В.Ф. Выдрин

НА ПУТИ К ЭЛЕКТРОННОМУ КОРПУСУ ЯЗЫКА БАМАНА: ОБОЗНАЧЕНИЕ ТОНОВ[1]

0. Сегодня, говоря о целостном описании языка, обычно подразумевают три компонента: полноценную грамматику, словарь и корпус текстов – конечно же, при этом имеется в виду именно электронный корпус. Эти компоненты теснейшим образом взаимоувязаны, и отсутствие одного из них неизбежно негативно сказывается на качестве остальных.

Бамана – один из «больших» языков Африки. Число говорящих на нём (в том числе как на втором языке) приближается к 10 млн., и ещё порядка 15 млн. человек говорят на близкородственных, взаимопонимаемых с бамана языках маninка и дьюла. В настоящее время имеются достаточно подробные грамматические описания бамана – в первую очередь, в этой связи нужно упомянуть диссертацию и грамматику Жерара Дюместра[2], а в целом счёт лингвистических публикаций по этому языку идёт на сотни. Имеются и неплохие словари – в первую очередь, Шарля Байоля[3], Жерара Дюместра[4], В.Ф. Выдрина и С.И. Томчиной[1] – правда, оба последних словаря опубликованы лишь частично.

---

[1] Данная работа выполнена в рамках проекта «Интегральное описание южных языков манде: словари, грамматики, корпуса глоссированных текстов», финансируемого РГНФ (номер проекта 08-04-00144а).

[2] *Dumestre G.* Le bambara du Mali: Essai de description linguistique: Thèse de Doctorat d'Etat / INALCO. Paris, 1987. 2e édiction: Paris: Les Documents de Linguistique Africaine, 1994. Tomes 1, 2; *Dumestre G.* Grammaire fondamentale du bambara. Paris: Harmattan, 2003.

[3] *Bailleul Ch.* Dictionnaire Français-Bambara. Bamako: Editions Donniya, 1997; *Bailleul Ch.* Dictionnaire Bambara-Français: 3e éd. Bamako: Editions Donniya, 2000.

[4] *Dumestre G.* Dictionnaire bambara-français. Fasc. 1–9 (A–N). Paris, 1981–1992.

Таким образом, главной лакуной остаётся электронный корпус бамана. При этом весьма желательно, чтобы такой корпус был глоссированным (языком глоссирования должен быть английский и/или французский, – последний является официальным языком Мали) – несомненно, отсутствие глоссирования резко снизило бы число тех, кто реально сможет пользоваться таким корпусом.

Очевидно, что создание представительного корпуса возможно лишь для развитого языка, на котором существует достаточное количество письменных текстов и который достиг определённой степени стандартизации – иначе говоря, необходимо, чтобы язык стал, хотя бы в какой-то степени, литературным. В ином случае невозможно говорить даже о самом скромном электронном корпусе (от 1 до 10 млн. слов). В какой степени этому условию удовлетворяет бамана?

В Мали, где бамана является наиболее значимым языком (не случайно его здесь иногда называют «малийским языком»), за время после получения независимости были предприняты ощутимые усилия для того, чтобы повысить его статус (и, в несколько меньшей степени, статус других местных языков). Ещё в период правления Мусы Траоре (1968-1991) начала издаваться ежемесячная газета Kibaru, постепенно увеличивавшая свой объём и тираж; началось преподавание бамана в первых экспериментальных школах. После установления демократического режима в стране движение в этом направлении ускорилось. В течение последнего десятилетия XX века преподавание на бамана и на других «национальных языках», по методу «конвергентной педагогики», ввели чуть ли не в половине малийских школ. Несмотря на то, что трудности сколаризации на африканских языках в Мали

[1] *Выдрин В.Ф., Томчина С.И.* Манден-русский словарь. Т. 1. СПб.: Издательство Дмитрий Буланин, 1999.

огромны[1], прогресс в деле превращения бамана в письменный язык довольно ощутим. Были созданы полноценные школьные учебники (в 1999 году уже появился учебник для 6 класса общеобразовательной школы); при участии французских и норвежских лингвистов создана грамматика для учителей. Выросло количество периодических изданий на бамана. Помимо многочисленных брошюр, изданных в рамках «функциональной алфабетизации» (по гигиене, о выращивании хлопка или арахиса, о рытье колодцев…), христианской и исламской литературы, появилась и художественная литература. Имеются довольно многочисленные издания фольклора – сказок, эпоса, пословиц, загадок. Даже если ознакомиться со списком публикаций на бамана на начало 1990-х годов, он оказывается довольно внушительным[2]; с тех пор число таких публикаций выросло многократно – несомненно, они составляют в совокупности несколько миллионов словоупотреблений.

