М. А. Косогорова

ОПЫТ ПОДГОТОВКИ ТЕКСТОВ ПУЛАР К АВТОМАТИЧЕСКОЙ РАЗМЕТКЕ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ[1]

Интерес исследователей-африканистов к языку пулар появился давно. Еще в середине XIX века появились первые работы европейских ученых, содержащие грамматики и словари. Затем, по мере продвижения лингвистики на новые уровни, стали появляться и сборники текстов, и более обширные и глубокие работы по грамматике. Пулар стал предметом изучения для диалектологии и типологии. Однако до недавнего времени корпусов на пулар не существовало.

Это упущение имеет несколько причин. Во-первых, сама структура языка оказалась слабо приспособленной для создания стандартно размеченного корпуса. В языке сложная морфонология: присутствуют случаи глубинной фузии, сложные аналитические формы, обширная система именных классов и специфичная система частичного отображения именного класса в типе начального корневого согласного.

Во-вторых, долгое время у языка не существовало стандартной системы письменности. Исторически первая письменность пулар использовала арабскую графическую систему – аджами. Такая система записи существует до сих пор, изучается в коранических школах и используется некоторыми носителями в повседневной жизни. Затем, с приходом европейцев, появилась другая система письма – на базе латиницы. Изначально эта система записи довольно сильно отличалась от той, что принята в 1966 году на конференции в Бамако и используется теперь, – измененной фонологической транскрипции на основе международного фонетического алфавита. Но даже в наши

[1] Работа выполнена в рамках проекта «Теоретическое и техническое обеспечение корпуса текстов на языке пулар» по программе Президиума РАН «Корпусная лингвистика».

дни, когда фонологическая система записи преподается в школах, на ней издаются газеты и книги и ее использование в целом превалирует, существуют разночтения в записи некоторых звуков. Так, можно сравнить запись одного и того же звука в разных источниках:

(1) hoɲɲ-u-gol ‘опалить шерсть’ [Bettison & Bettison 2006]
hoñ-a ‘пахнуть горелым’ [Зубко 1980]

Очевидно, что даже если обратить внимание на геминацию согласного, корень у этих глаголов общий, следовательно, один и тот же звук отображен разными способами. Таких примеров в современной литературе достаточно, чтобы назвать это системными разночтениями, а не разовыми прецедентами. Проблема разной записи решается вполне тривиально – путем унификации всех спорных случаев. В данном случае, в соответствии с конвенциями Бамако, в обоих случаях будет использован диграф ny.

И, наконец, в-третьих, язык пулар уникален своей дисперсной природой. В ходе исторических миграций фульбе, носители языка, оказались расселены очагами от Атлантического океана на западе до Голубого Нила на востоке, сохранив при этом максимально возможное единство языка. Существуют разные точки зрения на то, является ли язык пулар-фульфульде (фульфульде – название языка в его центрально-восточных диалектах) одним языком, поделенным на несколько диалектов, или это несколько близкородственных языков; однако фульбе, несомненно, причисляют себя к единому этносу. Разумеется, при таком дисперсном и широком распространении языка западный диалект фута-торо (Сенегал) будет заметно отличаться от восточного диалекта адамауа (Северный Камерун). Хотя с точки зрения типологии эти различия не достаточно значительны, чтобы определенно разделить ареал пулар-фульфульде на разные языки, но с точки зрения корпусной лингвистики эти различия принципиальны, поскольку создать единый корпус при такой диалектной вариативности очень сложно.

Тем не менее, попытка создания корпуса пулар была предпринята, и в настоящей статье описаны приемы, использованные при под-

готовке текстов к автоматической разметке с целью привести тексты пулар к единому стандарту.

Первая глобальная проблема – диалектологическая – была решена очевидным способом: для создания корпуса были взяты тексты одного диалекта, одного говора, полученные от одного носителя (диалект фута-джаллон (Гвинея), говор сану-лагорд-тарамбали, регион Лабе), что, в общем, исключает диалектное варьирование. Хотя такой прием и не вполне безупречен с лингвистической точки зрения, стоит заметить, что для начала работы он оказался действенным, а в процессе разметки выяснилось, что создать список инвариантов в рамках одного диалекта возможно. Что касается прочих диалектов, то к ним, возможно, будет применена та же тактика: ядро корпуса из одного говора, а дальше – расширение до рамок диалекта. Создавать наддиалектную норму не представляется нам необходимым, поскольку в ходе исторической дисперсии некоторые диалекты сильнее отдалились друг от друга, чем другие, и хотя для некоторого базового ядра языка возможно определить наддиалектное состояние, в остальном объем заимствований и изменений делает эту работу бесперспективной.

