ВОЗМОЖНОСТЬ ГЛОССИРОВАНИЯ АНАЛИТИЧЕСКИХ КОНСТРУКЦИЙ В ЯЗЫКЕ ПУЛАР: ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ И ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ

Косогорова М. А. (maria.kosogorova@gmail.com)
Институт языкознания РАН, Москва, Россия

Косогоров В. Н. (vadim.kosogorov@gmail.com)
ООО «Аплана.ЦР», Москва, Россия

В статье обсуждается возможность автоматического глоссирования аналитических конструкций в языке пулар. Основными препятствиями для этого являются омонимия аффиксов смыслового глагола и сложности с демонстрацией связи между составляющими конструкции. В теоретической части обсуждается несколько вариантов морфологической разметки таких конструкций, с указанием положительных и отрицательных факторов. В результате предлагается, на наш взгляд, оптимальная тактика обозначения аналитической пары. В практической части предлагается алгоритм, использованный для решения этой задачи в рамках программы LightParser.

Ключевые слова: пулар, автоматическое глоссирование, аналитические конструкции, снятие неоднозначности, морфология

THEORETICAL BASIS AND PRACTICAL IMPLEMENTATION OF INTERLINEARIZING THE ANALYTIC CONSTRUCTIONS IN PULAAR

Kosogorova M. A. (maria.kosogorova@gmail.com)
Institute of Linguistics, RAS, Moscow, Russia

Kosogorov V. N. (vadim.kosogorov@gmail.com)
Aplana Development Center, Moscow, Russia

The article presents a possibility of an automatic glossing of the analytic constructions in Pulaar. These constructions present two main problems. Firstly, the main verb affixes in an analytic construction are widely homonymous to the independent ones. Secondly, given this homonymy, it is very important to demonstrate the connection between the auxiliary and the main verb, because otherwise the two parts of the construction might just be accepted separately, which is a mistake. The theoretical part of the article presents several approaches to the solution of the problem with different ways to underline the connection. The advantages and the drawbacks of each approach are also discussed. The optimal tactics possible, given the Pulaar specialties is proposed. The practical part of the article offers a description of an algorithm implemented in the parsing software (LightParser) developed especially for Pulaar. This algorithm allows the automatic (in most cases) glossing of the analytic constructions.

Keywords: Pulaar, automatic glossing, analytic constructions, disambiguation, morphology

Фула (другие самоназвания включают в себя пулар, фулфульде, пёль, фуль, фулани и т. д.) — это один из самых известных языков Африки. Он относится к западно-атлантической языковой семье, но территориально выходит далеко за пределы её ареала.

Благодаря своему широкому распространению в Африке, высокому статусу и другим экстралингвистическим факторам, исследования языка фула начались около 150 лет назад, и к настоящему времени существует ряд весьма информативных словарей и грамматик разных диалектов, а также отдельные грамматические, синтаксические и лексикографические исследования. Однако для того, чтобы идти в ногу с современной лингвистической наукой и предлагать свои данные максимальному количеству исследователей, языку необходим корпус глоссированных текстов. Возможность поиска по корпусу с открытой разметкой позволит проводить исследования в различных областях языка. Также для работы с таким корпусом нет необходимости полностью владеть основами пулар. Для диалекта пулар фута-джаллон (Гвинея) была поставлена задача по созданию такого корпуса, и это привело к созданию независимого программного продукта LightParser для глоссирования именно этого диалекта с учётом всей его специфики.

Созданию независимого программного обеспечения предшествовали попытки использовать для этих целей уже существовавшие на тот момент программные пакеты, в том числе общеизвестные ToolBox и FieldWorks. Однако оба они оказались слабо применимы к лингвоспецифике языка в своём базовом виде, а отсутствие доступного открытого программного кода не позволило видоизменить их в соответствии с нуждами конкретного языка. Так, например, система классного циркумфикса (оформление имени и атрибута по категории именного класса одновременно с помощью аффикса и путём изменения начального корневого согласного) является практически уникальной для пулар, и поэтому не представлена в стандартных программных продуктах. Впоследствии была выпущена следующая версия пакета FieldWorks Language Explorer,

которая предоставляла больше возможностей и могла бы использоваться для работы с языком пулар; например, там была возможность, хоть и с ограничениями, маркировать аблаутные чередования. Однако в тот момент уже существовала пилотная версия программы LightParser, и, к тому же, настройка существующей системы под особенности языка с «нетиповой» лингвоспецификой и сложной морфонологией кажется нам немного более трудоёмкой задачей, нежели создание собственного продукта и его развитие по мере необходимости.