Итак, практические предпосылки для начала работы по созданию электронного корпуса бамана имеются. Однако остаются нерешёнными технические вопросы. Остановлюсь лишь на некоторых из них, а именно – на проблемах, связанных с тоновой нотацией. Эти проблемы сводятся к следующим пунктам[3]:

---

[1] См. на эту тему, в частности, *Dumestre G.* La classe école : une proposition radicale pour l'éducation au Mali // Mande Studies 8, 2006. P. 111–144; *Skattum I.* The introduction of the national languages into theeducational system of Mali: Objectives and consequences of the reform // Mande Studies 8, 2006. P. 95–109.

[2] *Dumestre G.* Bibliographie des ouvrages parus en bambara // Mandenkan, 26, 1993. P. 67–90.

[3] Я не буду рассматривать специально вопрос о качестве большинства публикаций на бамана – к сожалению, нередко оно оказывается ниже всякой критики; количество всевозможных ошибок в изданиях зачастую превышает все мыслимые пределы.

1) Обозначать ли тоны?

2) Если обозначать, то указывать ли лексические тоны или тоны поверхностной реализации?

1. Чтобы разговор об этом был содержательным, ему необходимо предпослать хотя бы краткий обзор особенностей тональной системы бамана[1].

1.1. Тональные классы

В бамана выделяются два уровневых тона (высокий и низкий). При этом подавляющее большинство слов распределяется по двум тональным классам: высокотоновый – все слоги несут высокий тон, и восходяще-тоновый – начальный сегмент слова несёт низкий тон, а конечный сегмент – высокий тон: *ǔ* ‘они’, *bùgɔ́* ‘бить’. Впрочем, тон конечного сегмента восходяще-тоновых слов меняется на низкий в некоторых контекстах, а именно, перед паузой и перед последующим высоким тоном:

1. Ù yé fàlì` bùgɔ̀
   3SG PFV осёл-ART бить
   ‘Они ударили осла’.

Поэтому их можно было бы рассматривать и как низкотоновые; в таком случае необходимо ввести правило повышения тона на конечном сегменте перед последующим низким тоном. Такой интерпретации препятствует факт наличия небольшой группы слов, у которых тон конечного сегмента всегда низкий – их-то и следует считать «истинно низкотоновыми» словами (в частности, в эту группу входит местоимение 3 лица единственного числа *à* и показатель инфинитива *kà*).

[1] По тональной системе бамана имеется довольно обильная литература. Здесь будут изложены лишь основные черты и правила функционирования тональной системы стандартного бамана – в той мере, в какой это затрагивает вопрос глоссирования текстов.

Помимо двух упомянутых больших тональных классов (а также маленького класса «истинно низкотоновых» слов), до 10% существительных образуют несколько «малых», или «нерегулярных», тональных классов; почти все такие существительные насчитывают три или более слога: *bǐlákòró* ‘необрезанный мальчик’ (мальчик, не прошедший инициацию), *wánjàlàká* ‘мифическое большое животное’, и др. Нарушают бинарную модель распределения по тональным классам и приставочные глаголы, а также глаголы, образованные путём префиксации именных основ; оба типа глагольной деривации подразумевают тоновую автономию морфем: *lájɛ̌* ‘собирать вместе’, *kǔnbɛ̌n* ‘встречать (гостя)’, и т.п. – заметим, что соответствующие существительные оказываются в «регулярных» классах: *lájɛ́* ‘собрание’, *kùnbɛ́n* ‘неделя’. Тональную автономию сохраняют корневые глагольные морфемы при редупликации: *bòlí* ‘бежать’ → *bòlíbòlí* ‘бегать туда-сюда’.

Если рассматривать оппозицию двух основных тональных классов, то можно считать, что тоны в них противопоставляются лишь на первом сегменте, в то время как тоны конечных сегментов обусловлены контекстом.

1.2. По крайней мере, для слов двух базовых тональных классов верно, что сегментной базой тонемы является целое слово. Сильным аргументом в пользу этого тезиса является **правило тональной компактности**, которое распространяется в бамана на некоторые виды синтагм (в первую очередь, на определительную непроизводную и на детерминативную с нереферентным первым компонентом). В соответствии с этим правилом, непервый компонент такой синтагмы утрачивает свой лексический тон, а тональный контур первого компонента распространяется на всю синтагму.

