионально ограниченной речи носителей русско-тюркского билингвизма. Соединение с системой метаразметки, принятой в корпусе [1, 2], расширяет возможности анализа за счет соотнесения типов отклонения с типами билингвизма, социокультурными типами говорящих.

Литература

1. *Резанова З.И.* Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. № 11. С. 105–118.

2. *Резанова З.И., Некрасова Е.Д., Миклашевский А.А.* Исследование психолингвистических и когнитивных аспектов языкового контактирования в проекте «Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур» // Русин. 2018. № 2 (52). С. 107–117.

3. *Зализняк А.А.* Грамматический словарь русского языка: Словоизменение. М. : Рус. яз., 1980. 880 с.

4. *Копотев М.* Введение в корпусную лингвистику. Praha : Animedia Company, 2014. 195 с.

5. *RLC.* Русский учебный корпус. URL: http://www.web-corpora.net/RLC/ (дата обращения: 14.01.2019).

6. *Корпус* контактно-обусловленной русской речи. URL: http://web-corpora.net/tsakorpus_russian_nonst/ (дата обращения: 14.01.2019).

7. *Земская Е.А., Китайгородская М.В., Ширяев Е.Н.* Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М. : Наука, 1981. 276 с.

8. *Русская* разговорная речь / отв. ред. Е.А. Земская. М. : Наука, 1973. 485 с.

9. *Русская* разговорная речь. Фонетика. Морфология. Лексика. Жест / отв. ред. Е.А. Земская. М. : Наука, 1983. 239 с.

10. *Русские* говоры Среднего Приобья. Ч. 1 / под ред. В.В. Палагиной. Томск : Изд-во Том. ун-та, 1984. 201 с.

11. *Русские* говоры Среднего Приобья. Ч. 2 / под ред. В.В. Палагиной. Томск : Изд-во Том. ун-та, 1989. 323 с.

12. *Араева Л.А.* Говоры Кузбасса в их современном состоянии // Координационное совещание по проблемам изучения сибирских говоров кафедр русского языка вузов Сибири, Урала и Дальнего Востока. Красноярск, 1991. С. 38–65.

13. *Банкова Т.Б.* Лексика томского городского просторечия (типология описания) : дис. ... канд. филол. наук. Томск, 1987. 164 с.

14. *Блинова О.И.* Просторечная лексика в системе местного диалекта // Лексикологический сборник. Барнаул, 1977. С. 68–83.

15. *Bacanlı E.* Inflectional suppletion in Turkic languages // Folia Linguistica Historica. 2011. № 32. P. 1–42.

16. *Будренюк Г.М., Григоревский В.М.* Языковая интерференция и методы ее выявления. Кишинев : Штиинца, 1978. 126 с.

17. *Сравнительно-историческая* грамматика тюркских языков. Лексика. М. : Наука, 2001. 288 с.

18. *Сравнительно-историческая* грамматика тюркских языков. Морфология. М. : Наука, 1988. 557 с.

19. *Гордеева О.И.* Некоторые закономерности влияния родного языка на усваиваемый язык в процессе становления двуязычия (на материале русского и татарского сибирских говоров) : дис. ... канд. филол. наук. Томск, 1965. 239 с.

20. *Гордеева О.И.* Об освоении татарами грамматической категории рода существительных в условиях русского старожильческого окружения // Лингвистический сборник. Томск, 1962. С. 29–34.

21. *Абдрахманов М.А.* К вопросу о закономерностях диалектноязыкового смешения (на материале тюркского говора дер. Эушта Томского района : дис. ... канд. филол. наук. Томск, 1960. 261 с.

The Oral Speech Corpus of Russian-Turkic Bilinguals of Southern Siberia: The Marking of Deviations from the Speech Standard

Voprosy leksikografii – Russian Journal of Lexicography, 2019, 15, pp. 127–140.
DOI: 10.17223/22274200/15/8

Zoya I. Rezanova, Tomsk State University (Tomsk, Russian Federation).
E-mail: rezanovazi@mail.ru

Keywords: bilingualism, corpus of texts, annotation, deviations from the speech standard, interference, Turkic languages, Russian, Siberian dialects.