Однако графические затруднения на этом не заканчиваются. Большинство современных текстовых редакторов не делают разницы между графическими воплощениями отдельных звуков [m] и [b] и двухфокусного преназализованного согласного [mb], который, наряду с [ng], [ny] и др., регулярно используется в пулар. Использование же большого количества диакритических символов делает эту систему записи менее удобной для повседневного использования. В печатных словарях алфавитный порядок требует, например, использования буквы mb после m, и так далее, и технически это вполне возможно осуществить. При создании встроенных словарей эта проблема обходится стороной, то есть, неупорядоченные по алфавиту языка вхождения словоформ не являются серьезным недочетом системы, так как поиск осуществляется автоматически (и никто не видит порядка словоформ во встроенном словаре), однако в перспективе адаптация программы по разметке к графическому отображению ко-

артикулированных согласных неизбежна, поскольку они широко используются в системе чередования начальных согласных корня.

Следующая задача, встающая перед любым создателем корпуса языка с еще не устоявшейся письменной традицией, уже не столь проста. Она заключается в приведении орфографии текстов к единому стандарту, что значительно упрощает работу парсера, хотя полностью привести язык к одному стандарту орфографии, разумеется, не требуется.. Для исследователя пулар эта задача усложняется, поскольку в языке используется система смыслоразличительных долгот и геминаций, выполняющая также ритмическую функцию. Такое явление непривычно европейскому уху, а для носителя оно является само собой разумеющимся, поэтому обоим требуется длительная тренировка, чтобы суметь различать долготы. Сравним, например, простой случай:

(2) waala ‘проводить ночь’
walla ‘помогать’

Простота этого случая – не только в реактивной геминации заударного согласного (геминированный согласный подчеркивает краткость предыдущего гласного), но и в том, что интересующий нас гласный находится в корне, поэтому он смыслоразличителен, и можно методом формоизменения установить его долготу. Однако в других случаях разрешить этот вопрос не так легко: в языке на законных основаниях могут существовать две формы одной лексемы, в которых фузионные преобразования затронули корень. В таких случаях приходится указывать оба варианта во встроенном словаре, а затем указывать связь между формами в формате постраничных сносок, как это предложено в издании [Малые языки и традиции 2008]. Но намного сложнее установить долготу гласного в глагольном видо-залоговом показателе.

(3) a.	o	yehi		ka	suudu	
	o	yeh-	-i	ka	suu-	-du
	3.sgO	идти-	-Act.Pfv.w	Prep	дом-	-sgNDU
	‘Он пошел домой’.					
b.	o	yehii		ka	suudu	
	o	yeh-	-ii	ka	suu-	-du
	3.sgO	идти-	-Act.Pfv.s	Prep	дом-	-sgNDU
	‘Он [уже] ушел домой’.					
(4) a.	himo	mari		barehun		
	himo	mar-	-i	bare-	-hun	
	Cop.3.sgO	иметь-	-Act.St	собака-	-sgKUN	
	‘Он [был] обладавший собачкой’.					
b.	o	mari		barehun		
	o	mar-	-i	bare-	-hun	
	3.sgO	иметь-	-Act.Pfv.w	собака-	-sgKUN	
	‘Он имел собачку’.					

Различие, приведенное в примерах 3a и 3b, может показаться незначительным, однако оно принципиально. Глагол пулар изменяется по залогам (Act, Md, Pass) и по видам, перфективному (Pfv) и потенциалисному (Pot). Виды, в свою очередь, имеют регистры: сильный, с оттенком результативности, и слабый. Подробнее о глагольной парадигме можно узнать из работы [Коваль 2003]. Принципиальным фактом является омонимичность показателей: во многих случаях она разрешима лишь семантически или контекстуально, см., например, пример 4, где показатель -i выражает статив лишь в сочетании со специальной формой местоимения, а при изменении подлежащего глагольная форма определяется как слабый перфектив, хотя ее вид не изменяется.

