На правах рукописи

КОСОГОРОВА МАРИЯ АЛЕКСАНДРОВНА

ПРИНЦИПЫ ГЛОССИРОВАНИЯ ДЛЯ КОРПУСА МЛАДОПИСЬМЕННОГО ЯЗЫКА: МОРФОЛОГИЧЕСКАЯ СТРУКТУРА ЯЗЫКА ПУЛАР

Специальность 10.02.21 – Прикладная и математическая лингвистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

Москва – 2012

Работа выполнена в отделе африканских языков
Федерального государственного бюджетного учреждения науки
Института языкознания Российской академии наук

НАУЧНЫЙ РУКОВОДИТЕЛЬ:

кандидат филологических наук,
ведущий научный сотрудник отдела африканских языков
ФГБУН «Институт языкознания РАН»
Антонина Ивановна Коваль

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:

доктор филологических наук, профессор,
профессор кафедры африканистики
Восточного факультета ФГБОУ ВПО
«Санкт-Петербургский государственный университет»
Валентин Феодосьевич Выдрин
кандидат филологических наук,
профессор факультета филологии
ФГАОУ ВПО НИУ «Высшая школа экономики»
Михаил Александрович Даниэль

ВЕДУЩАЯ ОРГАНИЗАЦИЯ:

ФГБОУ ВПО «Московский государственный университет
имени М. В. Ломоносова», кафедра теоретической
и прикладной лингвистики филологического факультета

Защита диссертации состоится «___» _____ 2012 года в ____ часов
на заседании диссертационного совета Д 002.006.03 при Институте
языкознания РАН по адресу: 125009 Москва, Б. Кисловский пер., д.1, с.1.

С диссертацией можно ознакомиться в читальном зале библиотеки Института языкознания РАН.

Автореферат разослан «___» _____ 2012 г.

Ученый секретарь
диссертационного совета /А.В. Сидельцев/

Общая характеристика работы

Объектом исследования являются актуальные проблемы, возникающие при создании текстового аннотированного корпуса младописьменного языка (на материале пулар). Создание корпусов текстов – это важнейшая задача современной лингвистики, поскольку они предоставляют исследователям практический материал для работы с различными языками мира.

Основная задача, которая стоит перед создателями глоссированного корпуса, – это информативная и экономичная разметка текстов, позволяющая осуществлять поиск по корпусу с максимальной точностью. Разработка качественного инвентаря глосс подразумевает детальное рассмотрение грамматических и лексических особенностей языка.

Язык пулар-фульфульде, данные которого выбраны в качестве эмпирического материала для исследования, относится к западно-атлантической языковой семье и является одним из крупнейших африканских языков. Он имеет много типологических особенностей, которые представляют значительный интерес для исследователей. Характерными особенностями языка, обеспечивающими ему неизменный интерес со стороны лингвистов, являются (помимо его масштабов, дисперсии и социолингвистических данных), в частности, сложная система именной классификации, процессы актантных преобразований в трёхзалоговой системе, связь фокализации со словоизменительной парадигматикой глагола, а также система чередований начальных согласных корня. Некоторые из этих типологических особенностей встречаются в других языках мира. Анализ этих особенностей и последующая разработка системы глоссирования для пулар-фульфульде может послужить опорой для создания аналогичных систем для других языков. Пример (1) предлагает три варианта глоссирования одной из таких особенностей – рамочной выделительной конструкции *[ko ... kon]*.

(1)

a. Минималистичный вариант: глоссирование без указания аналитической связи между элементами

o	faalaama		wadde	**ko**	o	faalaa		**kon**
o	faal-	aama	wadde	**ko**	o	faal-	aa	**kon**
3.sgO	хотеть-	Pass.Pfv.s	делать.Act.Inf	**Rel**	3.sgO	хотеть-	Pass.Pfv	**Def**

Она захотела сделать то, что хотела.

б. Вариант с неполным указанием связи между элементами

o	faalaama		wadde	**ko**	o	faalaa		**kon**
o	faal-	aama	wadde	**ko**	o	faal-	aa	**kon**
3.sgO	хотеть-	Pass.Pfv.s	делать.Act.Inf	**Rel**	3.sgO	хотеть-	Pass.Pfv	**Def.Rel**

Она захотела сделать то, что хотела.

в. Вариант с полным указанием связи между элементами и её типа

o	faalaama		wadde	**ko**	o	faalaa		**kon**
o	faal-	aama	wadde	**ko**	o	faal-	aa	**kon**
3.sgO	хотеть-	Pass.Pfv.s	делать.Act.Inf	***Rel**	3.sgO	хотеть-	Pass.Pfv	***Def{*Frame}**

Она захотела сделать то, что хотела.