The article describes approaches to the marking of deviations from the speech standard (error annotation) in the corpus of oral speech of Russian-Turkic bilinguals of Southern Siberia.

The texts of the corpus are recordings of oral spontaneous speech of Russian-Turkic bilinguals, as a result of which the corpus is bimodal, the sound of speech is synchronized with the transcription, which is realized using ELAN. Morphological annotation is performed automatically using the Mystem console program of the Yandex company with further manual correction.

Along with the traditional morphological annotation, the corpus contains the annotation of the so-called “errors” (error annotation). Two types of labeling “errors” are applied, the first in accordance with the levels of the language system, the second in accordance with their sources. The article describes tags, provides fragments of texts with these tags. The key corpus tags are: Phon – phonetics, Synt – syntax, Morph – morphology, Lex – lexis, Der – derivation, Disk – discourse, Sem – semantics, Acc – accent, Infl – inflexion, Aff – affix,; Decl – declension, Agr – agreement,; Gov – government, Id – idiom, Prep – preposition, preposition; Gen – Gender, Num – number, Con – construction; Red – reduction, etc.

The annotation of deviations from the speech standard also includes tags of the source of the deviation. This type of marking is determined by the focus of the created corpus of texts on fixing speech practices of bilinguals, on identifying the factors that determine the nature and degree of interference manifestation at all language levels. Translingual and intralingual influence is annotated in the corpus. The influence of the norms of the regional variant of the Russian language is marked as an intralingual one.

Regionally determined deviations from the speech standard are marked with a [Reg] – a regional tag. This tag unites all manifestations of regional variants – Siberian dialects, the Siberian variant of urban vernacular, regional variants of the literary language. Deviations from the speech standard, which are a manifestation of the influence of the features of the bilingual’s mother tongue, are indicated by [Int] – interference.

The peculiarities of oral colloquial speech in the corpus of texts of Russian-Turkic bilinguals are not marked with additional tags.

The article provides examples of annotation using two types of tags.

The morphological annotation and marking of deviations from the speech standard define a wide range of the search engine of the corpus of oral speech of the Russian-Turkic bilinguals of Southern Siberia. Connecting to the body corpus meta-markup system expands the search capabilities by matching the types of deviations with the types of bilingualism and the sociocultural types of speakers.

References

1. Rezanova, Z.I. (2017) Subcorpus of oral speech of Russian-Turkic bilinguals of Southern Siberia: typologically relevant signs. *Voprosy leksikografii – Russian Journal of Lexicography*. 11. pp. 105–118. (In Russian). DOI: 10.17223/22274200/11/7

2. Rezanova, Z.I., Nekrasova, E.D. & Miklashevskiy, A.A. (2018) Investigation of psycho-linguistic and cognitive aspects of language contacting in the project "Linguistic and Ethnocultural Diversity of Southern Siberia in Synchrony and Diahrony: Interaction of Languages and Cultures". *Rusin*. 2 (52). pp. 107–117. (In Russian). DOI: 10.17223/18572685/52/8

3. Zaliznyak, A.A. (1980) *Grammaticheskiy slovar' russkogo yazyka: Slovoizmenenie* [Grammatical Dictionary of the Russian Language: Word Change]. Moscow: Rus. yaz.

4. Kopotev, M. (2014) *Vvedenie v korpusnuyu lingvistiku* [Introduction to corpus linguistics]. Prague: Animedia Company.

5. RLC. *Russkiy uchebnyy korpus* [Russian Learner's Corpus]. [Online] Available from: http://www.web-corpora.net/RLC/. (Accessed: 14.01.2019).

6. *Korpus kontaktno-obuslovlennoy russkoy rechi* [Corpus of contact-influenced Russian speech]. [Online] Available from: http://web-corpora.net/tsakorpus_russian_nonst/. (Accessed: 14.01.2019).

7. Zemskaya, E.A., Kitaygorodskaya, M.V. & Shiryaev, E.N. (1981) *Russkaya razgovornaya rech'. Obshchie voprosy. Slovoobrazovanie. Sintaksis* [Russian colloquial speech. General issues. Word formation. Syntax]. Moscow: Nauka.

