Корпусные словари языков манден

Valentin Vydrin

ЯЗЫКОЗНАНИЕ

В.Ф. Выдрин

КОРПУСНЫЕ СЛОВАРИ ЯЗЫКОВ МАНДЕН[1]

В течение последнего десятилетия были созданы миллионные корпуса текстов маninка и бамана. Параллельно с этим ведется работа по созданию и совершенствованию словарей этих языков. Складываются условия для вывода лексикографической работы на принципиально новый уровень, который имел бы целью создание «корпусных словарей» (corpus-driven dictionaries). Это позволило бы гораздо детальнее и достовернее, по сравнению с «традиционными» словарями, представлять полисемию лексем, их сочетаемостные свойства, стилистические и региональнодиалектные характеристики каждого значения. В статье дается пробная разработка структуры словарной статьи глагола *bánban* на основании Справочного корпуса бамана.

Ключевые слова: язык бамана, язык маninка, манден, корпусная лексикография, полисемия.

Примерно десять лет назад была осторожно высказана мысль о необходимости создания больших аннотированных корпусов текстов для языков манден [Выдрин 2008]. Сейчас можно сказать, что эта задача уже решена, по крайней мере в первом приближении: объем Справочного корпуса бамана на начало лета 2017 г. превысил 4,1 млн слов, при этом подкорпус со снятой омонимией перевалил за 900 тыс. слов. Справочный корпус маninка составил более 3,5 млн слов (3,1 млн в подкорпусе нко и почти 0,4 млн — в латинском подкорпусе).

Эти корпуса аннотированы при помощи словарей Bamadaba (баманский корпус) и Malidaba (маninканский корпус). Эти сло-

[1] This work is related to the research strand 3 «Typology and dynamics of linguistic sysems» of the Labex EFL (financed by the ANR/CGI).

вари в настоящий момент очень различаются по степени проработанности: Bamadaba представляет собой полноценный бамана-французский и французско-баманский словарь, находящийся в открытом доступе на сайте баманского корпуса (http://cormand.huma-num.fr/bamadaba.html), он активно используется студентами и всеми теми, кто интересуется языком бамана, постоянно пополняется и уточняется в ходе работы по снятию омонимии. Словарь Malidaba находится на значительно менее продвинутой стадии: он был скомпилирован полуавтоматически из различных лексикографических источников и сейчас находится на этапе предварительной «чистки», которая состоит в устранении дублирующих словарных статей, подборе трехъязычных глосс, и, менее последовательно, введения определенного минимума грамматической информации и первичной проработки полисемии (подробнее см.: [Vydrin, Rovenchak, Maslinsky 2016]). Первичная проработка словаря Malidaba к началу лета 2017 г. была проведена примерно на 65 %, после завершения первого этапа чистки планируется помещение Malidaba на сайте Справочного корпуса манинка (http://cormand.huma-num.fr/cormani/).

Создание этих словарей само по себе — значительный шаг в развитии лексикографии языков манден. Bamadaba, хотя и отстает по объему от словаря Жерара Дюместра [Dumestre 2011][2], существенно превосходит его по уровню структурированности, а также обладает преимуществами онлайн-словаря: постоянно совершенствуется (в том числе пополняется новыми пластами лексики, хорошо представленной в текстах, но остававшейся в силу разных причин вне «традиционных» словарей), имеет прямой доступ к корпусу (каждая лексема имеет ссылку, по которой мгновенно создается полный конкорданс из подкорпуса со снятой омонимией), дает возможность поиска как в бамана-французском, так и во французско-баманском режиме. Malidaba, хотя и значительно

[2] Словарь Ж. Дюместра насчитывает 15 850 словарных статей, Bamadaba на начало сентября 2017 г. — почти 12 500. Впрочем, эти цифры нужно принимать с оговорками, поскольку принципы выделения лексем в этих словарях различаются в некоторых важных моментах.

уступает баманскому словарю и в количественном[3], и в качественном отношении, станет первым манинка-французским словарем, отвечающим хотя бы минимальным лексикографическим требованиям (обозначение тонов, разработка полисемии, наличие грамматических помет и т.д.).