Определительная непроизводная синтагма:

só ‘дом’ + *fĩn* ‘чёрный’ → *só fĩn* ‘чёрный дом’,

sɔ̌ ‘лошадь’ + *fĩn* ‘чёрный’ → *sɔ̀ fĩn* ‘чёрная лошадь’.

Детерминативная синтагма:
mùsó ‘женщина’ + *sɛ̌n* ‘нога’ → *mùsò sén* ‘женская нога’,
mùsó ‘женщина’ + *bóló* ‘рука’ → *mùsò bóló* ‘женская рука’.

1.3. Тоновая деклинация (down-drift) и тоновый артикль

В бамана действует правило тоновой деклинации, которое можно, несколько схематизируя, описать следующим образом: если во фразе наблюдается чередование высоких и низких тонов, то каждое понижение тона (переход от высокого к низкому тону) имеет большую амплитуду, чем повышение (переход от низкого к высокому). Вследствие этого общий интонационный контур фразы оказывается нисходящим, а высокий тон в конце фразы часто реализуется ниже, чем низкий тон в начале фразы.

В целом тональная деклинация – явление, характерное для очень многих языков мира (как тональных, так и нетональных); в ней проявляется автоматическая тенденция к ослаблению напряжения к концу фразы. Однако в бамана она оказалась фонологизованной в результате утраты сегментной составляющей референтного артикля. Исторически этот артикль, по-видимому, имел вид суффикса *–ò,* присоединявшегося к существительному или к последнему компоненту именной группы (такой вид этот артикль и сейчас имеет в идиомах северо-западной ветви группы манден и в некоторых идиомах на северо-западе Кот-д’Ивуара). В современном бамана (также как как в манинка, дьюла и большинстве других вариантов манден) сохранилась только тоновая составляющая артикля (в литературе принято обозначать это явление как «плавающий низкий тон»)[1]:

– во-первых, присоединяясь к восходяще-тоновому слову (синтагме), он препятствует понижению тона его конечного компонента, даже если это слово или синтагма оказывается перед паузой или перед высокотоновым словом:

[1] Здесь и далее плавающий низкий тон обозначается значком грависа без сегментной основы.

2. A̍ yé mùsó` wéelé. (а не:
A̍ yé mùsò` wéelé.)
1SG PFV женщина-ART звать
‘Он позвал женщину’;

– во-вторых, высокий тон после тонового артикля реализуется ниже, чем предшествующий ему высокий тон – как если бы между двумя высокотоновыми сегментами стояло низкотоновое слово. Ср.:

3a. A̍ má só` yé.
3SG PFV.NEG дом-ART видеть
‘Он не увидел дом’.

3b. A̍ má só yé.
3SG PFV.NEG дом видеть
‘Он не увидел (никакого, ни одного) дома’.

В примере (3b) тон *yé* реализуется на том же уровне, что и тон *só* (а фактически – даже немного выше), тогда как в (3a) тон *yé* оказывается ощутимо ниже, чем тон *só.*

1.4. В бамана выделяется несколько типов именных групп, которые различаются между собой

а) по наличию/отсутствию соединительного элемента (а также по характеру этого элемента);

б) по тому, слова каких частей речи могут заполнять ту или иную позицию в синтагме;

в) по типу связи между компонентами синтагмы (при отсутствии соединительного элемента):

– компактная (лексический тон не-начального компонента устраняется, тон первого компонента распространяется на всю синтагму);

– связанная (компоненты сохраняют свои тоны, при этом их тоны влияют друг на друга по общим правилам тонотактики бамана – иначе говоря, в таких синтагмах первый член не имеет

артикля, так что контакт между тональными контурами обоих слов оказывается непосредственным);

– свободная (тоны компонентов синтагмы не влияют друг на друга – в большинстве случаев по той причине, что первый из компонентов имеет тональный артикль[1].

* * *

Теперь попробуем найти ответы на вопросы, сформулированные в разделе 0.

2. Обозначать ли тоны?

В правилах нынешней орфографии бамана, принятой в Мали, обозначение тонов признаётся допустимым, но не обязательным (при этом правила тоновой нотации не оговариваются). Фактически же тоны обозначаются только в словарях и в научных лингвистических публикациях (и то не во всех), в то время как основной объём публикаций не тонирован. Это, фактически, и оказывается главным аргументом против указания тонов в будущем электронном корпусе бамана.