Как следует из примера, существует несколько, на первый взгляд равноценных возможностей аннотации рамочной конструкции. При выборе варианта, который будет использован в корпусе пулар, учитываются, главным образом, такие лингвистические факты, как контекст, функциональность, возможные случаи омонимии и вариативности, а также возможность технической реализации выбранного решения.

Актуальность исследования определяется тем, что на современном этапе построение и развитие языковых корпусов находится в русле важнейших задач, востребованных современным состоянием языковедения, в том числе и для . бесписьменных или младописьменных языков. Создание, на примере пулар, принципов морфоглоссирования таких языков является актуальной исследовательской задачей, так как представляет собой необходимый этап собственно перед созданием корпуса.

Целью исследования является разработка и реализация системы глоссирования для языка пулар-фульфульде с учётом типологических особенностей этого языка.

Поставленная цель определяет следующие конкретные задачи исследования:

– определение проблемных с точки зрения глоссирования фрагментов грамматики языка;

– описание общих подходов, целесообразных для аннотирования текстов на языке;

– инвентаризация конкретных явлений, требующих индивидуального подхода, на синтаксическом, лексическом, морфологическом и фонологическом уровнях языка;

– определение конвенций глоссирования для каждого случая;

– принятие решения о возможности технической реализации предложенных конвенций.

На защиту выносятся следующие положения.

1. Существует теоретическая и техническая возможность реализации аннотированного корпуса языка пулар.

2. Хотя морфологические особенности синтетического языка зачастую не позволяют создать адекватной аннотации в рамках синхронного корпуса, существуют способы их отображения в глоссировании, которые могут считаться удовлетворительными.

3. В случае, когда создаётся система глоссирования для типологически редкого языка, целесообразно в релевантных случаях дублировать необходимую информацию, что в иных ситуациях было бы избыточным.

Научная новизна исследования состоит в том, что здесь впервые создаётся последовательная система глоссирования для языка пулар с учётом его типологических особенностей. Впервые разобраны способы морфологического аннотирования таких явлений как, в частности, классный циркумфикс, аналитические конструкции, надклассные местоимения.

Теоретическая значимость исследования определяется тем, что в ходе работы были рассмотрены возможные варианты решения проблем, неизбежно возникающих при попытке создать морфологическую аннотацию текстов на языке синтетической природы с элементами аналитизма. Также в ходе работы были выделены теоретико-грамматические свойства языка, формирующие ту платформу, с опорой на которую выдвигаются и апробируются соглашения по представлению корпусных данных. Результаты исследования могут также быть применены для аналогичных явлений в других языках.

Практическая значимость настоящей работы состоит в том, что, во-первых, её результаты применены для создания аннотированного корпуса языка пулар, что делает этот язык более доступным для лингвистических исследований, в том числе квантитативных и квалитативных. Во-вторых, результаты работы могут быть использованы для разработки аннотированных корпусов других младописьменных языков. Наконец, результаты работы могут найти применение при подготовке учебных курсов по общей морфологии, корпусной лингвистике, по африканскому языкознанию, а также при разработке практических грамматик, нацеленных на преподавание языка пулар-фульфульде.

Основным материалом исследования послужили текстовые данные, собранные в ходе лингвистической экспедиции в Республику Кот д'Ивуар под руководством В.Ф. Выдрина (январь-май 2010 года). Работа по сбору текстов проводилась с носителем языка пулар фута-джаллон, говор Сану-Лагорд-Тарамбали (агломерация деревень к северу от Лабе, Республика Гвинея).

Тексты, изначально записанные в формате .WAV, были затем расшифрованы с помощью этого же носителя языка. В 2011 году для обработки собранных текстов пулар была разработана программа-парсер, направленная на морфоглоссирование данных с учётом лингвоспецифики языка. В результате этой работы в 2011 году был создан тестовый вариант корпуса пулар.

Апробация работы. Основные положения диссертации были представлены на XXV Международной конференции по источниковедению и историографии стран Азии и Африки «Востоковедение и африканистика в диалоге цивилизаций» (Санкт-Петербург, 22-24 апреля 2009 г.), на XII Конференции африканистов «Африка в условиях смены парадигмы мирового развития» (Москва, 24-26 мая 2011 г.), а также на XXVI Международной конференции «Модернизации и традиции» (Санкт-Петербург, 20-22 апреля 2011 г.), на Восьмой конференции по типологии и грамматике для молодых исследователей (Санкт-Петербург, ноябрь 2011 г.). Работа прошла обсуждение в отделе африканских языков ФГБУН «Институт языкознания РАН».

Структура работы. Работа состоит из введения, трёх глав, двух приложений и библиографии, насчитывающей более 100 наименований отечественных и зарубежных работ. Объём работы (без приложений) составляет 162 страницы машинописного текста.

