Электронный глоссированный корпус текстов языка бамана: первый этап

Valentin Vydrin

В. Ф. Выдрин

ЭЛЕКТРОННЫЙ ГЛОССИРОВАННЫЙ КОРПУС ТЕКСТОВ ЯЗЫКА БАМАНА: ПЕРВЫЙ ЭТАП[1]

0. Введение

В предыдущих публикациях, посвящённых электронному корпусу бамана [Выдрин 2008а; Выдрин 2008б; Vydrine 2008], были высказаны предварительные соображения о необходимости и возможности создания такого корпуса, а также намечались пути решения некоторых конкретных трудностей, которые неизбежно должны были возникнуть в этой работе. Эти идеи стали предметом обсуждения на Второй Международной конференции по языкам манде (СПб, сентябрь 2008) и были поддержаны коллегами из разных стран; обсуждение было продолжено на VI Всемирном конгрессе по африканской лингвистике (Кёльн, август 2009). Осенью 2009 года в Петербурге была создана рабочая группа по разработке модели электронного корпуса текстов бамана, в которую, помимо автора, вошли лингвист-программист Кирилл Александрович Маслинский и специалисты по языкам манде Анна Владимировна Эрман и Артём Витальевич Давыдов. К систематической работе группа приступила в марте 2010 года (после окончания очередной зимней экспедиции российских лингвистов в Гвинею и Кот-д'Ивуар). В качестве метаязыка корпуса был выбран французский, который является официальными языком и основным языком образования в Мали.

В данной статье мы постараемся представить обзор конкретных проблем, решением которых занималась рабочая группа в течение полугода (с марта по сентябрь 2010 г.), и обоснования принятых решений.[2]

[1] Данное исследование выполнено в рамках проекта «Разработка модели электронного корпуса текстов языков манден (маninка, бамана)», поддержанного грантом РФФИ № 10-06-00219-а.

[2] Здесь не будут рассматриваться проблемы, связанные с подбором текстов и их метаразметкой; им посвящена статья А. В. Давыдова в данном сборнике.

Общие соображения о значимости электронного корпуса текстов на языках манде были высказаны в уже упомянутых публикациях, что позволяет не излагать их здесь и сразу перейти к более техническим вопросам.

Напомним принцип действия всех программ автоматического анализа текста, предназначенных для создания языковых корпусов. Несколько упрощая ситуацию, можно сказать, что программное обеспечение состоит из «словарного» файла, а также из программы автоматического анализа (морфологического, синтаксического и др.), т. е. рабочего файла (или, скорее, совокупности файлов), содержащего в себе правила построения словоформ, их сочетаемости между собой и т. д.; «движок» связывает текстовый и словарный файлы. В словарный файл вносятся морфемы (или лексемы, или словоформы[3]) описываемого языка; при каждой морфеме (или словоформе), в другом поле, даётся её эквивалент на метаязыке (который может совпадать с описываемым языком, – в таком случае мы получаем одноязычное глоссирование, – а может быть иным; в нашем случае метаязыком является французский). Для служебных слов и морфем, как правило, даётся условный эквивалент, в соответствии с принципами Лейпцигских правил глоссирования. В особом поле даётся частеречная помета.

Несколько упрощая картину, принцип автоматического анализа можно описать так. Когда даётся команда «анализировать текст», программа-«движок» находит в словарном файле каждую лексему и морфему, представленную в тексте (при этом она членит слова на морфемы), создаёт в текстовом файле строку «парсинга» (поморфемной разбивки) и подставляет к каждой морфеме исходной фразы её эквивалент на метаязыке. Если же морфема в словарном файле не обнаруживается, то программа сигнализирует об этом, предлагая лингвисту различные варианты выбора: создать новую карточку в словаре; отметить слово как иноязычное вкрапление; устранить орфографическую ошибку в тексте.

[3] Последнее может быть необходимо для флективного языка. Поскольку языки манден изолирующие, с элементами агглютинации, то в дальнейшем возможность введения в словарь словоформ рассматриваться не будет.

Организационные структуры всех трёх компонентов корпуса – программы-анализатора, словаря и проанализированного текста – тесно взаимосвязаны. Тем не менее, в целях удобства изложения, эти компоненты и связанные с ними проблемы будут рассмотрены раздельно.

1. Программа-анализатор (парсер)[4]

На предварительном этапе я склонялся к тому, что компьютерная программа Toolbox – наиболее подходящее средство для создания электронного корпуса текстов на языках манден [Выдрин 2008б]. Однако в ходе дальнейших обсуждений с коллегами и в рамках рабочей группы стало очевидно, что некоторые недостатки этого программного продукта создают труднопреодолимые препятствия в работе над большим корпусом, насчитывающим миллионы словоупотреблений. Назовём лишь три таких недостатка:

1) закрытый характер этой программы, т. е. недоступность её исходных текстов. Это делает невозможным для разработчиков корпуса вносить изменения в программу парсинга для устранения конструктивных дефектов (с которыми доводилось сталкиваться, по-видимому, всем пользователям Тулбокса), а также с целью её приспособления к особенностям конкретного языка;

2) невозможность парсинга без ручного снятия омонимии. Эта особенность Тулбокса автоматически сводит его функцию к созданию микро-корпусов и делает невозможной обработку больших массивов текстов;

3) отсутствие в Тулбоксе средств для обработки внеязыковых вкраплений в текст на анализируемом языке – таких как слова или фразы из других языков (французские вкрапления нередки в текстах на бамана; в мусульманской религиозной литературе могут встречаться неадаптированные арабские слова и т. п.) и окказионализмы.

[4] Термин «парсер» чаще всего используется в значении «синтаксический анализатор», однако иногда его употребляют и в смысле «морфологический анализатор» (в частности, разработчики программы Toolbox). В данной статье этот термин будет применяться в основном в последнем значении.

В результате было принято решение о создании специальной программы-анализатора текста на бамана; разработкой этой программы занимается К. А. Маслинский, в режиме постоянных консультаций с остальными членами рабочей группы. В качестве языка программирования был избран Python, при этом рассматривается возможность перевода программы в дальнейшем на другой, более экономный язык.

На данном этапе речь идёт о разработке морфологического анализатора; разработка синтаксического анализатора текста на бамана – значительно более сложная задача, к решению которой предполагается перейти на более позднем этапе работы.

Для проверки работы парсера используется «пилотный корпус» разножанровых текстов на бамана, записанных в старой орфографии, объёмом в 102 тыс. слов (ок. 455 тыс. знаков). Этот файл был любезно предоставлен в наше распоряжение Жераром Дюместром.

В ходе разработки морфологического парсера были созданы следующие продукты:

1.1. Правила преобразования старой орфографии бамана в новую

Старая орфография основывалась на принципах, выработанных на совещании экспертов западноафриканских стран в Бамако в 1963 году и была официально принята в Мали в 1967 году. В 1986-1990 гг. она была замещена новой системой, основанной на африканской версии МФА. Эти две системы различаются в обозначении четырёх фонем (или шести, если учитывать вокалическую долготу): *è, ò, èe, òo, ny, ng* в старом написании соответствуют *ɛ, ɔ, ɛɛ, ɔɔ, ɲ, ŋ* в новом. Трудность представляют два диграфа, которые в старой орфографии не различали релевантные фонологические сущности: *ny* в серединной позиции в слове мог обозначать как носовой сонант /ɲ/, так и сочетание носового гласного с последующим палатальным сонантом, /Ṽy/; *ng* в начале знаменательной морфемы обозначал как носовой сонант /ŋ/, так и преназализованный велярный смычный /ng/. Эта неоднозначность не позволяет конвертировать тексты, имеющиеся в старой орфографии, путём простых автозамен. В то же время программа, предусматривающая обращение к словарю, решает эту проблему почти без остатка: в словаре Ш. Байоля

обнаруживается только одна минимальная пара, демонстрирующая оппозицию /ɲ/ и /Ṽy/ – *kέɲɛ* ‘препятствовать; терпеть неудачу’ : *kέɲyɛ* ‘выравнивать’ (поскольку тоны ни в старой, ни в новой орфографии бамана на письме не обозначаются, следует учитывать и тоновые квазиомонимы: *kɛ̀ɲɛ* 1. песок, *kɛ̀ɲɛ* 2. лобок, *kɛ̀ɲɛ* 3. воск).[5] Минимальных пар на оппозицию /ŋ/ : /ng/ в словаре нет.