Аргументы в пользу противоположного решения звучат значительно более весомо.

Во-первых, необозначение тонов резко увеличит графическую омонимию – количество минимальных тональных пар в бамана идёт на многие сотни. В частности, невозможно будет различить на письме местоимения 3 лица единственного числа *à* и 2 лица множественного числа *á;* показатель оптатива *ká* и показатель инфинитива *kà,* глаголы *fǎrá* ‘отделять’ и *fàrá* ‘присоеди-

[1] Подробное описание типов именных групп делает Дюместр: *Dumestre G.* Le bambara du Mali… P. 149–186. Следуя его интерпретации и терминологии насколько это возможно, я принципиально иначе рассматриваю выделение частей речи, благодаря чему существенно сокращается инвентарь типов ИГ в бамана.

нять, прибавлять', существительные *bɔ̀gɔ́* 'строительная глина' и *bɔ́gɔ́* 'гончарная глина', и многие другие.

Во-вторых, отсутствие тоновой маркировки сделает невозможным обозначать наличие или отсутствие артикля. Конечно, во многих случаях его употребление оказывается контекстно-заданным и, соответственно, предсказуемым. Но имеются и другие контексты (подобные 3a и 3b), в которых артикль действительно различает коммуникативные статусы слов.

В-третьих, некоторые типы именных групп различаются только тонами. Необозначение тонов приведёт к невозможности выразить на письме различные синтаксические значения[1].

Таким образом, если речь идёт о полноценном электронном корпусе, имеющем хоть какую-то научную и практическую ценность, необходимость обозначения тонов вряд ли может быть оспорена всерьёз.

3. Указывать ли лексические тоны или тоны поверхностной реализации?

3.1. Как это было показано в разделе 2, в словах двух основных тональных классов релевантным оказывается, фактически, лишь **тон начального компонента**, в то время как финальный тон порождается автоматически, в зависимости от левого и правого контекстов. Таким образом, обозначение только начального (лексического) тона слова позволяет сделать тоновую нотацию значительно более экономной (при этом утраты информации фактически не происходит). Это выражается не только в уменьшении количества диакритических знаков в тонированном тексте бамана, но и (что более существенно) в радикальном сокращении числа графических вариантов в глоссарии, который является основой для парсинга при глоссировании. В частности, если

[1] Фактически этот аргумент близок к предыдущему, о чём пойдёт речь ниже.

требовать обозначения тона конечного элемента, то для всех слов восходяще-тонового класса потребуется внесение в глоссарий двух графических вариантов, например: *bàlá* и *bàlà* ‘дикобраз’, *sĕn* и *sèn* ‘нога’. Если же обозначать только лексические тоны, то достаточно дать по одному варианту, соответственно *bàla* и *sèn.*

3.2. Вторая проблема, которая снимается при таком решении – **диалектное варьирование тонов**. Если распределение слов по тональным классам в бамана весьма устойчиво, то поверхностная реализация тонов может меняться весьма существенно, при этом некоторые правила поверхностной реализации тонов в диалектах могут быть факультативны. Если мы решаем обозначать тоны поверхностной реализации, мы всё равно вынуждены определить, какие из правил следует принимать в расчёт, а какие – нет; полный же учёт всех таких правил при создании электронного корпуса языка бамана – задача в принципе невыполнимая (по крайней мере, при нынешнем состоянии баманской диалектологии).

3.3. В то же время следует особо оговорить ситуацию с **«нерегулярными» тональными классами** существительных и с производными глаголами. В нынешней практике при публикации текстов на бамана с указанием лексических тонов, как правило, принадлежность слова к нерегулярному классу никак не отмечается – указывается лишь тон его первого компонента (в частности, это характерно практически для всех публикаций Жерара Дюместра последних лет). Такое решение вряд ли можно признать удовлетворительным – в отличие от тона конечного элемента любого слова из «регулярных» классов, здесь не-первые тоны не являются предсказуемыми, и они должны быть указаны в текстах. По-видимому, исключение можно сделать для крайне немногочисленных «истинно-низкотоновых» слов, таких как *à* (местоимение 3 лица единственного числа) или *kà* (показатель инфинитива): при крайней немногочисленности и высокой час-

тотности таких слов в текстах, особенности их тонального поведения можно особо оговорить, не вводя при этом специальной маркировки, отличающей их от восходящетоновых слов.

