Орехов Б. В. Проблемы морфологической разметки башкирских текстов // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. — Казань: Изд-во «Фэн» Академии наук РТ, 2014. — С. 135—140

ПРОБЛЕМЫ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ БАШКИРСКИХ ТЕКСТОВ

Б. В. Орехов
Национальный исследовательский университет «Высшая школа экономики»

В статье анализируются проблемы, связанные с морфологической аннотацией Башкирского поэтического корпуса. Формальный подход к грамматической аннотации, реализованный в Корпусе, требует пересмотра и адаптации инвентаря грамматических категорий, представленных в академических грамматиках башкирского языка.

Создание систем автоматического анализа морфологии тюркских языков — насущная задача, актуальная вовсе не формально. Задача эта прежде всего прикладная, но тесно связанная с теоретической областью. Во-первых, именно на теоретических описаниях строятся программные средства, а во-вторых, исследования, которые можно провести на уже реализованных автоматических системах могут серьёзно скорректировать наши теоретические представления о языке. То есть иногда язык оказывается вовсе не таким, как его описывают лингвисты.

В основе любого автоматического анализа морфологии лежат две составляющие: грамматический словарь и формализованное описание словоизменительной системы языка. На первый взгляд, башкирская лингвистика находится в этом отношении в выигрышном положении, так как обе эти задачи уже были ею какое-то время назад решены.

Лучше всего, как казалось, дело обстоит с формальным описанием словоизменения. Действительно, реализованная З. А. Сиразитдиновым модель порождения форм в башкирском языке отличается стройностью, упорядоченностью и последовательностью. Автор представляет основу изменяемых частей речи как элемент тензора [4: 20] и, отталкиваясь от этого, расписывает каждый вариант присоединения цепочки аффиксов. Это серьёзная теоретическая работа. Однако для наших практических целей у сделанного З. А. Сиразитдиновым описания два ощутимых недостатка. Во-первых, представленная модель слишком сложна и

объёмна, в цитированной книге она занимает порядка 80 страниц, что, конечно, заставляет искать более эффективные решения для программной основы морфологического анализатора. Во-вторых, моделированная З. А. Сиразитдиновым грамматика имеет порождающий характер и в своих практических перспективах имеет в виду прежде всего создание спелл-чекера (программы проверки орфографии). Интересующий нас анализ имеет обратное порождению направление. Переформатирование предложенной Сиразитдиновым модели под наши задачи было бы, в принципе, возможно, но потребовало бы дополнительных усилий, представляющихся чрезмерными в силу упомянутой в первом пункте выше сложности и объёмности грамматического описания.

Гораздо хуже обстоит дело с грамматическим словарём. Для начала скажем, что объём словаря М. Х. Ахтямова [6] не соответствует современным запросам. 20000 лексем, включая заимствования, — это чрезвычайно мало для построения адекватной современным реалиям системы анализа. Кроме того, в словнике мы обнаруживаем высокий процент малоупотребительных слов вроде *авиапорт*, *агронимия*, *мальтузиансылыҡ* и историзмов, таких как *ленинсы*. Отметим также, что в кратком грамматическом очерке, предваряющем словарь указывается на разницу в словоизменительном поведении исконных башкирских и заимствованных основ. Однако в самом словаре это различие никак не отражено, что, конечно, существенно затруднило бы практическое применение этого лексикографического опыта в компьютерных системах.