1.2. Правила обозначения тонов на письме

В ныне действующей практической орфографии бамана тоны не обозначаются,[6] а в научных публикациях в этом отношении царит анархия: практически каждый автор придерживается своих собственных правил.

В создаваемом корпусе предполагается сплошное тонирование текстов (за исключением, разумеется, иноязычных вкраплений), поэтому весьма актуальным становится формулирование правил, по возможности экономных, но в то же время не допускающих утраты релевантной для языковой системы информации.

Некоторые идеи относительно принципов тональной нотации в корпусе текстов на бамана были высказаны в статье [Выдрин 2008а]. Не повторяя здесь всей аргументации, ограничимся изложением самих правил.

Предлагается использовать следующие тональные диакритики: акут – высокий тон, гравис – низкий тон, гачек – восходящий тон (последний используется редко, только в словах трёх маломестных миноритарных классов,[7] для восходящего тона перед высоким – но

[5] Конечно, нельзя исключать возможность возникновения омонимии в результате продуктивного словосложения.

[6] Точнее, обозначение тонов допускается, но не считается обязательным; при этом правила тональной нотации не формулируются. В публикациях на бамана, предназначенных для малийцев, а также в школьном и университетском преподавании этого языка в Мали тоны не обозначаются практически никогда. Единственным исключением является, по-видимому, различение местоимений 3SG *à* и 2PL *á*: последнее пишется со знаком апострофа в постпозиции, и это правило соблюдается авторами текстов довольно последовательно.

[7] В принципе, можно было бы обойтись и без гачека, но тогда пришлось бы вводить значок плавающего низкого тона внутри слова

не перед низким; в последнем случае обходимся грависом, исходя из правила: «низкий перед низким реализуется как восходящий»).

В словах «стандартных» тональных классов обозначается только тон первого слога (высокий или низкий), вне зависимости от длины слова.

В префиксных глаголах тон (высокий или низкий) обозначается и на префиксе, и на первом слоге глагольной основы: *lákólo* ‘воспитывать’, *lákìrin* ‘вызывать обморок’, *màmìnɛ* ‘бронировать; обручаться’, *màgàn* ‘стараться’, *màjìra* ‘показывать’.

Этот же принцип применяется и в причастиях (образуемых суффиксами *-len/-nen, -tɔ, -ta, -bali*), которые сохраняют тоны исходных глаголов (*lákìrinnen, màmìnɛtɔ, lákìrinbali, màjìrata*), – но не в отглагольных именах (образованных по конверсии или при помощи суффикса *-li/-ni*), тоны которых становятся компактными (*lákirinni* ‘вызывание обморока’*, màminɛli, màminɛ* ‘помолвка’*, màgan* ‘усилие; прилежание’).

Для глаголов-компаундов (типов N+V, [N+Pref.]+V) тон обозначается и на первом слоге именной части, и на первом слоге глагольной основы: *kɔ́nɔnafìli* ‘тревожить’, *kɔ́nɔmíiri* ‘размышлять’, *kùnkɔrɔtà* ‘способствовать успеху’, *kùnnadá* ‘попрекать’.

В бамана, помимо глаголов-компаундов, менее 10% всех слов имеют нерегулярные тональные схемы; это почти исключительно существительные, наречия и служебные слова. Многие слова с нерегулярными тональными схемами малочастотны в текстах (в основном это названия биологических видов), и нередко их тональный контур варьируется от диалекта к диалекту (или даже идиолекту). Ниже даётся список тональных классов по [Dumestre 1987]; примеры также даются из диссертации Дюместра, при этом формы нередко отличаются от тонов соответствующих слов в словаре Ш. Байоля (H – высокий тон, L – низкий тон, R – восходящий тон).

или циркумфлекс, т.е вместо *nkárăngé* – *nká`ránge* или *nkârânge* ‘ловушка’.

Таблица 1. Тоновая нотация для слов различных тональных классов

Структура слова и тон. схема	Пример (в «полной» записи)	Предлагаемая орфография	Комментарий
CV: L	*kà* показатель инфинитива, *à* 3ед.	*kà, à*	только эти два слова
Двусложные:			
CV-CV: HL	*báwò* ‘потому что’, *kúnùn* ‘вчера’	*báwò*, *kúnùn*	не бывает на существительных
CV-CV: R-R	*gělǔ, gèélù* ‘маленький африканский филин’, *tèéndǎ, těndǎ* ‘молотоглав’	*gèlù, tèndà*	долгота гласного автоматическая
CVV-CV: R-L	*làálà* ‘возможно’	*làalà*	редкий
Трёхсложные:			
CV-CV-CV: H-L-H	*bámànán* ‘бамана’, *tásàlén* ‘чайник для омовений’	*bámànan*, *tásàlen*	самый многочисленный тип
CV-CVV-CV или CVV-CVV-CV: H-H-R	*náanáalěn* ‘ласточка’, *bákɔ́ɔnǐn* ‘большая ржанка’	*náanaalèn, bákɔɔnìn*	
CV-CV-CV: H-R-H	*nkárǎngé* ‘ловушка’, *dɔ́rɔ̌mɛ́* ‘5 франков’	*nkárǎngé*, *dɔ́rɔ̌mɛ́*	
CV-CVV-CV: L-H-R	*bòjáarǎ* ‘колючий молочай’, *bàɲáandǐ* ‘сенегальский ткачик’	*bòjárà*, *bàɲándì*	долгота второго гласного автоматическая
CVV-CV-CV: R-B-H	*jɔ̀ɔ́nkɔ̀mí* ‘чёрный скорпион’, *màángòró* ‘манго’	*jɔ̀nkɔ̀mi, màngòro*	долгота первого гласного автоматическая
Четырёхсложные (все - CV-CV-CV-CV)**:**			
H-H-L-H	*késékèlé* ‘зоб	*késekèle*,	частотный тип

Структура слова и тон. схема	Пример (в «полной» записи)	Предлагаемая орфография	Комментарий
	(птицы)', *fógónfògón* 'лёгкое'	*fógonfògon*	
H-H-R-H	*bábúgǔnín* 'муравьиный лев'	*bábugǔnín*	редкий тип; некоторые слова реализуют тон факультативно как H-R-H-H
H-H-H-R	*búnúnkóorǒ* 'шпорцевый гусь', *kɔ́lɔ́bóorǒ* 'коричневый цвет'	*búnunkoorò, kɔ́lɔborò*	редкий тип; долгота предпоследнего гласного автоматическая
H-L-L-H	*wánjàlàká* 'жираф'	*wánjàlaka*	редкий тип
H-L-H-H	*cóbàjírí* 'храбрец'	*cóbàjiri*	редкий тип
L-H-L-H	*dùnúnkàlé* 'оса-строительница', *kòlókòló* 'ощипанная курица'	*dùnunkàle, kòlokòlo*	частотный тип
L-H-R-H	*kànkálĭbá* 'напиток кенкелиба', *fàbúrĕmá* 'мелкий сорт батата'	*kànkálĭbá, fàbúrĕmá*	редкий тип
L-H-H-R	*gìngérénĭn* 'дневная хищная птица', *kàkílákǎ* 'фараонова курица'	*gìngérenìn, kàkílakà*	редкий тип
L-L-H-R	*tòrìmáanǎ* 'пиявка', *ngòròbáanĭn* 'капская горлица'	*tòrimanà, ngòrobanìn*	редкий тип; долгота предпоследнего гласного автоматическая

К сожалению, нельзя быть уверенным, что этот список исчерпывает все возможности, допустимые в различных

диалектах бамана. С другой стороны, поскольку подавляющее большинство подлежащих анализу исходных текстов не будет иметь тоновой нотации, для них эта проблема будет нерелевантной.