Программа LightParser, по сути, задумывалась как аналог ToolBox с небольшими изменениями, однако впоследствии она пробрела автоматическую направленность. В настоящее время она разбивает текст на строки по клаузам и на слова по ячейкам и загружает его в отдельное окно, где пользователь может редактировать текст, запускать пошаговую или автоматическую обработку. В случае пошаговой обработки пользователь инициирует глоссирование каждого слова отдельно, а автоматическая обработка запускается до конца текста или указанного отрезка текста. При пошаговой или поклаузной обработке омонимия снимается в режиме синхронного запроса, а при автоматической обработке — экспертом после окончания работы программы.

LightParser также предоставляет функциональность работы со словарями. При запуске программа автоматически загружает доступные словари. Формат словарей совместим с программой Toolbox, хотя и имеет небольшие отличия, предоставляющие расширенные возможности там, где это необходимо и исключающие не нужные для языка пулар опции. Важным, на наш взгляд, преимуществом словарей LightParser является отсутствие «базового варианта» морфемы и его обязательная демонстрация при нахождении в тексте альтернативы. Программа предоставляет удобный пользовательский интерфейс для редактирования словарей. Каждый словарь может быть открыт в отдельном окне, где имеются функции добавления, удаления, редактирования элементов словаря, а также функция поиска, как общего, так и специального. После сохранения изменений программа автоматически подгружает изменённый словарь для продолжения обработки текста.

Программа позволяет сохранять промежуточные и готовые тексты в файлах HTML или XML. Также тексты могут быть экспортированы в MSWord. Готовые тексты можно сохранять в файлы XML для обработки сторонним ПО (например, программой Search Too).

Программа LightParser разработана на платформе Microsoft .NET и языке C#. Её основой является библиотека классов .NET Framework, покрывающая весь спектр задач, с которыми сталкиваются разработчики современных программ. Язык C# позволяет в полной мере воспользоваться объектно-ориентированным программированием при помощи классов .NET Framework.

Для реализации оконного приложения Windows используется технология Windows Presentation Foundation (WPF), впервые разработанная Microsoft для .NET Framework 3.0 и существенно улучшенная в версии .NET Framework 4.0. Данная технология предусматривает описание оконного интерфейса при помощи языка XAML и последующую векторную визуализацию, которая не зависит от вида или разрешения устройства вывода.

Для создания удобного пользовательского интерфейса используется набор библиотек компонентов DevExpress. Данные библиотеки расширяют стандартную функциональность WPF и позволяют реализовать такие возможности как многооконный интерфейс с быстрым переключением между окнами, текстовый редактор, напоминающий MS Word, позволяющий редактировать тексты в табличном виде. Для работы с файлами HTML используется библиотека HtmlAgilityPack.

В настоящее время программой LightParser производится удовлетворительное автоматическое глоссирование большинства словоформ пулар. Важным достижением является также возможность глоссирования аналитических конструкций, которая невозможна, без серьёзных изменений, в известных нам системах обработки текстов.

Аналитическая форма определяется как составная форма, состоящая из служебного и знаменательного слов. В пулар аналитические конструкции используются для выражения ряда грамматических категорий, в том числе статива и прогрессива (глоссы St и Prog соответственно), оптатива (глосса Opt) и др. Работу аналитической конструкции и способы её глоссирования удобнее всего демонстрировать на стативных и прогрессивных формах, так как они наиболее стабильны, удобны и частотны.