Таким образом, создание и совершенствование миллионных корпусов манден уже оказало ощутимое влияние на лексикографию этих языков. Тем не менее, не будет преувеличением сказать, что лексикографический потенциал корпусов пока задействован лишь незначительно. Настало время всерьез задуматься о создании корпусных словарей (corpus-driven dictionaries) для языков манден с учетом накопленного опыта корпусной лексикографии для других языков. Такие словари, по сравнению с «традиционными», позволяют представить лексику языка гораздо более объемно, максимально приближая представление лексемы в словаре к ее реальному употреблению в языке.

Поскольку литература о корпусной лексикографии достаточно богата (см., в частности: [Kilgariff 1997; Kilgarriff, Rundell, Dhonnchadha 2006]), вместо изложения известных методологических подходов я попробую показать на примере одной статьи (глагол *bánban*), что дает корпусной подход к решению лексикографических задач для бамана.

В словаре Дюместра [Dumestre 2011] (далее DUM) указано, что эта лексема имеет также фонетический вариант *báman*, и даются (простым списком) следующие переводные эквиваленты: s'évertuer, persévérer, faire de son mieux, prendre au sérieux; prendre appui; enfoncer, fixer solidement BA[4]; centrer, concentrer; tendre, déployer, hisser; effort tendu KT, attention soutenue, sérieux (стараться, упорно делать, делать изо всех сил, относиться серьезно; опираться; втыкать, прочно закреплять BA; сосредотачивать, концен-

[3] По предварительной оценке, после завершения первого этапа чистки Malidaba будет содержать примерно 6300 лексем.

[4] Двухбуквенные индексы у Дюместра отсылают к источникам, список которых дается в конце его словаря.

трировать; натягивать, развертывать, поднимать (парус, флаг); напряженное усилие КТ, напряженное внимание, серьезность)[5].

Далее в словаре приводятся 16 примеров, иллюстрирующих употребление *bánban.*

В Манден-русском словаре [Выдрин, Томчина 1999] (далее ВТО) даны три формы: *bánban* (общеупотребительная), *bánba* (только бамана), *báman* (неуточненный диалект бамана). Значения глагола представлены с иерархизированной рубрикацией, даны пометы, отражающие актантную структуру глагола в каждом значении (переходный/ непереходный/ рефлексивный глагол; при наличии сильной валентности указывается послелог, вводящий косвенный член), в случаях, когда какие-то значения характерны не для всего ареала, вводятся соответствующие диалектные пометы. Таким образом, полисемия глагола *bánban* здесь выглядит следующим образом (иллюстративные примеры и идиоматические выражения опущены):

1. (m) *пг* натягивать, туго наматывать

2.1. *пг* прикреплять, укреплять, вбивать; прочно ставить, устанавливать

рг прочно встать

2.2. (m) *пг* делать прочным

3. *пг* упираться чем-л. (в — *lá*), надавливать чем-л.

рг опираться, упираться (в — *lá*)

4. *рг* усердствовать (в — *lá*), очень стараться (делать что — inf.)

5. (m) *пг* поддерживать, помогать

6. (bSegu) *рг* выпутаться, легко отделаться

Сравнение двух словарей[6] показывает, что они дополняют друг друга:

---

[5] Последние три значения именные. Дюместр рассматривает конверсию «глагол → существительное» как полисемию. Далее именные значения не анализируются.

[6] В словаре Шарля Байоля [Bailleul 2007] структура полисемии этого глагола представлена более сжато, она полностью перекрывается словарями DUM и ВТО и отдельно рассматриваться не будет.

— в DUM есть значения ‘сосредотачивать, концентрировать’, ‘поднимать (флаг, парус)’, ‘относиться серьезно’, отсутствующие в BTO;

— BTO, но не DUM, дает значения ‘*pг* прочно встать’, ‘упираться чем-л., надавливать’, ‘поддерживать, помогать’, ‘*pг* выпутаться, легко отделаться’ (диалектное значение); для манинка, но не для бамана, указаны также значения ‘*nг* делать прочным’ (которое, впрочем, можно объединить со значением ‘укреплять’, отделив последнее от 2.1. ‘прикреплять’ и т.д.).

Кроме того, BTO указывает значение ‘*nг* натягивать, туго наматывать’ с пометой «манинка». DUM подтверждает значение ‘натягивать’, но не ‘туго наматывать’ (которое, по-видимому, следует рассматривать не вместе с ‘натягивать’, а как отдельное значение или даже как вариант значения 2.1).

Можно также предложить «приблизить» манинканское значение (5) ‘поддерживать, помогать’ к семантически смежному значению 2.2. ‘делать прочным, укреплять’.