Если от существительного с нерегулярной тональной схемой образуется дериват или если такое существительное входит в составное слово, оно переходит в регулярный тональный класс (тон первого слога определяет тональный контур всего слова). Например: *nkárăngé* ‘ловушка’ ⇒ *nkárangenin* ‘ловушечка’, *bámànán* ‘бамана’ ⇒ *bámanankan* ‘язык бамана’.

В бамана выделяется особый класс существительных-композитов, которые, вслед за Жераром Дюместром [Dumestre 1987/1994, 261-285], принято называть «конгломератами». Эти существительные (в отличие от «обычных») образуются не по моделям именных групп и могут сохранять тоны своих компонентов. По наблюдению Ж. Дюместра, более краткие конгломераты (до 3 слогов) имеют сильную тенденцию к приобретению компактных схем, а более длинные – к сохранению исходных тонов компонентов. Нередко в произнесении конгломератов наблюдаются колебания между компактной и некомпактной тоновыми схемами, например: *sìgíńfĕ* (предлагаемая орфография: *sìgí-ń-fɛ̀*) ~ *sìgìnfɛ́* (*sìginfɛ*) ‘мигрант’, *táakàsègín* (предлагаемая орфография: *táa-kà-sègin*) ~ *táakáségín (táakasegin)* ‘хождение туда и обратно’, *jèníkàɲìmí* (предлагаемая орфография: *jèni-kà-ɲìmi*) ~ *jènìkàɲímí* (*jènikaɲimi*) ‘благоприятный случай’.

Конгломераты образуют открытый список; они образуются по различным моделям, причём каждая модель имеет самые разные варианты наполнения, в связи с чем исчисление тональных схем, допустимых для конгломератов, представляется невозможным.

Серьёзный вопрос представляет собой обозначение на письме тонового артикля бамана. В статье [Выдрин 2008а] он уже обсуждался и обосновывалась необходимость его фиксации. Однако реальность такова, что в огромном большинстве имеющихся текстов артикли (как и тоны в целом) не обозначены; соответственно, расстановку артиклей должен осуществлять или парсер, или человек (очевидно, тот, кто занимается ручным снятием омонимии). Однако правила употребления артикля ясны

далеко не во всех деталях[8] (их выявлению препятствует как раз малое количество имеющихся текстов, в которых артикли были бы обозначены), так что программировать автоматическую расстановку артиклей в настоящее время невозможно – парсер отразил бы, в лучшем случае, неполноту наших знаний по этой части и создал бы ложную иллюзию у пользователя. Очевидно, расстановку артиклей мог бы осуществить носитель языка бамана (обученный соответствующим образом) или лингвист, обрабатывающий текст при содействии носителя языка. Поэтому надеяться на то, что в обозримом будущем удастся получить значительное количество текстов с обозначенными тоновыми артиклями, вряд ли приходится.

Как уже говорилось в статье [Выдрин 2008а], игнорирование тонового артикля ведёт, в частности, к неразличению разных видов синтаксических отношений в рамках именной группы, что весьма нежелательно. Однако в письменной практике бамана можно отметить следующую тенденцию: именные группы генитивного типа, характеризующиеся компактным типом связи между своими составляющими (лексический тон неначального компонента устраняется, тон первого компонента распространяется на всю синтагму), часто пишут слитно, в одно графическое слово. Иначе говоря, такие ИГ трактуются скорее как сложные слова. При всей теоретической спорности такой трактовки,[9] подобное написание имеет практический смысл, указывая на отсутствие тонового артикля у первого компонента ИГ. Орфографическое правило слитного написания для тонально-компактных ИГ такого типа можно было бы предложить, по крайней мере, для нетонированных текстов.

[8] Можно упомянуть в этой связи работу [Creissels 2009], где для манинка р-на Кита упоминаются контексты употребления и неупотребления артикля, о которых не шла речь ни в предыдущих работах этого автора по языкам манден, ни в трудах других исследователей. Проверка аналогичных контекстов в бамана показала, что по крайней мере некоторые сформулированные Кресельсом правила можно распространить и на этот язык.

[9] Отметим при этом, что Дени Кресельс в своей грамматике манинка Кита [Creissels 2009], а также Шарль Байоль [2000] идут ещё дальше, предлагая рассматривать как сложные слова и иные типы тонально-компактных единств (прежде всего, атрибутивную синтагму).

1.3. Усовершенствование практической орфографии бамана

Как это ни удивительно, при всём внимании к вопросам орфографии со стороны малийских лингвистов, этот интерес фокусируется в основном на составе графем и лишь минимально затрагивает проблем слитного/раздельного написания и использования дефиса [Guide 1979; Guide 1993]. Во всяком случае, никакого детального свода правил орфографии, где затрагивались бы эти вопросы, по-видимому, опубликовано не было.[10]

Очевидно, что для автоматического анализа текста унификация орфографии необходима. Можно рассмотреть вопрос и с другой стороны: разработка корпуса глоссированных текстов на бамана – это хороший повод для стандартизации орфографии.

В ходе работы над Корпусом весной 2010 года в международной электронной рассылке прошла дискуссия относительно правил использования дефиса в бамана. Участники дискуссии сошлись во мнении, что дефис следует употреблять в двух случаях:

– при редупликации глаголов и прилагательных со значением интенсивности: *tíɲɛ* ‘портить’ – *tíɲɛ-tíɲɛ* ‘портить сильно и много’; *ɲùman* ‘хороший’ – *ɲùman-ɲùman* ‘очень хороший’. Дефис не употребляется при немотивированной редупликации (т. е. в тех случаях, когда соответствующая нередуплицированная форма в языке отсутствует), например, *wòroworo* ‘шуметь (о ветре)’ (в отсутствие соотносимой лексемы **wòro*);[11]

– для соединения компонентов конгломерата, если эти компоненты сохраняют свои исходные тоны (см. примеры в предыдущем разделе).

В ходе дальнейшей работы, несомненно, будут сформулированы и другие правила.

[10] В устных беседах в июне 2010 года в Бамако малийские лингвисты говорили о том, что когда-то такие правила обсуждались и вырабатывались и что свод этих правил должен где-то храниться. Однако его местонахождение нам установить не удалось.

[11] Это правило – впрочем, сформулированное не вполне чётко – имеется и в упомянутых малийских изданиях [Guide 1978; Guide 1993].

1.4. Упорядоченное представление словоизменительной и деривативной морфологии

Составление полного списка словоизменительных морфем абсолютно необходимо для автоматического анализа текста; без этого невозможно опознание очень многих текстовых словоформ даже в таком преобладающе-изолирующем языке как бамана. Однако, с точки зрения задач парсинга, различие между словоизменительными и регулярными словообразовательными морфемами (такими как суффиксы причастий и номинализации, суффиксы диминутива и огмантатива и др.) оказывается несущественным: регулярные дериваты также, как правило, не даются в словаре, и парсер должен уметь членить их на аффиксы и основы, даже если и не ставить задачи полного морфологического разбора словоформ.

Таким образом, стала очевидной необходимость полного списка аффиксов бамана, с указанием их алломорфов и правил сочетаемости, а также их стандартных глосс на метаязыке. Составление инвентаря аффиксов существенно облегчалось тем, что словоизменение и деривация в бамана изучены достаточно хорошо (можно упомянуть, в первую очередь, работы [Dumestre 1987/1994, 187-233, 281-321; Dumestre 2003]). Оставалось лишь свести аффиксы в таблицы, более формально представить их алломорфы и снабдить каждый аффикс глоссой. Кроме того, была составлена таблица служебных слов и их унифицированных глосс. Отдельная задача – выявление допустимых и запретных сочетаний аффиксов друг с другом и с основами, что позволяет существенно сократить количество вариантов разбора, предлагаемых парсером. Для непродуктивных аффиксов были составлены максимально полные (в идеале – исчерпывающие) списки образуемых с их помощью дериват. Таблицы служебных морфем и слов, с предлагаемыми стандартными глоссами (см. приложение к данной статье), выносились на обсуждение в международную электронную рассылку, так что предлагаемые глоссы можно теперь считать, применительно к бамана, международным стандартом.[12]

[12] В дискуссии приняли активное участие Жерар Дюместр, Эрвин Эберман, Дмитрий Идиатов, Клаудиа Домбровски.