8. Zemskaya, E.A. (ed.) (1973) *Russkaya razgovornaya rech'* [Russian colloquial speech]. Moscow: Nauka.

9. Zemskaya, E.A. (ed.) (1983) *Russkaya razgovornaya rech'. Fonetika. Morfologiya. Leksika. Zhest* [Russian colloquial speech. Phonetics. Morphology. Vocabulary. Gesture]. Moscow: Nauka.

10. Palagina, V.V. (ed.) (1984) *Russkie govory Srednego Priob'ya* [Russian dialects of the Middle Ob]. Pt. 1. Tomsk: Tomsk State University.

11. Palagina, V.V. (ed.) (1989) *Russkie govory Srednego Priob'ya* [Russian dialects of the Middle Ob]. Pt. 2. Tomsk: Tomsk State University.

12. Araeva, L.A. (1991) Govory Kuzbassa v ikh sovremennom sostoyanii [Kuzbass dialects in their present state]. In: Belousova, G.G. (ed.) *Koordinatsionnoe soveshchanie po problemam izucheniya sibirskikh govorov kafedr russkogo yazyka vuzov Sibiri, Urala i Dal'nego Vostoka* [Coordination meeting on the problems of studying Siberian

dialects of the Russian language at departments of universities in Siberia, the Urals and the Far East]. Krasnoyarsk: Krasnoyarsk State Pedagogical Institute.

13. Bankova, T.B. (1987) *Leksika tomskogo gorodskogo prostorechiya (tipologiya opisaniya)* [Vocabulary of Tomsk vernacular (a typology of the description)]. Philology Cand. Diss. Tomsk.

14. Blinova, O.I. (1977) Prostorechnaya leksika v sisteme mestnogo dialekta [Vernacular vocabulary in the system of the local dialect]. In: Arzhanykh, L.M. (ed.) *Leksikologicheskiy sbornik* [A lexicological collection]. Barnaul: Barnaul State Pedagogical Institute.

15. Bacanlı, E. (2011) Inflectional suppletion in Turkic languages. *Folia Linguistica Historica*. 32. pp. 1–42. DOI: 10.1515/flih.2011.002

16. Budrenyuk, G.M. & Grigorevskiy, V.M. (1978) *Yazykovaya interferentsiya i metody ee vyyavleniya* [Language interference and methods for its detection]. Kishinev: Shtiintsa.

17. Tenishev, E.R. (ed.) (2001) *Sravnitel'no-istoricheskaya grammatika tyurkskikh yazykov. Leksika* [Comparative historical grammar of Turkic languages. Vocabulary]. Moscow: Nauka.

18. Tenishev, E.R. (ed.) (1988) *Sravnitel'no-istoricheskaya grammatika tyurkskikh yazykov. Morfologiya* [Comparative historical grammar of Turkic languages. Morphology]. Moscow: Nauka.

19. Gordeeva, O.I. (1965) *Nekotorye zakonomernosti vliyaniya rodnogo yazyka na usvaivaemyy yazyk v protsesse stanovleniya dvuyazychiya (na materiale russkogo i tatarskogo sibirskikh govorov)* [Some regularities of the influence of the native language on the adopted language in the process of the formation of bilingualism (on the material of the Russian and Tatar Siberian dialects)]. Philology Cand. Diss. Tomsk.

20. Gordeeva, O.I. (1962) Ob osvoenii tatarami grammaticheskoy kategorii roda sushchestvitel'nykh v usloviyakh russkogo starozhil'cheskogo okruzheniya [On the development by the Tatars of the grammatical category of the gender of nouns in the context of the Russian old-timer environment]. In: Palagina, V.V. (ed.) *Lingvisticheskiy sbornik* [A linguistic collection]. Tomsk: Tomsk State University.

21. Abdrakhmanov, M.A. (1960) *K voprosu o zakonomernostyakh dialektnoyazykovogo smesheniya (na materiale tyurkskogo govora der. Eushta Tomskogo rayona* [On the patterns of dialectal language mixing (on the material of the Turkic dialect of the village of Eushta, Tomsk Oblast)]. Philology Cand. Diss. Tomsk.