Статив и прогрессив — это две парадигмы особого регистра, отличающиеся, в отношении принципа структурирования, «заметным единством на всём пандиалектном пространстве пулар-фульфульде» [Коваль 2003: 377 и далее про аналитические конструкции статива и прогрессива]. Общая формула этих аналитических конструкций может быть выражена как <предикативная связка + глагол>. В случае прономинально выраженного субъекта предикативная связка поглощается местоимением, образуя особый разряд «сложных», или, иначе, копулосодержащих местоимений. Такие местоимения при разметке получают пометку Cop (см. пример (1)). Звёздочки в строке глосс составляют внутреннюю сноску, подробнее о способе разметки конструкции см. ниже:

(1)

si	*tawii*	***midʼo***	***waawi***		*mi*	*okkete*	
si	*tawii*	***midʼo***	***waaw-***	***i***	*mi*	*okk-*	*ete*
если	*случилось*	***1Sg.Cop****	***мочь-***	***Act.Pfv*{*St}***	*1Sg*	*дать-*	*Act.Pot.s.DO.2Sg*

Если [вдруг] случится, что я могу, я дам тебе.

Связка — это регулярный элемент аналитической конструкции, в обязательном порядке упоминаемый в грамматиках всех диалектов (подробнее о связках аналитических конструкций и их диалектных особенностях см. [Miyamoto 1993]). В корпусе пулар помечается традиционная для этого диалекта связка no, указанная в том числе в грамматике [Diallo 2000]. Глагол же в дуративных (стативных и прогрессивных) конструкциях имеет морфологию, аналогичную «свободным» формам глаголов соответствующего вида — перфективного для статива и потенциалисного для прогрессива (см. пример (2)).

(2)

debbo		*on*	***no***	***Faalaa***			*warde*			*paykun*		*kun*
debb-	*o*	*on*	***no***	***faal-***	***Aa***		*war-*	*∅-*	*de*	*pay-*	*kun*	*kun*
женщина-	*sgO*	*Def.sgO*	***Cop****	***хотеть-***	***Pass.Pfv*{*St}***		*убить-*	*Act-*	*Inf*	*ребёнок-*	*sgKUN*	*Def.sgKUN*

Женщина хочет убить ребёнка.

Ср. интерпретацию faalaa вне стативной конструкции:

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

Она захотела сделать то, что она хотела.

Проблема автоматической разметки аналитических форм в полной мере стоит перед исследователями, например, английского языка. Следовательно, существует целый ряд решений, принятых разными исследователями для разных конкретных целей. Так, например, Йорк-Торонто-Хельсинкский корпус староанглийского языка присваивает общее значение конструкции вспомогательному элементу [http://www-users.york.ac.uk]; также возможно присвоение общего значения лишь основному элементу, или же обоим элементам конструкции. Три подхода к обозначению аналитических конструкций представлены в примерах (3а)–(3в).

Во всех трёх подходах есть и плюсы, и минусы. Бесспорно, то решение, при котором связь двух элементов не указывается вообще, неприемлемо для пулар, поскольку при этом аналитическая глагольная форма будет обозначаться через омонимичный независимый показатель соответствующего вида, а это в корне ошибочно. Поэтому такое решение подробно рассматриваться не будет.

(3)

а. *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***	*bare-*	*hun*
ребёнок-	*sgKUN*	*Def.sgKUN*	***Cop.St***	***быть-***	***Soc-***	***Act.St***	*собака-*	*sgKUN*

Тот мальчик жил с собачкой.

Решения типа (а) не являются оптимальными для пулар, поскольку основным элементом конструкции является глагольная форма, а она не зафиксирована как статив. Тактика обозначения лишь основного элемента (решение (б)) является, как нам кажется, допустимой в рамках небольшого корпуса пулар, для которого будет приемлемо отдельное (в виде сноски или в инструкции) разъяснение, указывающее на аналитическую природу конструкции. Именно это решение было использовано в предыдущей версии программы-парсера. Оно основано на идентичности вспомогательных элементов в различных парадигмах дуративного регистра, что представляет трудность для автоматической разметки. При использовании такой тактики программа-парсер помечает связку глоссой Cop, а с аффикса смыслового глагола программа снимает омонимию и помечает его глоссой St или Prog соответствующего залога (см. пример (4)). Преимущество такого подхода заключается в экономичности как с точки зрения пространства (что несущественно в условиях электронного корпуса, но в случае публикации на бумажном носителе значительно), так и с точки зрения программного ресурса.