1.5. Представление композитов

Большую сложность для автоматического анализа текста на бамана представляет обилие композитов, образующихся по продуктивным моделям (особенно если последовательно применять правило слитного написания всех тонально-компактных комплексов, о котором шла речь в 1.2) – и, соответственно, не представленных в словаре. Если парсеру ставится задача не только вычленять деривативные и словоизменительные аффиксы, но и пытаться анализировать каждую словоформу как композит (при этом в каждой его компоненте, в свою очередь, также могут выделяться деривативные аффиксы), то количество теоретически допустимых вариантов морфологического разбора словоформы резко возрастает. Чтобы сократить их количество (и, таким образом, повысить качество работы парсера), необходимо найти и сформулировать реально существующие ограничения на словосложение – или попытаться исчислить допустимые в бамана модели словосложения. Попытка такого исчисления была сделана; её результаты не приводятся здесь лишь из соображений экономии места. Однако следует быть готовым к умеренной результативности работы по этой модели – она осложняется, во-первых, возможностью рекурсивности в применении моделей словосложения, во-вторых – уже упоминавшейся анархией в отношении словоделения в большинстве публикуемых текстов на бамана. В целом композиты, по-видимому, будут представлять одну из главных трудностей для автоматического анализа баманского текста.

1.6. Поморфемное глоссирование

Как правило, в электронных корпусах различных языков программа автоматического анализа выдаёт, в качестве конечного продукта, лемматизированный текст, т. е. такой текст, где каждая словоформа снабжена пометами, отражающими её словоизменительные характеристики. Это совершенно оправдано для индоевропейских и иных языков с развитым словоизменением и достаточно чёткими словесными границами.[13]

[13] К сожалению, я не знаком с практикой корпусной лингвистики для языков Дальнего Востока и Юго-Восточной Азии, типа китайского или

Бамана же относится к языкам, где (а) словоизменение минимально, причём крайне немногочисленные словоизменительные морфемы – это аффиксы, присоединяемые агглютинативно; (б) очень развито словосложение, при этом провести границу между сложным словом и словосочетанием часто очень непросто. В таких языках лемматизация оказывается малоэффективной.

Поэтому наша рабочая группа приняла решение о двух уровнях глоссирования – лексемном (с представлением словоизменения) и поморфемном, в котором будет систематически отражаться морфемный состав каждого графического слова бамана. Таким образом, во-первых, отчасти снимается проблема разграничения словосочетаний и композитов; во-вторых, пользователь корпуса получает возможность поиска не только по лексемам, но и по морфемам (как служебным, так и знаменательным).

2. Словарное обеспечение

2.1. Основной словарь

Шарль Байоль, автор наиболее популярного бамана-французского словаря, неоднократно переиздававшегося в Мали (последнее издание – [Bailleul 2007]), предоставил электронную версию этого словаря в формате Toolbox в распоряжение рабочей группы по созданию корпуса, что существенно облегчило её задачу. В то же время, довольно быстро стало очевидным, что для использования в качестве программного продукта для электронного корпуса бамана этот словарь нуждается в весьма существенной доработке. Перечислим те параметры, которые затронула эта доработка.

2.1.1. Орфографическая конверсия. В словаре Ш. Байоля используется авторская версия тоновой нотации: низкий тон обозначается (знаком грависа) над каждой гласной; высокотоновые слоги остаются без тональных диакритик; восходящий тон (который фактически является аллотоном низкой тонемы на односложном сегменте, если за ним следует другая

вьетнамского. Можно ожидать, что принятые для этих языков подходы могут быть схожи с теми, которые мы выработали для бамана.

низкотоновая тонема) маркируется гачеком. Существительные и прилагательные даются в своей «артиклевой» форме (т. е. с повышением тона на конце низкотоновых слов), глаголы – с тоновым контуром позиции перед паузой (без повышения тона на конце низкотоновых слов).

Автоматическая трансформация такой нотации в принятую у нас оказалась возможной только для слов, принадлежащим двум основным тональным классам – «высокотоновому» и «низкотоновому». Автоматизация конверсии для миноритарных тональных классов потребовала бы такого сложного алгоритма, что более простым решением оказалась ручная замена.

2.1.2. Фонетические варианты и отсылочные статьи. Даже в письменной форме «стандартного бамана», на которую, в первую очередь, ориентируется проект по созданию электронного корпуса текстов, сохраняется достаточно высокая вариативность. Так, многие корни могут выступать в виде вариантов Ciɲɛ и Ciyɛn (*tìɲɛ ~ tìyɛn* ‘правда’, *bíɲɛ ~ bíyɛn* ‘печень’ и т. д.);[14] неустойчивой может быть назализация (*díla ~ dílan* ‘изготавливать’, *bùnte ~ bùnten* ‘размалывать в муку’ и т. д.) и гласные (*ɲé ~ ɲá* ‘глаз’, *mɔ̀gɔ ~ màa* ‘человек’) – при этом одна лексема может иметь достаточно большое количество вариантов. Конечно, в большинстве случаев фонетические варианты – диалектного происхождения,[15] однако их встречаемость в текстах на бамана (как устных, так и письменных) требует приведения таких вариантов в словаре. В то же время, учёт в словаре всех диалектных вариантов невозможен: во-первых, очень многие слова будут представлены в таком случае десятками вариантов, что, к тому же, резко увеличит омонимию и затруднит парсинг. Во-вторых, надеяться на полное представление в словаре всех диалектных вариантов всё равно не приходится – хотя бы потому, что диалекты бамана для этого недостаточно полно описаны. В-

[14] Следует отметить, что Guide de transcription рекомендует в данном случае формы типа Ciɲɛ, однако на практике авторы текстов очень часто используют и альтернативные варианты.

[15] В «стандартном бамана» за основу взят диалект столицы Мали, Бамако, однако выходцы из различных районов страны привносят формы из своих диалектов, что облегчается слабостью политики кодификации языка со стороны государственных органов.

третьих, в диалектном континууме манден трудно провести границы между говорами бамана, манинка, дьюла и т. д., так что стороннику тотального включения диалектных вариантов в словарь было бы провести границу между языками.

В словаре Байоля последовательно представлены формы трёх локальных диалектов бамана; разумеется, отражён и стандартный бамана. Наша рабочая группа приняла решение сохранять имеющиеся в словаре варианты, но с некоторыми оговорками. В частности, иногда словарь Байоля даёт формы из периферийных диалектов, появление которых в текстах на стандартном бамана маловероятно, при этом такие формы создают омонимию с употребительными словами. Например, среди форм лексемы *dílа, dílan* ‘изготавливать’ даётся и южная форма *bílа.* Последняя оказывается омонимичной (в отсутствие тоновой нотации) весьма употребительному глаголу *bìla* ‘класть’. Если учесть, что словоформа *bìla* может быть также проанализирована как сочетание основы *bì* (диалектный вариант глагола *bìn* ‘падать’) с суффиксом прогрессива *-la,* то количество вариантов анализа каждой встретившейся в тексте словоформы *bila* превосходит все рамки здравого смысла. В то же время, словарь Байоля не даёт аналогичные диалектные формы для многих других слов, например, *blɔ̀* (стандартный бамана: *dɔ̀lɔ*) ‘пиво’, *blɔ̀ki* (стандартный бамана: *dùlɔkī*) ‘рубаха’ и др. В этой ситуации представляется предпочтительным убрать такие диалектные формы, которые сильно увеличивают «шум» и затрудняют работу парсера.