Основное содержание работы

Глава 1. Обзор литературы и общая информация

Пулар-фульфульде – это один из самых известных языков Африки. Он относится к западно-атлантической языковой семье (северная группа, сенегамбийские языки), но территориально выходит далеко за пределы её основного ареала. Кочевое скотоводство, которое традиционно являлось основным ремеслом носителей языка – фульбе, обусловило распространение этого языка на весьма обширной территории к югу от Сахары, ограниченной Атлантическим океаном с запада и Голубым Нилом с востока. Особенностью языка является то, что он расположен в своём ареале не континуально, а дисперсно, занимая кое-где обширные области, а где-то образуя небольшие фулаязычные анклавы. В настоящее время можно выделить три диалектные зоны языка – западную (Сенегал, Гвинея), центральную (Мали, Буркина-Фасо, Нигер) и восточную (Нигерия, Камерун, Чад, Судан). Полидиалектные корпусные исследования пулар-фульфульде интересны отчасти из-за сложности

происходивших в течение многих веков миграций фульбе и сложной системы взаимоотношений между разными группами этого и других народов, получившейся в результате, а отчасти из-за того, что, по мнению исследователей, фульбе от Сенегала, Гамбии и Гвинеи до Нигера и Камеруна говорят на одном языке.

Исследование диалектов языка пулар-фульфульде имеет более чем полуторавековую историю, начиная с первых опытов в грамматических описаниях и до капиталистических трудов, оказавших влияние на дальнейшее развитие фуланистики (А. Гаден, А. Лабуре, А. Клингенхебен, Д. Арнотт, Г.В. Зубко, Д. Ноа, П. де Вольф и др.) К настоящему времени создано значительное количество грамматик и разномасштабных словарей, специализированных исследований по грамматике, лексике, диалектологии. С материалами языка пулар-фульфульде работали также известнейшие африканисты как К. Майнхоф и Д. Вестерманн, а также теоретики языка и типологи (Э. Сепир, Дж. Гринберг, И.А. Мельчук, Г. Корбетт, В.А. Плунгян). Существование доступного аннотированного корпуса текстов на этом языке обеспечит данными по языку гораздо более широкий круг исследователей.

С этой целью было принято решение о создании такого корпуса, в тестовом варианте включающего в себя тексты лишь одного диалекта. Формируемый корпус можно квалифицировать как имеющий устное происхождение, поскольку, хотя для языка пулар и была разработана письменность на базе латиницы, устная форма существования всё же имеет более важную роль и пока является гораздо более надёжным источником текстов, нежели письменная. Корпус содержит тексты преимущественно фольклорного жанра: в его жанровый состав входят сказки (около половины всех текстов), социокультурные зарисовки из жизни информанта (менее половины всех текстов), а также присказки и сказания (несколько текстов).

В ходе первичной обработки собранных текстов возникла необходимость создания независимой программы-парсера, и эта задача была со временем успешно решена. В результате была разработана программа LightParser, предназначенная для автоматизированного глоссирования текстов. В качестве входной информации парсер принимает цельные тексты и словари для их обработки. Далее пользователь может запустить как автоматический, так и пошаговый режим обработки слов тек-

ста. В случае возникновения неоднозначностей при обработке программа предоставляет пользователю возможность выбрать необходимый вариант вручную. Парсер имеет удобный пользовательский интерфейс, снабжён редактором текстов и словарей. Программа обеспечивает совместимость с другими программами, используемыми в данной области.

В перспективе предполагается включить в корпус данные по другим диалектам пулар-фульфульде, что, разумеется, потребует доработки парсера, а также изменения словаря. Также в перспективе планируется включение в состав корпуса строки перевода на другие языки: в данный момент рабочими языками корпуса являются пулар и русский, что осложняет его использование зарубежными исследователями.

Глава 2. Первичная подготовка текста и синтаксический уровень языка

Вторая глава посвящена проблемам, которые возникают при глоссировании и связаны с синтаксическим уровнем языка, а также с характерными для языка явлениями, находящимися на стыке синтаксиса и морфологии. В текущих технических условиях эти проблемы решаются ещё на этапе предварительной подготовки текста, на данный момент ручной, в перспективе – автоматической. Разумеется, перечисляются лишь основные решения, менее же значительные вопросы остаются за рамками обсуждения.