Но главная претензия к «Грамматическому словарю башкирского языка» в том, что это крайне ненадёжный источник. В нём находятся десятки опечаток, причём такого свойства, которые заставляют усомниться не только в том, что рукопись побывала в руках хорошего редактора, но и в том, что весь объём словаря был создан одним автором. Тот вид, в котором появляются в словнике некоторые заимствования из русского, приводят к мысли, что часть работы была выполнена малограмотными помощниками. Приведём избранный список таких ошибок с указанием страниц: *форфор* (с. 66, следом за словом фарфор), *дисерт* (с. 67), *дефтонг* (с. 71), *гравёра* (с. 71, почему-то форма род. п. выдана за начальную), *маникен* (с. 75), *кононир* (с. 76), *мородёр* (с. 76, ср. правильное *мародёрлыҡ* на с. 104), *билитёр* (с. 76), *архидея* (с. 80), *галерия* (с. 80), *клодовой* (с. 82), *мередиан* (с. 87), *ленолиум* (с. 87), *кординал* (с. 87), *аппонент* (с. 90), *делитант* (с. 90), *гоноррея* (с. 92), *абонимент* (с. 99), *храноґраф* (с. 99), *аранжерия* (с. 100), *рефлекция* (с. 101), *бутофория* (с. 101), *форресплав* (с. 102, очевидно, имелся в виду *ферросплав*), *браковчица* (с. 102), *фармоколог* (с. 102), *пиниципиллин* (с. 105), *инквизатор* (с. 105), *рецедивист* (с. 106), *велиончель* (с. 106), *дезиртирлыҡ* (с. 109), *суверинитет* (с. 110), *гроссмейстр* (с. 110), *бомдардировка* (с. 115, это ошибочное написание даже выделено полужирным), *травмотология* (с. 116), *аккомпонимент* (с. 117).

Итак, проблему формального описания словоизменительной системы башкирского языка пришлось решать заново. Формальный подход с прицелом на практический результат имеет другую идеологическую платформу, нежели подход академический, поэтому в процессе нам пришлось столкнуться со множе-

ственными несоответствиями академического описания (например, в [1]). Теоретическая наука, к примеру, склонна рассматривать прилагательные и наречия в башкирском языке как самостоятельные части речи. Однако такое разграничение проводится исключительно по семантическим (и отчасти словообразовательным) признакам, в то время как с точки зрения словоизменения разницы между этими классами лексем не наблюдается. Очевидно, что на такое разделение повлияла традиция описания европейских языков, где разница между прилагательными и наречиями выражена морфологически. О таком эффекте в своё время писал Л. В. Щерба: «Так, грамматики большинства известных нам языков находятся под тем или другим влиянием латинской грамматики, от которой они лишь с великим трудом и только очень постепенно освобождаются. В этих условиях я позволю себе утверждать, что при изучении языков у подавляющего большинства лингвистов получается смешанное двуязычие и изучаемый язык в той или иной мере воспринимается ими в рамках и категориях родного. В связи с этим особенности структуры изучаемых языков или стираются, или фальсифицируются» [5: 41]. Особое внимание академик Щерба как раз призывал уделить грамматикам народов СССР, которые, по его мнению, не должны быть похожи на русскую грамматику: «грамматики разных национальных языков в пределах Союза [ССР] должны прежде всего сбросить с себя иго русской грамматики. Грамматика и словарь каждого языка должны быть составлены совершенно независимо от других языков и вовсе не должны представлять из себя сколка с латинской, немецкой, русской грамматики или словаря» [5: 318]. Поставленная проблема продолжает оставаться актуальной до последнего времени: «Раньше лингвисты использовали "универсальную" грамматику латинского таким образом, теперь используют "универсальную" грамматику английского. Если мы непреклонно настроены находить наши универсалии в каждом языке, мы без сомнения будем универсально удачливы — можно указать хотя бы аллофонные правила для глухих шумных в испанском (...). Этот "успех", нам кажется, скорее обнаруживает нашу непреклонность, чем эмпирические особенности всех языков. Другими словами, это все еще аналитическая истина в кантианском смысле, но не синтетическая» [3: 25].

В процессе работы над башкирским морфологическим анализатором у нас сформировалось прикладное представление о составе грамматического словаря. Например, дублирующие основы у существительных и прилагательных зачастую создают и дублирование разборов (вследствие повторения схем разборов этих частей речи). Поэтому мы включаем в словарь основ прилагательных только качественные, т. е. потенциально могущие иметь специфические для прилагательных словоизменительные формы сравнительной степени. Выделение модальных и звукоподражательных слов в отдельную категорию также не до конца очевидно и требует, наш взгляд, дополнительной аргументации.

Упомянутый выше печатный грамматический словарь склонен дублировать лексемы в разделах существительных и прилагательных в том случае, если существительное в ряде контекстов выступает в определительном значении. Однако

такой подход на уровне прикладного грамматического словаря кажется нам неверным и создаёт лишние омонимические сущности при разборе. На наш взгляд, из раздела прилагательных все дублирующие (идентичные по форме и близкие по значению) субстантивные основы должны быть убраны и оставлены только в разделе существительных.