В словаре Байоля принят принцип подачи каждого фонетического варианта на своём алфавитном месте в виде особой статьи, с отсылкой к основной статье. Впрочем, в реальности тут много непоследовательного: по своему оформлению отсылочные статьи часто мало отличаются от основных и содержат полный набор информации о лексеме; иногда отсылочная статья содержит информацию, в главной статье не представленную. Не так уж редко лексема, имеющая фонетические варианты, оказывается представлена в словаре двумя полноценными статьями, не содержащими эксплицитных отсылок друг к другу.

Надо сказать, что для парсера отсылочные статьи не нужны вовсе, поскольку он может осуществлять поиск по всем

фонетическим вариантам, упомянутым в основной статье, без обращения к отсылочной статье. Более того – упоминание фонетического варианта и в основной статье, и в отсылочной лишь осложняет его работу, продуцируя «фиктивную омонимию», поскольку парсер учитывает оба упоминания этого варианта (в главной и в отсылочной статьях).

В такой ситуации наиболее простым путём для упорядочивания информации было признано уничтожение всех отсылочных статей (а также дублирующих статей), с обязательным перенесением, в случае необходимости, всей содержательной информации в главную статью.

2.1.3. Подбор французских эквивалентов и проблема полисемии. Выбор переводного эквивалента при глоссировании нередко оказывается весьма непростым делом, особенно если идёт речь о большом корпусе текстов. Изначально рабочая группа приняла технически простое решение: если в статье в словаре Байоля поле \ge (предназначенное для французского эквивалента) встречается более одного раза,[16] то программа берёт в качестве глоссы для баманской лексемы содержимое первого по порядку поля \ge. При этом исходили из того, что при описании семантики полисемичного глагола лексикограф ставит на первое место, по умолчанию, наиболее прототипическое значение лексемы, из которого легче всего вывести все остальные.

В ходе дальнейшей работы выявились две главные трудности; первая из них (подбор эквивалента) – субъективного характера, вторая (проблема полисемии) – объективного.

2.1.3.1. Подбор эквивалента. В словаре Байоля (как, впрочем, и в очень многих других) граница между толкованием значения и собственно эквивалентом на метаязыке (т. е., в идеале, – слова, которое можно использовать в тексте на языке перевода) оказывается нечёткой, а распределение информации по полям базы данных (которой является программа Toolbox) – довольно

[16] Toolobox представляет собой базу данных, приближенную по структуре к текстовому процессору; в частности, в пределах одной карточки может быть несколько полей с одинаковым названием. Эта особенность программы даёт большую свободу маневра лексикографу, но сильно затрудняет конвертацию словарного файла в формате Toolbox в форматы других баз данных, более строгих по структуре.

произвольным. Очень часто обнаруживается, что в первом по счёту поле \ge оказывается не один эквивалент, а два, например:

\lx nɔ̀.ra.da
\va nɔ̀nada
\va nwána
\ps n
\ge cadet, puîné

Там же может оказаться, помимо эквивалента, также и толкование или его часть – при этом вторая часть толкования оказывается нередко перенесённой во второе поле \ge:

\lx npàana
\va pàana
\ps v
\ge écarter (les jambes
\ge les bras ...)

Наконец, предлагаемый автором словаря эквивалент может быть просто слишком длинным и потому неудобным для глоссирования текста:

\lx npóko
\va nfúku
\ps n
\ge taon noir à la piqûre cuisante

Эти и некоторые другие особенности организации исходного словаря заставили думать о необходимости его тотального просмотра и доработки с точки зрения потребностей парсера. В результате интенсивной работы всей рабочей группы в июле-августе 2010 г. было проведено упорядочивание словаря по перечисленным выше параметрам, а именно: в первое по порядку следования поле \ge внесён один эквивалент, по возможности краткий[17] и представляющий прототипическое значение лексемы,

[17] В тех случаях, когда подобрать односложные эквиваленты оказалось невозможным, части многосложного эквивалента даются без пробела и разделяются точками, например:
\lx jáfulate
\ge arbre.Hannoa.undulata.

а все остальные данные из этого поля устранены; устранены отсылочные статьи.

2.1.3.2. Проблема полисемии. Если словарь показывает, что идентифицированное в бамана слово полисемично, то встаёт вопрос выбора между его значениями. Какое из значений должна отражать глосса? Всегда ли использовать в качестве глоссы данной лексемы один и тот же эквивалент или, в зависимости от контекста, использовать разные эквиваленты (отражающие разные значения)?

Технически несравненно проще считать одну глоссу «постоянным представителем» одной лексемы, в каком бы из своих значений эта лексема ни выступала в тексте. Исходя из потребностей глоссирования этого типа и проводилась адаптация электронной версии словаря Байоля. Иное решение потребовало бы разработки семантически чувствительного парсера, что практически эквивалентно созданию достаточно совершенной программы машинного перевода с бамана на французский. Конечно, о такой задаче можно и нужно думать, но вряд ли она стоит в ближайшей повестке дня.

Возможно и компромиссное решение (хорошо известное в компьютерно-интернетовской практике): лексема всегда представлена одной и той же глоссой, но пользователю предлагается опция «показать полисемию», при выборе которой во всплывающем окне показываются все зафиксированные в словаре значения лексемы (иначе говоря, содержимое всех полей \ge, имеющихся в словарной карточке). Так, для глагола *dún* в качестве основной глоссы фигурирует *manger,* а при включении опции «показать полисемию» будут продемонстрированы также значения *dépenser, rouler qn.*

С точки зрения устройства парсера такое решение не представляет особых трудностей, но оно требует значительно более глубокой доработки словаря, чем та, которая была осуществлена на настоящем этапе, поскольку некоторые лексемы бамана имеют многие десятки значений, а их подача в словаре Байоля пока что очень далека от той, которая необходима для автоматизированного представления полисемии. Доработку словаря в этом направлении имеет смысл планировать на следующем этапе работы (предположительно, в 2011-2012 гг.).

2.1.3.3. Поморфемное членение. В словаре Байоля лексемы-дериваты и композиты обычно даются с указанием членения на морфемы, а в специальном поле, \lt, приводится покомпонентный перевод. Однако при более тщательном рассмотрении оказалось, что

1) морфемное членение приводится далеко не всегда – нередко в слове указывается только одна морфемная граница из двух или трёх (*màakɔrɔ.ba* ‘vieillard’ – ср. полное членение: *màa.kɔrɔ.ba*), и достаточно систематически не приводится морфемное членение в фонетических вариантах лексемы (что, действительно, может считаться избыточным для «бумажной» версии словаря, но совершенно необходимо для парсинга) – например:

\lx màa.dolo
\va mɔ̀gɔdolo
\ge Orion;

2) иногда вычленяемые автором словаря знаменательные морфемы не представлены в словаре – таким образом, они оказываются «отсылками в никуда»;

3) предлагаемый в поле \lt покомпонентный перевод плохо соотносится с эквивалентами вычленяемых знаменательных морфем (см. раздел 2.1.3.1.).

Таким образом, мы пришли к необходимости второй систематической переработки словаря, которая и была осуществлена силами нашей группы в сентябре 2010 года. В результате все лексемы в словаре (в каждом из своих фонетических вариантов) теперь представлены с полным морфемным членением, при этом каждая вычлененная корневая морфема снабжена стандартным переводным эквивалентом, совпадающим с тем её эквивалентом, который даётся в основной статье, посвящённой этой морфеме. В качестве эквивалентов деривационных морфем даны стандартные глоссы из списка, который приводится в Таблице 3 в Приложении.

2.2. Дополнительные словари

В словаре Ш. Байоля представлены, за единичными исключениями,[18] только нарицательные существительные языка бамана. При этом очевидно, что в текстах имена собственные составляют достаточно большой процент всех словоупотреблений. На момент начала работы над Корпусом у меня имелись словари географических названий, личных имён и клановых имён бамана, в основном в рукописной форме. А. В. Давыдов осуществил компьютерный набор этих словарей (в формате Toolbox), а в ходе экспедиции в Мали в июне-июле 2010 года протонировал их.[19] На данный момент эти словари ни в коей мере не претендуют на исчерпывающий характер (насколько вообще возможно говорить о достижении предела в расширении таких словарей), они будут пополняться в ходе работы по ручному снятию омонимии.