В примере 3, однако, форма различна, однако слаборазличительная, финальная позиция маркирующего гласного не всегда позволяет с уверенностью определить долготу. Существуют дополнительные факторы, помогающие определить качество показателя, например, в придаточном условия используется сильный регистр, но в отсутст-

вие таких факторов решение принимается исследователем при расшифровке аудиозаписи.

Систематизация долготы звуков – довольно сложная задача, часто решаемая (при отсутствии авторитетного мнения носителя языка) статистическим методом, с помощью комбинации мнений различных словарей диалекта. В случае, если эти данные недоступны, приходится принимать независимое решение, а затем придерживаться его на протяжении всей работы.

Следующий этап подготовки текста на языке пулар к автоматической разметке полностью подчинен технической стороне вопроса. Исходный текст, помимо аффиксов, выраженных графически, содержит также нулевые аффиксы, местоположение которых программа сама определить не в состоянии. Речь идет, например, о неличном подклассе класса O, объединяющем в себе многие заимствования. Как следует из названия, у класса есть и второе, личное воплощение, которое объединяет в себе имена людей, профессий и т. д. Подклассы класса O различаются, впрочем, не только по семантическому признаку. Личный O-подкласс оформляется материальным показателем соответствующей ступени, неличный же оформляется нулевым аффиксом, тем не менее согласуясь по классу O (см. пример 5).Также нулевой аффикс регулярно используется в глагольной словоформе. Чтобы отобразить значимые нули в размеченном тексте, исследователь должен добавить их в исходный вариант текста, а после использования программа удалит их.

(5) a.	nyaametee			on
	nyaam-	-etee-	-∅	on
	есть-	-Pass.Pfv[PCP]-	-sgO	Def.sgO
	‘еда’			
	nyaameteeji			dì
	nyaam-	-etee-	-ji	dì
	есть-	-Pass.Pfv[PCP]-	-plDI	Def.plDI
	‘[много видов] еды’			

b.	anniya		makko	on
	anniya-	-∅	makko	on
	желание-	-sgO	Poss.3.sgO	Def.sgO
	‘ее желание’			

Доказательств существования нулевого аффикса в этой позиции можно привести два. Во-первых, структура имени предопределяет его наличие, и он появляется при попытке, например, поставить имя во множественное число (см. пример 5a). Во-вторых, при согласовании имен неличного O-подкласса атрибут, который должен дублировать показатель класса вершины, всегда получает показатель класса O (см. пример 5b). Таким образом, «для диалекта фута-джаллон, как и для других западных диалектов, может быть сформулировано правило: если ПК [показатель класса] =∅, то имя будет принадлежать к согласовательному классу O» [Коваль 1979 : 11].

Полуавтоматический способ разметки нулей, представленный выше, представляется нам наиболее экономичным с точки зрения затрат ресурсов и в то же время наиболее информативным. В нем гармонично сочетаются использование знаний исследователя и автоматические действия, исключающие возможность случайной ошибки.

Следующая проблема, встающая перед исследователем, – это проблема разделения текста на отрывки. Она отнюдь не уникальна и имеет несколько решений. Одним из них считается разделение текста на элементарные дискурсивные единицы – минимальные кванты дискурса, которые невозможно подразделить. Это разделение, без сомнения, наиболее подробно, однако в рамках корпуса, целью которого не являются специализированные дискурсные исследования, мы считаем эту подробность излишней. Другой вариант – разделение текста на элементарные предложения. Такая тактика оставляет оптимальный объем фразы, при этом позволяя глоссировать достаточно длинные тексты. Однако при всех плюсах названных критериев разбиения, существует один существенный минус: из-за особенностей языка такая тактика не всегда реализуема. Одной из этих особенностей, в частности, является субъектная анафора. Это правило дейст-

вует на общефульском уровне и позволяет использовать в сложносочиненном предложении нулевую анафору в случае, если при первом предикате используется полнолексемный субъект (см. пример 6a). В случае, если эта позиция заполнена субъектным местоимением, то при дальнейших повторах в том же предложении необходим повтор этого местоимения (см. пример 6b).