Сегментирование текста на клаузы

Вопрос последовательной сегментации текстов на небольшие отрезки должен решаться с учётом синтаксических особенностей языка. Язык пулар-фульфульде допускает значительную длину предложений, поэтому стандартное решение, где каждый отрезок равен одному предложению, не является оптимальным для текстов на этом языке. Однако формат корпуса, в котором потенциальный пользователь будет использовать синтаксические данные, делает невозможным и другое крайнее решение – разделение текста на элементарные дискурсивные единицы, поскольку в таком случае синтаксическая структура предложения может стать неочевидной. С

учётом этих доводов было принято решение разделять текст на предложения с возможностью дальнейшего их подразделения на сложносочинённые и сложноподчинённые части в случае необходимости. В настоящее время технические возможности позволяют провести лишь последовательную нумерацию каждой клаузы (пример 2а), однако в перспективе планируется ввести многоуровневую нумерацию клауз (пример 2б).

(2)

а.

(28)	njaatigi	ɓeyngu	sari	inni	
(28)	njaatigi	ɓeyngu	sari	inn-	i
(28)	приятель.sgO	жена.sgNGU	заяц	говорить-	Act.Pfv.w

Приятель зайчихи сказал:

(29)	ko	hondʊ̄n		wondi		
(29)	ko	hon-	dʊ̄n	won-	d-	i
(29)	Foc	3sg.Inter-	sgDUN	быть-	Soc-	Act.Pfv.Sb

"Что у вас случилось?"

б.

(28)	njaatigi	ɓeyngu	sari	inni	
(28)	njaatigi	ɓeyngu	sari	inn-	i
(28)	приятель.sgO	жена.sgNGU	заяц	говорить-	Act.Pfv.w

Приятель зайчихи сказал:

(28′)	ko	hondʊ̄n		wondi		
(28′)	ko	hon-	dʊ̄n	won-	d-	i
(28′)	Foc	3sg.Inter-	sgDUN	быть-	Soc-	Act.Pfv.Sb

"Что у вас случилось?"

Проблема нулевой анафоры

Важная синтаксическая особенность языка связана с различием полипредикатных структур языка с нулевой и ненулевой анафорой кореферентного субъекта. В примере (3а) при первом предикате субъект имеет полнолексемную реализацию, то есть в данном случае выражен именем существительным. Такая реализация обусловливает возможность нулевой анафоры этого субъекта при последующих пре-

дикатах в рамках того же предложения. В примере (3б) позиция субъекта при первом предикате заполнена прономинальной единицей, и такое воплощение не обладает достаточной силой референции, чтобы позволить использование нулевой анафоры. Оппозиция "нулевая vs. ненулевая анафора" ставит перед разработчиками корпуса вопрос: стоит ли указывать нулевой субъект при глоссировании. Теоретически это возможно и будет иметь вид, представленный в примере (3в). Однако такое решение технически трудно реализуемо, поэтому на данный момент нулевая анафора в разметке не указывается.

(3)

а.

debbo		**on**	immii		yalti			ka	yaasi	
debb-	**o**	**on**	imm-	ii	yalt-	i		ka	yaasi	
женщина-	**sg.O**	**Def.sgO**	встать-	Md.Pfv.w	выходить-	Act.Pfv.w		Prep	наружу	

Женщина встала, вышла на двор.

б.

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 двор.

в.

debbo		**on**	immii		**Ø**	yalti			ka	yaasi
debb-	**o**	**on**	imm-	ii	**Ø**	yalt-	i		ka	yaasi
женщина-	**sg.O**	**Def.sgO**	встать-	Md.Pfv.w	**3Sg.O**	выходить-	Act.Pfv.w		Prep	наружу

Женщина встала, вышла на двор.

Повторы, оговорки и вариативность

Тексты, использованные при создании корпуса пулар, имеют звуковое происхождение, то есть они были записаны на аудионоситель, а затем расшифрованы. При расшифровке возможно было бы исключить из записи все повторы и дискурсивные маркеры, поскольку при автоматической разметке, встретив нестандартную или ошибочную словоформу, программа-парсер может допускать ошибки. Однако, при использовании точной стенограммы аудиозаписи появляется возможность ис-

пользовать эту запись в корпусе в дальнейшем. Наличие звукового сопровождения глоссированного текста позволяет проводить исследования для гораздо большего числа областей науки, в частности, фонетики, дискурса, просодии. Для этой цели в тексте сохраняются все оговорки, повторы и ошибки, но для удобства понимания они расшифровываются в строке перевода. Также для исследований в этой области сохраняются и отмечаются все хезитации и дискурсивные маркеры, однако их классификация в текущие задачи не входит.

Также принципиальное решение принято в области орфографической вариативности. Письменная традиция (на латинице) возникла сравнительно недавно, поэтому (а также из-за большой диалектной вариативности) орфографическая и произносительная норма заметно варьирует, и возникает вопрос возможной унификации. Объективные причины диктуют отрицательный ответ, и встроенный словарь программы-парсера имеет из-за этого возможность указания контекстно-вариативных лексем.