3.4. Очевидным образом, обязательным является обозначение **тонового артикля** – как отмечалось в разделе 1.3., его наличие оказывается предсказуемым далеко не всегда, так что игнорирование артикля ведёт к утрате существенной информации, содержащейся в тексте. Главной проблемой здесь является то, что при обработке письменных нетонированных баманских текстов (а также в тех тонированных текстов, где артикль игнорируется) задача установления факта наличия/отсутствия артикля целиком ложится на плечи человека, осуществляющего разметку текста. В ситуации когда правила употребления артикля вряд ли можно считать установленными в должной степени, это требует от оператора и очень хорошего знания языка бамана (фактически на уровне носителя языка), и хорошей лингвистической подготовки (во всяком случае, знания основ тонологии языка бамана – что скорее нетипично для подавляющего большинства современных малийских лингвистов).

3.5. Наконец, одна из самых непростых проблем связана с **тонированием не-первых компонентов компактных синтагм**, т.е. таких словосочетаний, не-начальные элементы которых утрачивают свои лексические тоны (см. раздел 1.2). Этот вопрос формулируется так: указывать ли для не-начальных слов в таких синтагмах их лексические тоны (несмотря на то, что в реальности эти тоны нейтрализуются) или «тоны нейтрализации», т.е. те тоны, которые они реально несут? (При втором решении тоны на таких словах можно не указывать вообще, поскольку они являются предсказуемыми, будучи автоматически выводимыми из контекста).

Рассмотрим аргументы в пользу каждого из этих решений.

а) Обозначение лексических тонов не-первых компонентов компактных синтагм

– облегчает их идентификацию (поскольку при этом сохраняется графическое единство лексемы в разных контекстах);

– резко снижает количество графических вариантов слова, которые необходимо вносить в глоссарий, что существенно облегчает работу при разметке корпуса. Так, если мы решаем обозначать не лексические тоны, а «тоны нейтрализации», то мы обязаны включить в глоссарий такие варианты двух слов, составляющих в бамана минимальную пару, противопоставленную по тону:

bàlá ‘дикобраз’ – *bàlá, bàlà, bálá;*
bálá ‘балафон’ (африканский ксилофон) – *bálá, bàlà.*

Очевидно, что такой принцип делает глоссарий значительно более громоздким.

б) Обозначение «тонов нейтрализации»

– более адекватно отображает звучание устной речи;

– предположительно, этот принцип позволяет более адекватно отразить природу синтаксических отношений в пределах именной группы.

Первый из этих аргументов существенен, однако он, очевидно, менее весом, чем приведённые выше аргументы в пользу противоположного решения. Второй же аргумент заслуживает более подробного рассмотрения.

Фактически он может быть переформулирован следующим образом: не отражая на письме тоновую нейтрализацию, мы рискуем упустить разницу между различными типами именных групп. Действительно ли это так?

Для того, чтобы отличить ИГ со «свободным» типом связи между компонентами от ИГ двух других типов, достаточно обозначения тонового артикля. Например:

mùso` kùn` [*mùsó` kǔn`*] ‘голова женщины’ («свободный» тип связи) ≠ *mùso kùn`* [*mùsò kún`*] ‘женская голова’ («компактный» тип связи).

wùlu` bòlilen* [*wùlú` bòlílèn] ‘когда/поскольку собака убежала…’ ≠ *wùlu bòlilen`* [*wùlú bòlílén`*] ‘сбежавшая собака’.

Опасность неразличения могла бы возникнуть только между ИГ с «компактным» и ИГ со «связанным» способами связи компонентов (см. 1.4). Однако в реальности такие ИГ различаются одновременно и классами слов, способными в такие синтагмы входить: «связанный» тип соединения с существительными характерен для числительных, причастий, большинства производных прилагательных (кроме прилагательных на *–man*); «компактный» – для прилагательных на *–man,* непроизводных прилагательных. Значит, эта опасность оказывается лишь гипотетической, в реальности же она практически никогда не реализуется.

Таким образом, аргументы в пользу обозначения лексических тонов оказываются существенно более значимыми.

4. Способ тонирования глоссированных баманских текстов – лишь одна из многих проблем (языковых, технических, организационных, финансовых), которые предстоит решить на пути к созданию работающей модели электронного корпуса этого языка. Однако, как представляется, именно она оказывается ключевой.