По результатам этого сравнения полисемия глагола может быть представлена следующим образом:

1.1 *nг* натягивать, развертывать

1.2 *nг* поднимать (парус, флаг)

2.1. *nг* прикреплять, туго наматывать; вбивать; прочно ставить, устанавливать

pг прочно вставать

2.2. (m) *nг* делать прочным, укреплять

2.3. (m) *nг* поддерживать, помогать

3. *nг* упираться чем-л. (в — *lá*), надавливать чем-л.

pг опираться, упираться (в — *lá*)

4.1 *pг* усердствовать (в — *lá*), очень стараться (делать что — inf.)

4.2 *pг* относиться серьезно (к — *lá*)

4.3 *nг* концентрировать, сосредотачивать (на — *kàn*)

5. (bSegu) *pг* выпутаться, легко отделаться.

В мае 2013 года, когда Справочный корпус бамана насчитывал около 1,5 млн слов, я провел корпусную проверку полисемии это-

го глагола. На тот момент в корпусе обнаружилось 123 употребления формы *banban*, из которых 12 (именные употребления, омонимы) были отсеяны. Таким образом, к интересующему нас глаголу на тот момент относились 111 употреблений (в это число вошли и регулярные дериваты, значения которых автоматически выводимы из глагольных, и случаи контекстной номинализации глагола в составе именных композитов). Почти половину корпуса (точнее, около 700 тыс. слов из 1,5 млн) на тот момент составляли тексты Ветхого Завета. Первым интересным результатом стало то, что из 111 употреблений *bánban,* 76 (т.е. 68,5 %) происходили из Ветхого Завета, иначе говоря, в этом источнике глагол *bánban* оказался вдвое более частотным, чем в остальных текстах Корпуса. Интересно также, что фонетический вариант *baman,* отмеченный в словарях BTO и DUM, не встретился ни разу.

В сентябре 2017 г. я снова проанализировал употребление *bánban* по действующей версии Справочного корпуса бамана, которая превосходила по объему (4,1 млн слов) версию 2013 г. в 2,73 раза; соответственно, доля текстов Ветхого Завета в Корпусе снизилась до 17 %.

Общее число употреблений формы *banban* с учетом дериват и композитов составило 253. После отсева омонимов, нерегулярных дериват (каузатив *lábanban*) и именных употреблений это число сократилось до 226. Таким образом, прежняя тенденция сохранилась — частотность *bánban* в Ветхом Завете и сейчас в два раза выше, чем в остальной части корпуса.

Несмотря на увеличение выборки, вариант *baman* так и не появился (встретившаяся пять раз форма *bamanna* оказалась во всех случаях не перфективной интранзитивной формой интересующего нас глагола, а неправильным написанием слова *bámàna* ‘бамана’); очевидно, его следует вынести за пределы современного письменного узуса бамана.

Ниже будет проанализирована полисемия глагола *bánban* на основании корпусных данных. Последовательность значений в целом соответствует той, что приведена выше; новые (под)значения вводятся (в соответствии с логикой, принятой в BTO) рядом с теми значениями, которые наиболее близки к ним семантически.

При каждом значении сразу после валентностной пометы (*пг, нг, рг*) указано в скобках число употреблений глагола *bánban* в данном (под)значении.

При количественной оценке веса каждого значения следует иметь в виду, что учитывались не только собственно глагольные употребления, но и регулярные дериваты, а также композиты, в которых глагольное значение просматривается достаточно отчетливо; для некоторых значений доля таких дериват и композитов оказалась весьма высокой. Нужно отметить, что в значениях *bánban,* для которых характерна лабильность, у дериватов и композитов противопоставление по переходности/ непереходности/ рефлексивности (укреплять — укрепляться, прочно ставить — прочно вставать) нейтрализуется, что создает трудность для статистического подсчета. Другая трудность — вопрос об отнесении пассивных употреблений к переходному или непереходному подзначению (этот вопрос я решил в пользу переходных значений, т.е. ‘быть установленным’ считается вместе с ‘прочно устанавливать’, а не ‘прочно вставать’).

Другой важный момент — это синкретизм значений. Зачастую оказывается не так просто определить, в каком именно из смежных значений употреблен глагол в том или ином предложении, например 2.1 ‘прочно ставить’ или 2.2 ‘делать прочным, укреплять’? 4.1 ‘усердствовать в чем-л.’ или 4.2 ‘относиться серьезно к’?