(4)

debbo		*on*	***no***	***faalaa***		*warde*			*paykun*		*kun*	
debb-	*o*	*on*	***no***	***faal-***	***aa***	*war-*	*∅-*	*de*	*pay-*	*kun*	*kun*	
женщина-	*sgO*	*Def.sgO*	***Cop хотеть-***	***Pass.St***		*убить-*	*Act-*	*Inf*	*ребёнок-*	*sgKUN*	*Def.sgKUN*	

Женщина хочет убить ребёнка.

Однако у приведённого решения есть существенный недостаток: для пользователя, который не полностью владеет тонкостями морфологии пулар, такого обозначения аналитической формы будет недостаточно — он может не заметить и не учесть связи между элементами и допустит ошибку, обращаясь с дуративом как с синтетической формой. Такая ошибка исключена при использовании третьего варианта решения — двойном маркировании аналитической формы.

Тактика двойного маркирования широко используется при маркировании аналитических форм не случайно: это наиболее информативный и удобный, с точки зрения пользователя, метод их обозначения. В примере (3в) демонстрируется один из вариантов такой тактики. Впоследствии эта тактика была признана нами слишком громоздкой: дублирование показателя не удваивает информативности разметки. Но на основе этого варианта был разработан формат, который используется в текущей версии программы; он представлен в примере (5).

(5)

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

Собачка знает.

Как следует из примера, маркирование аналитических форм дуратива производится по принципу микро-сноски, где индексом служит астериск (как наиболее частотный символ для таких целей), а при финальном элементе конструкции

вместо указания регистра помещена расшифровка. Ещё одним преимуществом выбранного подхода является возможность указывать в глагольном формообразующем аффиксе не только залог, но и вид, что при прочих подходах привело бы к существенному разрастанию строки глосс. Аналогичным способом маркируется и оптатив (см. пример (6)), но с оговоркой, что вместо связки оптативную конструкцию дополняет частица. Ранее (см. пример (7)) частица маркировалась глоссой-переводом «пусть», а глагол был снабжён, помимо залога, глоссой Opt, формально омонимичной императиву единственного числа (глосса Imp).

(6)

njano		*on*	*inni*			***yo***	*ɓe*	***yaltu***		
njan-	*o*	*on*	*inn-*		*i*	***yo***	*ɓe*	***yalt-***		***u***
большой-	*sgO*	*Def.sgO*	*говорить-*		*Act.Pfv.w*	***Part****	*3.plƁE*	***выходить-***		***Act*{*Opt}***

Старший сказал, чтобы они вышли.

(7)

njano		*on*	*inni*			***yo***	*ɓe*	***yaltu***	
njan-	*o*	*on*	*inn-*		*i*	***yo***	*ɓe*	***yalt-***	***u***
большой-	*sgO*	*Def.sgO*	*говорить-*		*Act.Pfv.w*	***пусть***	*3.plƁE*	***выходить-***	***Act.Opt***

Старший сказал, чтобы они вышли.

Такая же тактика, при необходимости, используется и в других аналитических формах, например в аналитическом отрицании, рамочных выделительных конструкциях (см. пример (8)):

(8)

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.Scnd*		***Def*{*Rel}***

Она захотела сделать то, что она хотела.

Однако такая тактика плохо применима к аналитическим конструкциям с фокусом контраста из-за их не вполне аналитической природы. В максимальном варианте фокализованная конструкция будет включать в себя фокализованный элемент с частицей-фокализатором ko, а также глагол a специальной приконтрастивной форме (см. пример 9а). В минимальном же варианте она ограничится лишь одним аутофокализованным глаголом (см. пример 9б).

(9)

а.

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

Они договорились, что [это именно] они поженятся.

б.

mi	***wi'u***		*kono*	*mi*	*waɗaali*	
mi	***wi'-***	***u***	*kono*	*mi*	*waɗ-*	*aali*
1Sg	***сказать-***	***Act.Pfv.Praed***	*но*	*1Sg*	*делать-*	*Act.Pfv.Neg*

Я [только] сказал, но не сделал.