Аналитические формы

Аналитические глагольные формы – это регулярно использующиеся в языке пула-фульфульде конструкции, которые, несмотря на свою внешнюю простоту, представляют значительную трудность для автоматического глоссирования, поскольку каждый элемент большинства конструкций может использоваться и отдельно. Наиболее распространённые глагольные аналитические конструкции – это статив и прогрессив (парадигмы «дуративного» регистра), оптатив, а также рамочные выделительные конструкции. Аналитические конструкции состоят из обязательного вспомогательного связочного элемента и смысловой формы. Основной проблемой при маркировании таких конструкций стала обязательная демонстрация связи между двумя элементами. Используя существующий опыт, можно предложить три варианта маркирования одного и того же предложения со стативной конструкцией (представлены в примере (4)).

(4)

a. Значение конструкции (St) присваивается вспомогательному элементу

paykun		kun	**no**	**wondi**			e	barehun		
pay-	kun	kun	**no**	**won-**	**d-**	**i**	e	bare-	hun	
ребёнок-	sgKUN	Def.sgKUN	**Cop.*St***	**быть-**	**Soc-**	**Act**	Prep	собака-	sgKUN	

Тот мальчик жил с собачкой.

б. Значение конструкции (St) присваивается основному элементу

paykun		kun	**no**	**wondi**			e	barehun		
pay-	kun	kun	**no**	**won-**	**d-**	**i**	e	bare-	hun	
ребёнок-	sgKUN	Def.sgKUN	**Cop**	**быть-**	**Soc-**	**Act.St**	Prep	собака-	sgKUN	

Тот мальчик жил с собачкой.

в. Значение конструкции (St) присваивается обоим элементам конструкции

paykun		kun	**no**	**wondi**		e	barehun		
pay-	kun	kun	**no**	**won-**	**d-**	**i**	e	bare-	hun
ребёнок-	sgKUN	Def.sgKUN	**Cop.St**	**быть-**	**Soc-**	**Act.St**	Prep	собака-	sgKUN

Тот мальчик жил с собачкой.

На данный момент корпус использует вариант, представленный в примере (4б), поскольку он наиболее легко реализуется технически. Однако в перспективе развитие программы позволит маркировать аналитические конструкции с помощью способа, разработанного на базе решения (4в). Он представлен в примере (5) и представляет собой систему внутренних сносок, связывающих элементы конструкции, более адекватную и экономичную, чем решение (4в).

(5)

barehun		kun	**no**	**humpitii**	
bare-	hun	kun	**no**	**humpit-**	**ii**
собака-	sgKUN	Def.sgKUN	***Cop**	***иметь.сведения-**	**Md.Pfv{*St}**

Собачка знает.

Конструкции с фокусом контраста

Морфология глагола пулар отражает типологически примечательную связь с синтактико-коммуникативной организацией высказывания. Благодаря этой связи "глаголы получают особое морфологическое оформление в случае, если высказы-

вание построено говорящим с контрастивным выделением 1) субъекта, 2) глагольного предиката (сказуемого), 3) прочих аргументов предиката (дополнения или обстоятельства)" [1]. Контрастивная конструкция в диалектах пулар (западная диалектная зона языка пулар-фульфульде) состоит, как правило, из двух элементов – собственно глагольной формы, состоящей из двух обязательных элементов – глагольной основы и видо-залогового аффикса соответствующей приконтрастивной парадигмы, и собственно фокализованного элемента. В случае с контрастивно выделенным субъектом или второстепенным членом предложения, фокализованный элемент может вводиться лексемой ko, помеченной глоссой Foc (см. пример (6)).

(6)

ɓe	haldi			**ko**	**kamɓe**	jombinndirta		
ɓe	hal-	d-	i	**ko**	**kamɓe**	jomb-	inndir-	ta
3.plꞌBE	говорить-	Soc-	Act.Pfv.w	**Foc**	**Emph.3.plꞌBE**	жениться-	Recp-	Act.Pot.**Sb**

Они договорились, что [это именно] они поженятся.

Однако в случае с аутофокализованным глаголом вокализуемая составляющая содержится в самом глаголе. Это затрудняет определение фокализованной конструкции и, таким образом, её обработку программой-парсером. Также наличие аутофокализованного глагола предопределило ещё одно важное решение, принятое при создании корпуса: фокализованные субъект и объект решено было не выделять специальным образом, оставив лишь маркировку прифокусной глагольной парадигмы и, при его наличии, фокусный элемент *ko*.