Таким образом, статистику распределения употреблений глагола по значениям следует рассматривать как приблизительную.

1.1 *пг* (4) натягивать, растягивать

Это значение в BTO помечено как «манинканское»; впрочем, значение ‘tendre’ упоминается и в DUM. Из четырех примеров один — из газеты *Jɛkabaara*, в абстрактном контексте описания отношений между людьми; три — из художественной литературы. Таким образом, «манинканская» помета может быть убрана, однако эти значения в бамана оказываются нечастыми.

1.2 *пг* (0) поднимать (парус, флаг).

Встает вопрос о том, нужно ли выделять это значение, или считать его вариантом значения 1.1.

2.1. *пг* (22) прикреплять, туго наматывать, вбивать; прочно ставить, устанавливать;

нг (3) прочно вставать;

рг (0) прочно вставать.

Примеров на рефлексивный вариант этого значения в корпусе не найдено, зато обнаружились (пусть и не очень многочисленные) непереходные употребления, не предусмотренные в имеющихся словарях. Впрочем, из трех таких употреблений бесспорным является только одно, еще в двух случаях мы имеем форму результативного причастия *bánbannen,* употребленную в идентичных контекстах (1). Таким образом, непереходный вариант значения также можно считать маргинальным.

(1)	*Dùgukolo`*	*bánban-nen*		*bɛ́*	*kɔ́nùman*	*ànî*
	земля\ART	натягивать-PTCP.RES		быть	хорошо	и
à	*tɛ*	*sé*	*kà*	*lámàga.*		
3SG	IPFV.NEG	мочь	INF	двигать		

‘Земля стоит/установлена прочно, ее нельзя пошевелить’.

Интересно также, что из 22 переходных употреблений *bánban* в этом значении 19 приходится на Ветхий Завет (почти всегда — в описании создания мира и его компонентов Богом), два — на газету *Jɛkabaara*, один — на газету *Kibaru*. Получается, что довольно высокая цифра для данного значения (около 10 % от всех корпусных употреблений глагола *bánban*) — следствие описанного Адамом Килгарифом «эффекта морской улитки»[7], тогда как за пределами Ветхого Завета это значение оказывается редким.

2.2. *пг* (55) делать прочным, укреплять;

нг (31) укрепляться, закрепляться, укореняться, становиться устойчивым;

рг (3) укрепляться, закрепляться.

[7] «The whelks problem» [Kilgarriff 1997: 138–139]. Имеется в виду ситуация, когда в выборке оказывается текст, специально посвященный какому-то в целом редко упоминаемому объекту или явлению (например, морской улитке, *whelk*), в результате чего суммарная частотность этого слова в выборке оказывается завышенной.

Это значение (‘делать прочным’) в ВТО помечено как «манинканское», однако оказалось, что в баманском корпусе оно (особенно его абстрактный вариант ‘укреплять’) — самое частотное. Обнаружилось немало употреблений и непереходного варианта этого значения, не отмеченного в ВТО[8], а также рефлексивный вариант, более редкий. Распределение по источникам для этого значения оказывается относительно равномерным (точнее, перевес в пользу Ветхого Завета наблюдается и здесь, но он значительно меньший, чем для значения 2.1).

Таблица 1. **Распределение употребления 2.2 по источникам**

	Ветхий Завет	Jɛkabaara	Kibaru	другие
пг	25	21	4	5
нг	7	24	0	0
рг	2	0	1	0

2.3 *пг* (3) поддерживать кого-л., помогать кому-л.

Это значение в ВТО также отнесено к манинка, однако оно отмечено также в баманском Ветхом Завете, из которого и происходят все три обнаруженных употребления.

3.1 *пг* (0) упираться чем-л. (в — *lá*), надавливать чем-л.;
рг (0) упираться (в — *lá*).

Это значение, упомянутое в ВТО как вариант значения ‘опираться’ (однако в действительности достаточно дистантное от него), в корпусе не обнаружено; по-видимому, имеет смысл исключить его из словаря.

3.2 *рг* (22) опираться (на, о — *kàn*);
нг (12) опираться (на, о — *kàn*).