Для реализации глоссирования аналитических конструкций описанным выше способом в программу добавлена дополнительная функциональность, позволяющая автоматически находить такие конструкции в тексте и обрабатывать надлежащим образом.

С точки зрения программы, глоссирование аналитической конструкции должно состоять из распознавания глагола-связки, поиска соответствующего ему смыслового глагола, и снятия омонимии с приписыванием паре специальных обозначений, обозначающих аналитическую конструкцию.

Для выполнения первого этапа, то есть распознавания глагола-связки, в программу добавлен конфигурируемый список обозначений связок. При выполнении второго этапа, то есть, поиска смыслового глагола, используется реализованный ранее функционал определения частей речи. Для корректного глоссирования смыслового глагола в интерфейс редактирования словаря программы добавлена возможность задания специальных значений формообразующего глагольного аффикса, помеченных как составляющие аналитической конструкции. При нахождении нужной морфемы в составе смыслового глагола соответствующей конструкции, вместо обычного значения программа использует эту специальную глоссу, которая также содержит в себе общее значение аналитической конструкции.

Алгоритм обработки аналитических конструкций выглядит следующим образом. При глоссировании текста программа ищет в строке значений обрабатываемой словоформы специальную пометку, (например, ***Cop, Part***). При нахождении такой словоформы производится поиск смыслового глагола среди нескольких последующих слов (по умолчанию проверяются три последующих слова). В случае, если в заданном промежутке находится глагол, при его глоссировании программа использует специальное значение для его формообразующего аффикса, которое отмечено в словаре как относящееся к аналитической конструкции (например, ***-Act.Pfv****). Для пометки глагола-связки в его глоссу добавляется астериск.

Для идентификации глагола используется значение части речи, которое задаётся в словаре для каждой морфемы. В случае наличия неоднозначности сначала применяется механизм снятия омонимии через соответствие частеречных помет каждой морфемы, а при невозможности однозначного автоматического разбора — запрос пользователю на проведение этой операции вручную. Затем применяется механизм распознавания аналитических конструкций.

В результате работы программы аналитические конструкции автоматически распознаются и глоссируются соответствующим образом: глагол-связка помечается астериском (например, ***Cop****), а смысловой глагол глоссируется как часть аналитической конструкции (например, ***Act.Pfv****). В конце последней составляющей (т. е. глагола) конструкции присваивается общее значение в формате внутренней сноски (например, ***{*St}***).

Предложенное решение позволило существенно упростить процедуру разметки аналитических конструкций, доведя её до полного автоматизма в стандартных ситуациях. Мы надеемся, что по мере введения в работу программы статистического анализатора, от вмешательства оператора при работе парсера в этой области возможно будет полностью отказаться. Также данное решение может быть применимо при глоссировании аналогичных конструкций на других языках.

Литература

1. *Diallo A.* Grammaire descriptive du pular du Fuuta Jaloo (Guinée). Frankfurt am Main: Peter Lang, 2000.
2. *Miyamoto R.* A Study of Fula Dialects: Examining the Continuous/Stative Constructions. // Senri Ethnological Studies № 35. Unity and Diversity of a People: the Search for Fulbe Identity. Eguchi, P. K. & Azarya, V. (eds). Osaka: National Museum of Ethnology, 1993.
3. *Коваль А. И.* Контрастивность как морфологическая категория пулар-фульфульде. // Основы африканского языкознания. Глагол. Под ред. В. А. Виноградова и И. Н. Топоровой. М.: Издательская фирма «Восточная литература» РАН, 2003. С. 357–459.

References

1. *Diallo A.* (2000). Grammaire descriptive du pular du Fuuta Jaloo (Guinée), Peter Lang, Frankfurt am Main
2. *Koval A. I.* (2003) Contrastivity as a Morphological Category in Pular-Fulfulde [Kontrasticnost' kak morfologicheskaja kategorija pular-fulfulde], in Osnovy afrikanskogo jazykoznanija. Glagol [Foundations of African Linguistics. The Verb], Vostochnaja literature, pp. 357–459.
3. *Miyamoto R.* (1993) A Study of Fula Dialects: Examining the Continuous/Stative Constructions, Senri Ethnological Studies № 35, pp. 215–230.