(6) a. paykun kun ƴetti nyiiri
pay- -kun kun ƴett- -i nyii- -ri ndin
ребенок- -sgKUN Def.sgKUN брать- -Act.Pfv.w каша- -sgNDI

ndin ∅ yehi ∅ jasi
∅ yeh- -i ∅ jas- -i
Def.sgNDI 3Sg идти- -Act.Pfv.w 3Sg копать- -Act.Pfv.w

ngayka
ngay- -ka
яма- -sgKA
‘Мальчик взял кашу, пошел, вырыл яму…’

b. o immii o yalti ka yaasi
o imm- -ii o yalt- -i ka yaasi
3.sgO встать- -Md.Pfv.w 3.sgO выходить- -Act.Pfv.w Prep наружу
‘Она встала, она вышла на улицу…’

Возможно, последовательным решением было бы указывать нулевую анафору в предложении и делить текст на элементарные предложения. Тем самым из примера 6a получилось бы три отрезка, а из примера 6b – две. Однако мы находим это решение недостаточно экономичным, и к тому же не очень удачным с точки зрения возможного пользователя-синтаксиста, который предпочел бы, чтобы предложения не разбивались на более мелкие части. Таким образом, остается третий вариант разделения текста – на предложения. Такое разделение имеет очевидный недостаток: в зависимости от жанра предложения могут быть достаточно объемными, что затруднит поиск, то есть именно то, ради чего в итоге и производится деление текста на базовые единицы. Однако при переносе текстов с аудионо-

сителя в бумажный вариант существует возможность в ряде случаев (хотя и не везде) искусственно регулировать длину предложения, что нивелирует указанный недостаток. В остальном же деление текста на предложения является, на наш взгляд, оптимальным: каждый отрезок текста получается грамматически и семантически цельным, и общее их число не превышает порога адекватности, после которого использование излишне члененного текста станет неудобным.

Как уже указывалось ранее, в примере 4а стативная конструкция в предложении возможна лишь при условии использования определенного разряда местоимения. Это подводит нас к следующей проблеме, которую в перспективе придется поставить перед собой создателям корпуса пулар, – проблеме аналитических форм.

В текущей версии корпуса аналитические формы не отображаются в разметке, что может навести пользователей на ложную мысль о полной омонимии, например, форм слабого перфектива и глагольной формы в стативе. Это не так: статив образуется аналитически, с помощью формы слабого регистра и глагола-связки (в футаджаллоне эта связка имеет вид no), который может присоединяться к специальным (т.н. копулосодержащим) местоимениям, а в случае полнолексемной реализации субъекта использоваться отдельно (см. пример 7).

(7)	debbo		on	no	faalaa		warde		
	debb-	-o	on	no	faal-	-aa	war-	-∅-	-de
	женщина-	-sgO	Def.sgO	Cop	хотеть-	-Pass.St	убить-	-Act-	-Inf

paykun		kun
pay-	-kun	kun
ребенок-	-sgKUN	Def.sgKUN

‘Женщина хотела убить мальчика’.

Аналитические формы не ограничиваются лишь стативом и дуративом: также аналитически выражается оптатив, изменение фокуса контраста, некоторые виды отрицания и др. Это довольно распространенное в пулар явление, и игнорирование таких конструкций может быть лишь временным уходом от решения проблемы. Вариант

решения существует (пример 8), и его реализация представляется неизбежной, возможно, в несколько видоизмененной форме. Это решение заключается в использовании системы микро-сносок, где индексом служит «звездочка» (как наиболее частотный символ для таких целей), а при финальном элементе конструкции вместо указания регистра помещена расшифровка.

(8)	barehun		kun	no	humpitii	
	bare-	-hun	kun	no	humpit-	-ii
	собака-	-sgKUN	Def.sgKUN	*Cop	*иметь.сведения-	-Md.Pfv{*St}
	‘собачка знает’					

Однако такое решение также требует разметки текста на этапе подготовки – отметку о том, что две словоформы реализуются в комплексе, составляя аналитическую конструкцию, необходимо представить программе в момент первичного обращения к тексту; впоследствии она, как и в ситуации с нулевыми аффиксами, будет удалена.

И еще одна проблема, которая до сих пор не была решена в рамках корпуса, однако решение которой необходимо, – это проблема разметки ступеней начальных согласных корня. Система изменения анлаута в именах существительных и их атрибутах, функционирующая с большей или меньшей эффективностью во всех диалектах языка, представляет собой сложный механизм, который связывает аффиксальный показатель класса со ступенью начального корневого согласного. Согласные, задействованные в этой системе, организованы в три подгруппы: щелевые, чистые смычные и преназализованно-смычные (в случае глухого согласного две последние подгруппы совпадают), – и каждый именной класс требует, помимо определенного аффиксального показателя класса, также использования определенного вида согласного в корне. Более того, в именной группе согласование последовательно реализуется путем уподобления в атрибуте обоих параметров вершины. Эта гармоничная система, однако, технически трудно реализуема в условиях корпуса. Тот факт, что в диалекте фута-джаллон, с которого началась работа над корпусом,

эта система потеряла продуктивность и почти не используется, является исторической случайностью, однако благодаря ей решение этой проблемы стало возможным отложить.