Отдельного разговора заслуживают и другие случаи экстраполяции русскоязычных грамматических привычек на башкирскую морфологию. Так, если русским эквивалентом слова *күпме* является «сколько», то это вовсе не значит, что и в башкирском эта словоформа должна считаться местоимением.

Наконец, приходится признать и неполноту академического грамматического описания, обнаруживаемую при сплошном анализе текстов. Так, в академической грамматике есть аффиксы деепричастий -майынса/мәйенсә, но нет -майса/мәйсә (см. у поэта С. Габидуллина *тайшанмайса*). В академической грамматике аффикс -гәсә понимается как осложнённый аффикс местоимений, то есть использование этого аффикса для словоизменения существительных не может быть объяснена. Тем не менее, контексты демонстрируют хотя и редкое, но функционирование упомянутого аффикса в составе форм существительных: *парижгәсә*.

В итоге, при морфологическом разборе башкирских текстов анализатором bashmorph принята следующая аннотация (помета, международное соответствие, русское соответствие):

S	substantive	существительное
V	verb	глагол
ADJ	adjective	адъектив
NUM	numeral	числительное
SPRO	pronoun	местоимение
PART	particle	частица
POST	postposition	послелог
CONJ	conjunction	союз
INTJ	interjunction	междометие
PL	plural	множественное число
SG	singular	единственное число
1,2,3SG	first (second, third) person, singular	первое (второе, третье) лицо единственного числа
1,2,3PL	first (second, third) person, plural	первое (второе, третье) лицо множественного числа
UNCERT	uncertainty	клитика со значением неуверенности
UNDEF	undefiniteness	клитика со значением предположительности
INTERROG	interrogative	клитика со значением вопросительности
REQUEST	request	клитика со значением просьбы
PRED	predicative	сказуемость
POSS	possessive	принадлежность
IPOSS	impersonal possessive	неличная принадлежность
NOM	nominative case	основной падеж

GEN	genitive case	родительный падеж
DAT	dative case	дательный падеж
ACC	accusative case	винительный падеж
ABL	ablative case	исходный падеж
LOC	locative case	местно-временной падеж
ABE	abessive case	абессив
DERIV.ABSTR	abstract noun	абстрактное существительное
DERIV.AGENS	agens noun	со значением деятеля
ASSIM	assimilation	уподобление
DIST	distinctive	выделительное (числительное)
COL	collective	собирательное
ORD	ordinary	порядковое
APRX	approximative	приблизительное
DIVIS	divisive	разделительное
COMP	comparative	сравнительная степень
ACYCL	acyclic	непериодическое действие
GER	gerund	деепричастие
IMP	imperative mood	повелительное наклонение
REFL	reflexive genus	возвратный залог
PASS	passive genus	страдательный залог
RECP	reciprocal genus	взаимный залог
CAUS	causative genus	понудительный залог
NEG	negation	отрицание
PRES	present tense	настоящее время
PST.DEF	past definite tense	прошедшее определённое
PST.INDF	past indefinite tense	прошедшее неопределённое
FUT.INDF	future indefinite tense	будущее неопределённое
FUT.DEF	future definite tense	будущее определённое
DESI	desiderative mood	наклонение намерения
COND	conditional mood	условное наклонение
OPT	optative mood	желательное наклонение
INF	infinitive	инфинитив
SUP	supin	имя действия

Список литературы

1. Грамматика современного башкирского литературного языка / отв. ред. А. А. Юлдашев. М.: Наука, 1981. 496 с.
2. Орехов Б. В., Слободян Е. А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28—31 октября 2010 г.) / отв. ред. В. А. Баранов. Уфа; Ижевск: Вагант, 2010. С. 167—171.
3. Пильх Г. Язык или языки? Предмет изучения лингвиста // Вопросы языкознания. 1994. № 2. С. 5—28.

4. Сиразитдинов З. А. Моделирование грамматики башкирского языка. Словоизменительная система. Уфа: АН РБ, Гилем, 2006. 160 с.
5. Щерба Л. В. Языковая система и речевая деятельность. М.: Едиториал УРСС, 2004. 432 с.
6. Әхтәмов М. Х. Башҡорт теленең грамматика һүҙлеге. Һүҙ үҙгәреше. Өфө: «Башкортостан» нәшриәте, 1994. 216 с.