Только предстоит создать словарь аббревиатур (отметим, что большинство аббревиатур, встречающихся в баманских текстах, – французские, а не собственно баманские: SIDA – syndrome de l'immunodéficience acquise, CMDT – Compagnie malienne du développement des textiles, ODIPAC – Office de Développement Intégré pour les Productions Arachidières et Céréalières и т. д.)

Ещё одна категория словоупотреблений, которые являются источником трудностей для парсинга, – неадаптированные французские слова (при том что адаптированные заимствования, по-видимому, следует включать в основной словарь). Для их частичной идентификации предполагается использовать метод поиска нетипичных в языке бамана позиций и сочетаний графем (сочетание двух гласных; согласные в конце слова и т. п.).

2.3. Пополнение словарей в ходе ручной разметки Корпуса

Странно было бы ожидать, что все лексемы из текстов бамана, включаемых в Корпус (даже если не учитывать

[18] Такими как традиционные «порядковые имена» детей у бамана (*Ncì, Ngòlo ~ Ŋòlo, Nsǎn* и т. д.) и, окказионально, прозвища некоторых кланов (*nɔ̀gɔ.ka* appellation des « Tarawele »).

[19] Нужно иметь в виду, что имена собственные у народов манден допускают повышенную тоновую вариативность. Впрочем, для автоматического анализа нетонированных текстов это обстоятельство дополнительных трудностей не создаёт.

неадаптированных иностранных слов), будут содержаться в уже имеющихся словарях. Поэтому предполагается, что работа над Корпусом станет важнейшим источником пополнения словаря языка бамана. Это пополнение может осуществляться на этапе ручного снятия омонимии в текстах – т. е. на том этапе, который следует за метаразметкой и автоматическим парсингом. Поскольку ручное снятие омонимии, по крайней мере на начальных этапах работы, предполагается проводить силами российских (или, шире – европейских) студентов и специалистов по языку бамана, т. е. теми, для кого бамана не является родным языком, можно предвидеть, что создание новых словарных статей в словаре может вызвать у них затруднение. По-видимому, имеет смысл предусмотреть такой алгоритм работы: 1) устанавливается, что слово, не опознанное парсером, не является скорее всего именем собственным, аббревиатурой, иностранным словом или результатом опечатки; 2) такое слово вносится в некий временный словарь; 3) слова из временного словаря (в контекстах, в которых они встретились в текстах Корпуса) проверяются с информантами, для которых язык бамана является родным, после чего принимается решение о внесении (или невнесении) их в основной словарь.

3. Структурирование анализируемого текста: Уровни представления текста и глоссирования

Обработанный и глоссированный текст – это то, с чем, в обычном случае, будет иметь дело пользователь Корпуса. Рассмотрим, каким образом предполагается организовать этот текст. Сразу оговоримся, что:

а) Корпус будет открытым для доступа любому пользователю Интернета;

б) пользователь не будет иметь доступа к полным текстам документов, включённым в Корпус (это ограничение связано с охраной авторских прав);

в) не планируется устанавливать ограничений на количество фразовых примеров, которые пользователь получает по запросу (ср. практику подобных ограничений, скажем, в Британском Национальном Корпусе). Доступ к полному списку примеров,

обнаруженных в Корпусе, необходим для углублённых исследований.

Всякий текст в Корпусе будет представлен на нескольких уровнях анализа.

1) Исходный вид. Текст воспроизводится в том виде, в котором он представлен в источнике – с сохранением орфографии, пунктуации, опечаток и описок. Это необходимо для осуществления контроля: если программа-парсер или разметчик (человек, осуществляющий ручное снятие омонимии) допускает ошибку (например, принимает французское вкрапление за баманское слово, написанное с опечаткой), эта ошибка может быть обнаружена при обращении к исходной форме текста. Кроме того, особенности текста, в том числе опечатки и пунктуация, могут сами по себе являться предметом исследования лингвиста, и было бы неразумным закрыть эту возможность для пользователей Корпуса.

2) Запись в «нормализованной орфографии», с тоновой нотацией. При переходе на этот уровень осуществляется автоматическая конвертация старой орфографии в новую, ручное исправление орфографических ошибок, автоматическая идентификация словоформ, обозначение тонов в соответствии с принятыми принципами. Если в исходном тексте тоны указаны, то осуществляется автоматическое преобразование исходной тоновой нотации в ту, которая принята в Корпусе.

3) Представление текста с вычленением словоизменительных морфем.

4) Представление с полным поморфемным разбиением (отделение словообразовательных морфем, расчленение композитов на составляющие).

5) Представление текста с синтаксической разметкой: обозначение границ именных групп; связывание финитных глаголов с предикативными показателями; связывание глаголов (финитных и нефинитных форм) с управляемыми ими послелогами; обозначение границ клауз и т. д. Этот уровень представления предполагается обеспечить на более поздних этапах проекта.

6) Строка лемматизации: каждой лексеме и каждой словоизменительной морфеме бамана дан в соответствие французский эквивалент.

7) Строка глоссирования: каждой морфеме бамана (как словоизменительной, так и словообразовательной) дан в соответствие французский эквивалент.

8) Литературный перевод на французский.

Примечание: 1) В представлении пользователю уровни со 2 до 5 могут быть, по-видимому, объединены без ущерба для содержания. 2) Литературный перевод текста на французский может быть добавлен только вручную.

Корпус бамана планируется сделать неоднородным по степени анализа. Наименьшую его долю будут составлять тексты со снятой вручную омонимией и с проставленными тоновыми артиклями (как уже отмечалось, расстановка артиклей должна производиться теми, для кого язык бамана является родным, или, во всяком случае, с участием таких информантов). Небольшим будет также подкорпус с литературным переводом на французский. Следующий, более широкий круг, будет являть собой подкорпус со снятой вручную омонимией. Наконец, все остальные тексты в Корпусе будут только автоматически обработаны парсером; даже при сохранении неснятой омонимии такие тексты могут дать пользователю Корпуса много полезной информации.

Соответственно, пользователь сможет осуществлять поиск только по каким-то из этих подкорпусов или по всему корпусу в целом – в зависимости от того, нужно ли ему максимально возможное количество примеров (какое-то количество которых при этом может оказаться неправильным) или он предпочитает получить меньшее количество более надёжных примеров (без «шума»).

4. Некоторые перспективы проекта «Корпус текстов бамана»

В июне-июле 2010 мы с А. В. Давыдовым совершили поездку в Гвинею и Мали, главной целью которой был сбор материалов для Корпуса, а также налаживание контактов с лингвистами (и другими заинтересованными кругами) этих стран, которые могли бы быть полезными в ходе работы над проектом. Попытаюсь обобщить впечатления от этой поездки.

4.1. Мали

Реакция лингвистов была позитивной; идею создания Корпуса поддержали все наши собеседники. Особенно заинтересовала их перспектива использовать результаты корпусного проекта для упорядочивания орфографии языка бамана и, в перспективе, для создания программы автоматической проверки орфографии. Другое дело, что на нынешнем этапе участие малийцев в работе над проектом может быть лишь весьма ограниченным – в частности, требуется их помощь в получении электронных версий книг и газет, публикуемых на бамана. В дальнейшем, когда удастся добиться некоего минимального уровня качества работы парсера и приступить к созданию полноценного корпуса текстов, они могут быть привлечены к снятию омонимии. Очень желательной была бы помощь малийцев в транскрибировании аудиозаписей – это позволило бы создать подкорпус устной речи бамана.

Очевидно, что для налаживания сотрудничества в этой области потребуются дополнительные финансовые ресурсы, превышающие рамки исследовательского гранта РФФИ.