Отделение этого значения от 3.1 (с которым оно объединено в ВТО) подкрепляется и тем, что его косвенное дополнение требует послелога *kàn,* а не *lá* (вопреки тому, что указано в ВТО; впро-

[8] Отметим, впрочем, что из 31 употребления, отнесенного при подсчете к непереходным, лишь 20 приходятся на финитные, остальные 11 — причастия и отглагольные имена, которые соотнесены с непереходным вариантом этого значения лишь предположительно.

чем, вполне возможно, что тут речь не о разном управлении при разных значениях, а об ошибке в BTO или различии между манинка и бамана по этому параметру).

Практически во всех отмеченных случаях употребления (как непереходных, так и рефлексивных) речь идет об абстрактном варианте значения (‘опираться на кого-то в своих действиях’, ‘опираться на правила’ и т.п.), физическое действие имеется в виду лишь в одном случае.

Различие между рефлексивным и непереходным употреблениями обусловлено, за редкими исключениями, типом подлежащего: при одушевленном подлежащем глагол употребляется рефлексивно, при неодушевленном — непереходно, что соответствует общему правилу [Vydrine, Coulibaly 1995: 71]. Нужно оговориться, что приведенная здесь статистика условна, поскольку более половины употреблений *bánban* в этом значении — в составе композитов, в таких случаях рефлексивный или непереходный характер исходной глагольной конструкции может быть определен только по типу смыслового подлежащего.

Что касается распределения этих употреблений по источникам, то здесь отмечено лишь одно рефлексивное употребление в *Kibaru* и четыре рефлексивных употребления в Ветхом Завете, все остальные происходят из *Jɛkabaara*[9], причем почти все они принадлежат перу одного автора — Тумани Ялама Сидибе[10], главного редактора и основного автора этой газеты. Таким образом, частое употребление *bánban* в этом значении можно считать особенностью авторского стиля Т.Я. Сидибе.

4.1 *рг* (20) усердствовать (в — *fɛ̀, kàn*), стараться (делать что — *fɛ̀, kàn,* Inf., Sbjv).

Вопреки BTO, косвенный член (объект приложения стараний) вводится послелогами *fɛ̀* (при обозначении сферы деятельности) или *kàn* (если обозначается цель деятельности), а не *lá*.

---

[9] При том что объемы текстов из обеих газет, *Jɛkabaara* и *Kibaru*, в корпусе примерно одного порядка, см.: http://cormand.huma-num.fr/sc_non_desambig.html, http://cormand.huma-num.fr/sc_desambig.html.

[10] Т.Я. Сидибе является выходцем из г. Кита, т.е. из зоны распространения центрально-западного варианта языка манинка.

(2) *Án ka kán k’ án bánban*
1PL QUAL.AFF равный INF 1PL натягивать
kɔ́ɔri-sɛnɛ ìn fɛ̀…
хлопок-возделывать DEF с
‘Мы должны усердно растить хлопок...’ [jekabaara116_03sidibe-kita_makoci].

(3) *Án ka dá án yèrɛ̂ lá; k’ án*
1PL SBJV ложиться 1PL сам в INF 1PL
bánban án ka ké-ta-w ké-cogo
натягивать 1PL POSS делать-PTCP.POT-PL делать-способ
ɲuman` kàn.
хороший\ART на
‘Мы должны верить себе и стараться как следует выполнять намеченное’ [jekabaara113_01sidibe-kalo_laadilikan].

Достаточно часто глагол в этом значении употребляется и без косвенного члена.

Что касается распределения употреблений *bánban* в этом значении, то и здесь газета *Jɛkabaara* лидирует с большим отрывом (13 употреблений из 20, в том числе девять примеров принадлежат перу Т.Я. Сидибе), остальные случаи распределены равномерно.

4.2 *рг* (14) относиться серьезно (к — *kàn, mà*).

Это значение отмечено в DUM и отсутствует в BTO; корпусная проверка показала, что оно может быть отнесено к семантическому ядру глагола *bánban.* Чаще всего косвенный член вводится послелогом *kàn,* встретился лишь один случай, когда в этой функции употреблен послелог *mà* (возможно, речь идет о каком-то особом нюансе значения, но для проверки этого предположения пока недостаточно данных); еще один пример с послелогом *mà* приводится в DUM.

По распределению употреблений здесь также абсолютно доминирует *Jɛkabaara* (11 употреблений).

4.3 *пг* (11) концентрировать, сосредотачивать, нацеливать (на — *kàn*);

рг (6) сосредотачиваться, концентрироваться (на — *kàn*).