Другая сложность такой конструкции с точки зрения автоматической разметки состоит в частичной или полной омонимии базисных и контрастивных парадигм. Если в предложении присутствует контрастивно выделенный член, омонимию можно разрешить достаточно легко, однако при отсутствии фокализующего *ko* приходится ориентироваться на второстепенные критерии, как-то: связь с соседними предикатами, акцентно-просодические средства, общий смысл фразы и др. Очевидно, что машина не в состоянии выполнить тонкий анализ, который зача-

[1] Коваль А.И., Нялибули Б.А. Глагол фула в типологическом освещении. М.: Ин-т Языкознания РАН, Ин-т Русского Языка РАН, 1997. С. 26.

стую требуется для снятия омонимии форм из приконтрастивных глагольных парадигм. В корпусе пулар встречаются такие спорные моменты, и в случае, когда определить тип парадигмы невозможно, было принято решение использовать базисную маркировку.

Глава 3. Морфологический уровень языка и реализация строки глоссирования

Третья глава посвящена проблемам, возникающим на морфологическом уровне при сегментации текстов, а также таким вопросам оформления строки глосс, как способ указания заимствований, вид глагола, необходимая и избыточная лингвистическая информация.

Морфемная граница

При глоссировании текстов на пулар возникает ряд проблем, не в последнюю очередь из-за весьма сложной структуры языка, препятствующей морфочленению. Существует общая языковая тенденция в сторону упрощения и унификации сложных морфонологических узлов, и в результате этих процессов на морфемном шве зачастую происходят процессы, делающие разбиение на морфемы практически невозможным. В случае, если глоссируется регулярно встречающаяся двуморфемная форма, состоящая из корня и аффиксального показателя (именного класса, видозалогового и др.), проблем при разметке возникает существенно меньше, чем в случае более распространённой ситуации, когда простая аддитивная модель состоит из трёх и более морфем, многие из которых отделимы в диахронической перспективе, но на синхронном уровне уже неразделимы. Это ставит перед создателями корпуса важную задачу: позиционируя корпус как синхронный, отделить те основы, которые, тем не менее, стоит указать как составные, от тех, чей состав уже принадлежит к области диахронического анализа. Также морфочленение затрудняют разнообразные и многочисленные процессы на стыке морфем, как-то: ассимиляции, возникновение вокалической прослойки и др. Случаи фузионного сращения морфем также составляют значительную трудность при морфочленении. При-

мер (7) демонстрирует глагольный показатель, слитый с прямообъектным местоимением и геминатой просодического происхождения.

(7)

o	yennimmi	
o	yenn-	immi
3.sgO	бранить-	Act.Pfv.w.DO.1Sg

Он бранил меня.

Значимый морфологический нуль

Нулевая морфема – это весьма распространённое явление в языке пулар. К числу нулевых морфем, которые решено было обязательно отмечать в тексте, относится, главным образом, видо-залоговый аффикс (только активного залога). Этот регулярный элемент глагольной словоформы может иметь нулевое воплощение в некоторых формах индикатива (см. пример (8)).

(8)

kun	nannoo			fow
kun	nan-	Ø-	noo	fow
3.sgKUN	слышать-	Act.Pfv.w-	Retr	всё

Она всё слышала.

Также нулевую морфему возможно отмечать в императиве единственного числа активного залога. Отметим также, что омонимичная форма, в сочетании с соответствующей копулой, используется для образования аналитической формы оптатива.

(9)

aritoy			
ar-	it-	oy-	Ø
приходить-	Iter-	Dist-	Act.Imp.2Sg

Приходи в другой раз.

И, наконец, закрывает список нулевых морфем, которые эксплицитно присутствуют в корпусе пулар, нулевой аффикс имён существительных, входящих в неличный подкласс лично-сингулярного класса O, объединяющего в себе многие заимствования. Как следует из терминологии, у класса есть и основное, личное воплощение, которое объединяет в себе имена родства, названия профессий и т.д. Личные имена класса O нормально оформляются материальным показателем соот-

ветствующей ступени, а имена неличного O-подкласса часто оформляются нулевым аффиксом, тем не менее, требуя согласования по классу O (см. пример (10)).

(10)

batoo		yordo		
batoo-	∅	yor-	∅-	do
лодка-	sgO	быть.сухим-	Act.Pfv[PCP]-	sgO

сухая лодка

Но периодически в текстах на языке пулар, независимо от диалекта, можно встретить не оформленную по классу именную основу. Имена с такими основами имеют односоставную структуру (только корень), употребляются лишь в контекстах, не требующих согласования по классу (т.е. без атрибута), а также закономерно не содержат грамматического значения. Такие основы вполне могут быть оформлены по определённому именному классу с помощью аффикса, и тогда они потеряют генерическое значение, но приобретут численно-грамматические характеристики (см. пример (11)).