4.2. Гвинея

В Конакри, столице страны, и в Канкане, административном центре населённой манинка области Верхняя Гвинея, мы провели серию встреч с гвинейскими лингвистами, а также с активистами культурно-образовательного движения н’ко. В частности, мы присутствовали на специальном заседании *Ńkó*` *Dúnbu*` ‘Академии н’ко’, по своим функциям сходной с Французской Академией. Члены Академии занимаются регламентацией орфографии н’ко, а также проводят большую лексикографическую работу: пополняют одноязычный словарь манинка (первое издание которого насчитывает около 32 500 словарных статей), готовят к изданию н’ко-французский словарь; они переводят на н’ко законодательные тексты Гвинейской Республики и т. д. – причём вся эта работа проводится без какого бы то ни было финансирования со стороны государственных органов или международных организаций.

Наш рассказ о проекте Корпуса вызвал у членов Академии большой энтузиазм; они выразили свою готовность к сотрудничеству. Но в данном случае речь идёт не просто о работе

с текстами на другой графической основе, но и с другим языком: языки маninка и бамана, хотя и близки друг к другу, различаются всё же достаточно сильно для того, чтобы парсер и словарь бамана можно было применять к текстам на маninка (тем более если говорить о работе с текстами на «литературном н'ко», который отличается от письменного «стандартного бамана» ещё больше, чем разговорные варианты маninка и бамана). При этом, несомненно, наработки по баманскому корпусу сильно облегчат процесс создания корпуса маninка.

Если всё же иметь в виду перспективу создания корпуса текстов на маninка, то необходимо иметь в виду препятствия, которые имеются на этом направлении. На настоящий момент можно, в частности, упомянуть (помимо, само собой разумеется, проблемы получения финансирования) следующие трудности:

– плохая обеспеченность Конакри (и, тем более, других гвинейских городов) электроэнергией, что существенно затрудняет работу с компьютером;

– отсутствие маninка-французского словаря (аналогичного бамана-французскому словарю Шарля Байоля); это означает, что такой словарь надо создавать заново.

5. Заключение

В целом можно отметить, что работа над электронным корпусом текстов бамана пока что идёт по оптимистическому сценарию:

– к моменту написания данной статьи практически готова первая рабочая версия парсера и необходимый для её функционирования словарь;

– ясны конкретные задачи по совершенствованию этих инструментов, стоящие перед рабочей группой;

– имеется достаточно большое количество текстов в электронном виде, готовых для введения в Корпус;

– работа над Корпусом встречает понимание и поддержку коллег из разных стран, что открывает хорошие перспективы для международного сотрудничества в данной области.

Литература

Выдрин В. Ф. На пути к электронному корпусу языка бамана: обозначение тонов // Труды международной конференции «Корпусная лингвистика – 2008». СПб.: Санкт-Петербургский государственный университет, 2008а. С. 122–134.

Выдрин В. Ф. Электронные корпуса африканских языков: завтра или послезавтра? // А. Ю. Желтов (ред.). Петербургская африканистика. Памяти Андрея Алексеевича Жукова. СПб.: Санкт-Петербургский государственный университет, 2008б, С. 29–39.

Давыдов А. В. Электронный корпус языка бамана: Комплектование и принцип метатекстовой разметки // Настоящий сборник.

Bailleul Ch. Cours pratique de bambara. Bamako, Editions Donniya, 2000.

Bailleul Ch. Dictionnaire Bambara-Français. 3e édition corrigée. Bamako : Donniya, 2007.

Creissels D. Le malinké de Kita. Köln: Rüdiger Köppe Verlag, 2009.

Davydov A. Towards The Manding Corpus: Texts Selection Principles and Metatext Markup. Eds. Guy De Pauw, H. J. Groenewald, and Gilles-Maurice de Schryver. Proceedings of the Second Workshop on African Language Technology (AfLaT 2010). Valletta, Malta: European Language Resources Association (ELRA), 2010, P. 59–62. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W5.pdf

Dumestre G. Le bambara du Mali: Essai de description linguistique. Thèse de Doctorat d'Etat. INALCO. Paris, 1987. 2e édition : Paris : Les Documents de Linguistique Africaine, 1994, Tomes 1, 2.

Dumestre G. Grammaire fondamentale du bambara. Paris: Karthala, 2003.

Guide de transcription et de lecture du Bambara. Bamako: DNAFLA, 1979.

Guide de transcription et de lecture du Bambara. 2e édition, revisé par Demba Konarɛ, Moussa Diaby, Soumana Kanɛ. Bamako: DNAFLA, 1993.

Vydrin V. Glossed electronic corpora of Mande languages: A perspective that we cannot avoid // Mande languages and linguistics. 2nd International Conference, St. Petersburg (Russia), September 15–17, 2008. Abstracts and Papers. V.Vydrin (ed.). St. Petersburg, 2008, P. 15–22.

Приложение. Стандартные глоссы для аффиксов и служебных слов бамана

Таблица 2. Словоизменение

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Значение**
-la/-na	*-na* после носового гласного, *-la* в остальных случаях	PROG	v	суффикс прогрессива (в сочетании с pm *bɛ́* (утвердительная конструкция), *tɛ́* (отрицательная конструкция))
-ra/ -la/ -na	*-na* если предшествующий слог содержит назальный звук; *-la* если предшествующий слог неносовой и содержит *r, l;* *-ra* во всех остальных случаях	IPFV.INTR	v (v.i)	показатель имперфектива для интранзитивных глаголов
-w		PL	n, adj, dtm, prn	показатель множественного числа (присоединяется к последнему слову ИГ)
-`	проявляется как даунстеп последующего высокотонового слова	ART	n, adj	артикль (присоединяется к последнему слову ИГ)

Таблица 3. Словообразование

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Часть речи деривата**	**Значение**
		Суффиксы			
-ba		AUGM	n, adj, ptcp	= исходной	аугментатив
-baa/-baga	варианты *-baa и -baga* в свободном варьировании	AG.OCC	v	n, (adj)	имя окказионального деятеля
-bali		PTCP.PRIV	v	ptcp	привативное причастие
-ka		GENT	n	n	суффикс имени жителя какого-л. места или выходца из этого места («гентильный»)
-la/-na	*-na* после носового гласного, *-la* в остальных случаях	AG.PRM	v	n	суффикс имени деятеля
-la/-na	*-na* после носового гласного, *-la* в остальных случаях	LOC	n	n	суффикс имени места
-la/-na	*-na* после носового гласного (факультативно – и после слога «носовой согласный + неносовой гласный») , -	PRICE	num	n	суффикс имени стоимости («количество товара стоимостью в X»)

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Часть речи деривата**	**Значение**
	la в остальных случаях				
-la/-na	распределение между алломорфами отчасти факультативное, отчасти лексикализованное	MNT1	v, n, pp	n	суффикс имени ментальной деятельности или её результата
-lata/ -nata	распределение между алломорфами отчасти факультативное, отчасти лексикализованное; отличия от MNT1 минимальны	MNT2	v, n, pp	n	суффикс имени ментальной деятельности или её результата
-lama/ -nama	*-nama* после носового гласного, *-lama* в остальных случаях	STAT	n	adj	суффикс отыменных прилагательных со значением «под видом X», «в качестве X», «сделанный из X», «будучи X»
-lan/ -nan	*-nan* после носового гласного, *-lan* в остальных случаях; *-ran* – редкий лексически распределённый вариант *-lan*	INSTR	v	n	суффикс имени инструмента
-len/ -nen	*-nen* после носового гласного, *-len* в	RES	v	ptcp	суффикс результативного причастия

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Часть речи деривата**	**Значение**
	остальных случаях				
-li/-ni	*-ni* после носового гласного, *-li* в остальных случаях	NMLZ	v	n	суффикс отглагольного имени
-ma		COM	n	adj, (n)	суффикс отыменного прилагательного с комитативным/орнативным значением
-ma		RECP.PRN	n	n	суффикс взаимности отношений
-ma		DIR	v	v	непродуктивный суффикс, сочетающийся главным образом с основами глаголов направленного действия, часто не меняя исходного значения
-man		ADJ	vq	adj	адъективизатор квалитативных глаголов
-nan		ORD	num	adj	суффикс порядковых числительных
-nin		DIM	n, adj, ptcp	= исходной	диминутив
-ntan		PRIV	n	adj, (n)	суффикс отыменного привативного прилагательного
-nci		AG.EX	n, adj, v	n	суффикс «имени неумеренного