Принципиальной деталью отображения этой системы в корпусе является указание ступени анлаута не в строке поморфемного перевода, а в строке морфемного членения, что является необычным с точки зрения разметки текстов (поскольку, строго говоря, не является морфемным разбиением), однако с точки зрения языка представляется логичным (пример 9). Технически это несложно реализовать, однако это потребует указания необходимой ступени анлаута на этапе подготовки текста исследователем, поскольку в некоторых диалектах система работает нестандартным способом, например, не включает в себя заимствованную лексику. Такие особенности исключают возможность полного доверия программе.

(9)	paykun		kun	ƴetti		nyiiri		
	^{Occ}pay-	-kun	kun	ƴett-	-i	nyii-	-ri	
	ребенок-	-sgKUN	Def.sgKUN	брать-	-Act.Pfv.w	каша-	-sgNDI	

ndin	yehi		jasi		ngayka		
ndin	yeh-	-i	jas-	-i	^{Pren}ngay-	-ka	
Def.sgNDI	идти-	-Act.Pfv.w	копать-	-Act.Pfv.w	яма-	-sgKA	

‘Мальчик взял кашу, пошел, вырыл яму’.

Гораздо бóльшую сложность представляет реализация связи между аффиксальным показателем класса и ступенью анлаута. Необходимость демонстрации этой связи желательна, однако технического решения этой проблемы пока нет.

Подводя итог, можно разделить список задач при подготовке текстов к автоматической разметке на три группы. В первую, собственно лингвистическую, входят задачи по унификации графики и орфографии текстов, принятие и последовательное соблюдение решений о долготе гласных и геминации согласных и определение сложных форм (аналитических, "нулевых", а также ступеней анлаута) для последующей обработки программой. Во вторую, техническую, группу входят такие задачи, как корректное маркирование

сложных форм, разбиение текста на более элементарные единицы и техническая проверка текста. И, наконец, третья группа задач – нерешенные: их решение необходимо в перспективе, однако в данный момент без него можно обойтись.

Таким образом, можно считать, что главная задача предварительной подготовки текста – корректное функционирование программы-парсера, не требующее дельнейшего существенного вмешательства специалиста – успешно выполняется. При этом баланс между качеством выполненной работы и затраченными ресурсами сохраняется, что, безусловно, свидетельствует об успехе данного этапа разработки корпуса языка пулар.

Сокращения

3sgX – «третье» лицо – местоимение сингулярного именного класса X
Act – активный залог
Cop – копула, предикативная связка
Def – определенность
Inf – инфинитив
Md – медиальный залог
Occ – смычная ступень начального согласного
Pfv – перфектив, перфективный вид
Pfv.s – сильный перфектив
Pfv.w – слабый перфектив
Pl – множественное число
Prcn – преназализованно-смычная ступень начального согласного
Prep – предлог, предложно-союзное слово
Pass – пассивный залог
Sg – единственное число
sgKA, sgNDI, sgNDU ... – сингулярные именные классы
sgKUN – диминутивно-сингулярный именной класс
sgO – (лично-)сингулярный именной класс
St – статив

Литература

Bettison Jim & Karen. Dictionnaire Pular-Francais. 2e edition. Labe: Traducteurs Pionniers de la Bible, 2006.

Коваль А.И. О значении морфологического показателя класса в фула // Морфонология и морфология классов слов в языках Африки. Имя, местоимение. М: Наука, 1979. С.5-100.

Коваль А.И. Контрастивность как морфологическая глагольная категория в пулар-фульфульде // Основы африканского языкознания. Глагол. М.: Восточная литература, 2003.

Зубко Г.В. Фула-русско-французский словарь. М.: Русский язык, 1980.

Малые языки и традиции: существование на грани. Вып. 2. Тексты и словарные материалы. Под ред. А.Е. Кибрика. М., 2008.