(11)

а) gerto 'курица~куры', но

б) gerto-gal 'курица-sgNGAL'

gertoo-de 'курица-plDE'

Стоит отметить, что прочие случаи, где морфологический нуль не будет иметь выражения в разметке, единичны и спорны, и это обусловило отказ от дополнительной сегментации по умолчанию. Технически оформление обязательного морфологического нуля осуществляется путём его вставки в исходный текст и последующего анализа программой.

Классный циркумфикс

Обширная система именных классов является, как уже было сказано, характерной особенностью языка пулар-фульфульде (свыше двадцати именных классов). Разграничение существительных по именным классам производится с помощью комплекса из трёх критериев: синтаксического, морфологического и семантического. При создании аннотированного корпуса языка пулар наибольшее внимание,

очевидно, уделяется морфологическому критерию разделения на именные классы. Согласно ему, существительное, принадлежащее к какому-либо именному классу языка пулар-фульфульде, как и согласованный с этим существительным атрибут, оформляются единовременно двумя способами, формирующими так называемый классный циркумфикс. Во-первых, каждый класс имеет определённый набор классных аффиксов завершающих словоформу, а во-вторых, каждый класс, в релевантных случаях, требует использования одного из трёх типов начального корневого согласного – на ступени смычного, несмычного или преназализованно-смычного. Такое оформление составляет общеязыковую модель, от диалекта к диалекту допускающую большее или меньшее варьирование. Факт наличия этой системы должен быть отражён в разметке. Образец типичного оформления классного циркумфикса в согласованных атрибутах показан в примере из диалекта масина (12).

(12)

taali			da6ɓi		
taal-	i		^{Oc}da6ɓ-	i	
сказка-	plƊI		^{ƊI}короткий-	plƊI	

‘сказки короткие’

nyiikoy			ndanewoy		
nyii-	koy		^{PrenOc}ndane-	woy	
зуб-	plKOY		^{KOY}белый-	plKOY	

‘беленькие зубки’

nyannde			wootere		
nyan-	nde		^{NOc}woote-	re	
^{NDE}день-	sgNDE		^{ƁE}один-	sgNDE	

‘один день’

Технически такой подход пока не реализован, поскольку диалект пулар футаджаллон, на основе текстов которого закладывается корпус, использует циркумфикс лишь частично – в этом диалекте утрачена продуктивная система анлаутного согласования. Благодаря этому в корпусе пулар на данный момент используется минималистский подход к оформлению циркумфикса: согласование в именной

группе глоссируется без указания ступени анлаута вообще. То же относится и к анлауту глагола, способному в большинстве диалектов – но не в диалекте фута-джаллон – служить средством согласования глагола с подлежащим по числу. Разумеется, в реализуемом подходе есть существенные минусы в общеизвестном плане, но он удовлетворительно отражает ситуацию в пулар фута-джаллон. В перспективе же должна быть разработана система, позволяющая с большей или меньшей степенью автоматизма проводить разметку именной и глагольной групп с указанием, в релевантных случаях, ступени анлаута и связи её с именным классом или глагольной парадигмой, а также конвертировать результат в формат xml.

Заимствования

При создании аннотированного корпуса пулар информация о заимствованиях была опущена, поскольку при всей безусловной привлекательности её с описательной точки зрения, она выходит за рамки задачи по синхронному представлению языка. Однако совершенно исключать заимствования из поля зрения не следует, в основном из-за их несоответствия стандартным моделям языка.

Интегрированные заимствования, впрочем, никакой особой сложности для автоматической разметки не представляют, поскольку они полностью подчинены правилам языка, и неисконность корня является лишь «академическим фактом», неспособным как-либо повлиять на процесс разметки. По-другому обстоит дело с варваризмами и не вполне интегрированными заимствованиями.

В качестве примеров не вполне интегрированного и интегрированного заимствований хорошо подходят имена, представленные в примере (13). Словоформа под литерой а) представляет собой недавно пришедшую в язык словоформу. Стоит заметить, что имена, пришедшие в язык пулар из других языков, в большинстве своём, попадают в неличный подкласс класса О, принимающий заимствования. В диалектах пулар-фульфульде принимать заимствования способны и другие классы, но неличный О-подкласс является средоточием заимствований на всём пространстве пулар-фульфульде. Лексика, попадающая в этот подкласс, трактуется как получающая формальный нулевой аффиксальный показатель класса, что обеспечивает способность единиц участвовать в процессе согласования и прономинализации.

Под литерой б) находятся интегрированные заимствования, то есть, те, которые локализуются в каком-то другом классе, но выбор этого класса осуществляется преимущественно по формальным. В первом случае классификация произошла по семантическому признаку, то есть, понятие «финиковая пальма» семантически относится в «класс деревьев» KI, и, соответственно, оформляется по этому классу. Во втором же случае важную роль в классификации сыграл фонетический фактор: из-за наличия в языке общеизвестного класса NGOL, существительное было ложно переразложено и отнесено именно к этому классу, хотя семантически ничего общего с его полем не имеет.