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Часть речи деривата**	**Значение**
					деятеля»
-ɲɔgɔn/ -ɲwan/ -ɲwaan	варианты – разного диалектного происхождения	RECP	v, n	n	суффиксоид «имени партнёра по деятельности»
-ta		PTCP.POT	v	ptcp	суффикс причастия с потенциальным значением
-tɔ		PTCP.PROG	v	ptcp	суффикс прогрессивно-проспективного причастия (прогрессив – от непредельных глаголов, проспектив – от предельных)
-tɔ		PTCP.ST	n	n, adj	имя субъекта состояния (чаще – неблагоприятного)
-ya		DEQU	vq	n, v	суффикс, образующий динамические глаголы и имена качеств от квалитативных глаголов
-ya		ABSTR	n, adj, (v)	n, (v)	суффикс имени статуса или состояния (от имён, обозначающих лиц и некоторых животных), имени качества (от производных прилагательных); (редк.) суффикс глаголов с инхоативным значением

Аффикс	**Алломорфия**	**Глосса**	**Частеречная сочетаемость**	**Часть речи деривата**	**Значение**
Глагольные префиксы					
lá-/ná-	*ná-* факультативно после носового гласного, *lá-* в остальных случаях	CAUS	v	v	каузативный префикс (часто – с лексикализованным нерегулярным значением)
mà- ~ màn-	алломорф *màn-* только в единичных глаголах	SUPER	v	v	префикс с затемнённой семантикой (этимологически, очевидно, суперэссивной)
rá-/rɔ́-	не в стандартном бамана; фонетические варианты – разного диалектного происхождения	IN	v	v	префикс с затемнённой семантикой (этимологически, очевидно, инэссивной)
sɔ̀-		EN	v	v	непродуктивный префикс (3 глагола перемещения), восходит к слову *sɔ̀n* ‘сердце’

Комментарии:
В графе «Частеречная принадлежность деривата» в скобках указывается второстепенное образование по конверсии (более или менее лексикализованное).

Таблица 4. Служебные слова

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
à	3SG	pers	любая ИГ	неэмфатическое местоимение 3 лица ед. числа	
á	2PL	pers	любая ИГ	неэмфатическое местоимение 2 лица мн. числа	
ánw	1PL.EMPH	pers	любая ИГ	эмфатическое местоимение 1 лица мн. числа	
án	1PL	pers	любая ИГ	неэмфатическое местоимение 1 лица мн. числа	
áw	2PL.EMPH	pers	любая ИГ	эмфатическое местоимение 2 лица ед. числа	
bɛ́	BE	cop	после ИГ подлежащего	копула неглагольного локативного предложения	
bɛ́ ~ bí ~ bé	IPFV.AFF	pm	после ИГ подлежащего	показатель утвердительного имперфектива	диалектные варианты
bɛ́ kà	PROG.AFF	pm	после ИГ подлежащего	показатель	

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
				утвердительного прогрессива	
bɛ́ka ~ bɛ́ga ~ bága ~ búga	PFV.TR.AFF	pm	после ИГ подлежащего	показатель перфектива при переходном глаголе	редкая диалектная форма, синоним *yé*
bɛ́nà ~ bínà ~ bénà	FUT.AFF	pm	после ИГ подлежащего	показатель утвердительного будущего	
bìlen ~ bìle ~ bèlen	COND.NEG	pm	после ИГ подлежащего; иногда сопровождается предикативным показателем *yé* или *má*	показатель отрицательного условного наклонения	архаичный и редкий показатель
dè	FOC	prt	после фокализуемого слова	показатель контрастивного фокуса	
dòn	PRES	cop	после ИГ подлежащего	копула неглагольного презентативного предложения	
dùn	TOP.CNTR	prt	следует за ИГ субъекта или иной ИГ, вынесенной в крайне левую позицию	показатель контрастивной топикализации подлежащего	
é`	2SG.EMPH	pers	любая ИГ	эмфатическое	

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
				местоимение 2 лица ед. числа	
í	2SG	pers	любая ИГ	неэмфатическое местоимение 2 лица ед. числа	
í	REFL	pron	любая несубъектная ИГ; субъектная ИГ придаточного предложения	рефлексивное местоимение	
ìn	DEF	dtm	стоит после ИГ	«новый определённый артикль»	
kà	INF	pm	перед ИГ прямого дополнения; в её отсутствие – перед глаголом	показатель инфинитива	
ká	OPT	pm	после ИГ подлежащего	показатель оптатива	
ká	POSS	conj	после ИГ посессора	посессивная связка	
ká	QUAL.AFF	pm	после ИГ подлежащего	показатель утвердительного квалитативного предложения	
kàná ~ kánà	PROH	pm	после ИГ подлежащего	показатель прохибитива	
kɔ̀ni	TOP	prt	после топикализуемой ИГ	показатель контрастивного топика	

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
mà ~ màa	DES	pm	после ИГ подлежащего, представленной словом *Ala* ‘Бог’; глагол присоединяет суффикс *–ra/-la/-na* PFV.INTR	предикативный показатель в предложении, обозначающем благопожелание	
má	PFV.NEG	pm	после ИГ подлежащего	показатель отрицательного перфектива	
mán	QUAL.NEG	pm	после ИГ подлежащего	показатель отрицательного квалитативного предложения	
mána ~ máa	COND.AFF	pm	после ИГ подлежащего	показатель утвердительного кондиционалиса	*máa* – форма в северных диалектах
mín	REL	dtm, pron	после релятивизируемой ИГ в левосторонней придаточной клаузе; в позиции ИГ в правосторонней придаточной клаузе	маркер релятивизации	
nà ~ ná	CERT	pm	после ИГ подлежащего	показатель уверенного будущего	

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
nìn	DEM	dtm, pron	вместо, перед или после ИГ	указательное местоимение	
nɔ́gɔn	RECP	pron	любая несубъектная ИГ	взаимное местоимение	
ó`	DISTR	conj	между двумя ИГ	показатель дистрибутивной связи	
ò	ANAPH	pron	замещает ИГ	анафорическое местоимение	
òlú	ANAPH.PL	pron	замещает ИГ	плюральное анафорическое местоимение; эмфатическое местоимение 3 л. мн. ч.	
tɛ́	COP.NEG	cop	после ИГ подлежащего	копула неглагольного отрицательного локативного предложения	
tɛ́ ~ tí ~ té	IPFV.NEG	pm	после ИГ подлежащего	показатель отрицательного имперфектива	диалектные варианты
tɛ́ kà	PROG.NEG	pm	после ИГ подлежащего	показатель отрицательного прогрессива	
tɛ́ka ~	PFV.TR.NEG	pm	после ИГ подлежащего	показатель	редкая диалектная

Форма	**Глосса**	**Часть речи**	**Позиция**	**Значение**	**Алломорфия**
tɛ́ga				отрицательного перфектива при переходном глаголе	форма, синоним *yé*
tɛ́nà ~ tɛ́nà ~ tínà	FUT.NEG	pm	после ИГ подлежащего	показатель отрицательного будущего	
tùn	PST	prt	чаще всего перед pm или cop	показатель ретроспективного сдвига	
wà	Q	prt	в конце предложения	частица общего вопроса	
yé	PFV.TR	pm	после ИГ подлежащего	показатель утвердительного переходного перфектива	
yé	EQU	cop	после ИГ подлежащего	копула в эквативном неглагольном предложении	
yé	IMP	pm	следует за ИГ подлежащего, выраженного местоимением 2 мн.	показатель императива при подлежащем во 2 мн.	
yé kà	RCNT	pm	после ИГ подлежащего	показатель недавнего прошлого	малоупотребительный