Однако не всем заимствованиям удаётся подпасть под один из этих признаков и покинуть «свалочный класс», как его называют типологи. Некоторые заимствованные словоформы, даже старые, так и остались оформленными по неличному O-подклассу, или же вовсе в форме неоформленной по классу основы.

(13)

а)

karsiini on
karsiini- ∅ on
бензин- sgO Def.sgO

б)

tamaroohi kin
tamaroo- hi kin
финик- sgKI Def.sgKI
‘финиковая пальма’

lekkol ngol
lekk- ol ngol
школа -sgNGOL Def.sgNGOL
‘школа’

При глоссировании заимствований возникает лишь одна проблема: вне контекста невозможно отличить не вполне интегрированные заимствования от интегрированных по фонетическому признаку. Зачастую для этого требуется экспертная оценка носителя. Если же экспертная оценка невозможна, то в неочевидных случа-

ях решено было определять спорные заимствования как неинтегрированные, чтобы избежать более грубой ошибки – определения имени в неправильный класс.

При этом интегрированные заимствования не представляют сложности для разметки и рассматриваются как стандартная лексика.

В Заключении подводятся итоги проведённого исследования. Основные результаты работы состоят в следующем:

- разработана система программного обеспечения, рассчитанная на использование с языком пулар-фульфульде;
- разработана система и конвенции глоссирования языка пулар фута-джаллон с учётом его лингвоспецифики и типологических особенностей;
- предложен способ транскрибирования устной речи с учётом перспективы последующего глоссирования;
- предложен формат представления глоссированных текстов на языке пулар фута-джаллон;
- исследованы синтаксические явления языка, представляющие трудность для глоссирования, и выработана стратегия их обработки;
- рассмотрены морфологические процессы в языке и предложены способы их адекватного отображения в морфологической разметке;
- определены случаи надклассных местоимений и предложены способы их отображения в глоссировании;
- составлен список грамматических категорий языка пулар фута-джаллон и обосновано частичное использование этих категорий в морфологической разметке;
- внесены предложения по технике перевода некоторых видов лексем, в частности, введено понятие квазиперевода.

Публикации

По теме диссертации опубликованы следующие работы.

- **Косогорова М.А. К вопросу о согласовании по категории именного класса в пулар-фульфульде // Труды Института лингвистических исследований РАН. Т. VII. Ч. 3. Отв. ред. Д.В. Герасимов. СПб.: Изд-во «Наука», 2011г. Сс. 322-326.**
- **Косогорова М.А. Christiane Seydou. L'épopée peule de Boûbou Ardo Galo: Héros et rebelle. Paris: Éditions Karthala, 2010. (Рецензия) // Вопросы языкознания №3. 2012. Сс. 147-149.**
- **Коваль А.И., Косогорова М.А. К проблемным вопросам морфоглоссирования текстов пулар-фульфульде // Вопросы филологии №36. М. 2012.**
- Косогорова М.А. Анализ именных согласовательных моделей в авторском литературном тексте на фульфульде // Исследования по языкам Африки. Вып. 3. М.: Ин-т Языкознания РАН, 2009. Сс. 177-191.
- Косогорова М.А. К проблеме начально-корневых чередований в пулар-фульфульде в диалектологическом аспекте // Африканский сборник-2009. СПб.: Кунсткамера. 2009. Сс. 204-213.
- Косогорова М.А. К проблеме корневых чередований в пулар-фульфульде в диалектологическом аспекте // «Востоковедение и африканистика в диалоге цивилизаций»: XXV Международная конференция по источниковедению и историографии стран Азии и Африки. Тезисы докладов. Отв. Ред. Н.Н. Дьяков. СПб. 2009. С. 119.
- Коваль А.И., Косогорова М.А. К принципам морфоглоссирования текстов пулар: проблемные вопросы // «Модернизация и традиции»: XXVI Международная конференция по источниковедению и историографии стран Азии и Африки. Тезисы докладов. Отв. ред. Н.Н. Дьяков и А.С. Матвеев. СПб. 2011. Сс. 282-283.
- Косогорова М.А. К проблеме глоссирования имён существительных в языке фула // «Модернизация и традиции»: XXVI Международная конференция по источниковедению и историографии стран Азии и Африки. Тезисы докладов. Отв. Ред. Н.Н. Дьяков и А.С. Матвеев. С Пб. 2011. Сс. 283-285.
- Косогорова М.А. К вопросу о семантических воплощениях именных классов в языке пулар (диалект фута-джаллон) // XII Конференция африканистов «Африка в условиях смены парадигмы мирового развития». М. 2011. Сс. 185-186.