**DOI: 10.24411/2308-8079-2018-00011**

**УДК 81'33+004.6**

# ЗАГОЛОВОК КАК ЛИНГВИСТИЧЕСКИЙ ОБЪЕКТ В ПОЛНОТЕКСТОВОЙ БАЗЕ ДАННЫХ РЕГИОНАЛЬНЫХ БЕЛОРУССКИХ СМИ

**Станкевич А.Ю.**

В статье дана краткая характеристика полнотекстовой базы данных региональных белорусских СМИ (РБМ). Описана структура разметки заголовков по языку. Автором также определены инструменты лемматизации заголовков. Приведены договоренности, используемые при выверке русскоязычных лемм. Описаны структура и направления использования лингвистической базы данных «Заголовки РБМ».

**Ключевые слова:** лингвистическая база данных, заголовок, региональные газеты, двуязычные газеты, лемма.

# HEADLINE AS A LINGUISTIC OBJECT IN A FULL-TEXT DATABASE OF REGIONAL BELARUSIAN MASS MEDIA

**Stankevich A.Y.**

The article gives a brief description of the full-text database of regional Belarusian mass-media (RBM) and describes the structure of a language annotation of headlines. The author also defines tools for lemmatization of headlines and describes agreements for Russian lemmas correction as well as the structure and uses of the linguistic database “Headlines of RBM”.

**Keywords:** linguistic database, headline, regional newspapers, bilingual newspapers, lemma.

*Статья подготовлена в рамках проекта Государственной программы научных исследований «Экономика и гуманитарное развитие белорусского общества» на 2016-2020 гг. (договор № А70-16 от 04.01.2016).*

**Общая характеристика ресурса**

Полнотекстовая база данных региональных белорусских СМИ (далее РБМ) разрабатывается на основе находящихся в открытом доступе электронных версий региональных печатных изданий. Объем текущей версии ресурса – 4,5 млн словоупотреблений (с/у); ресурс включает метаразмеченные тексты по пяти областям Беларуси: Гродненской (2,29 млн с/у по текстам газет «Астравецкая праўда», «Бераставіцкая газета», «Вечерний Гродно», «Воранаўская газета», «Іўеўскі край», «Перспектива», «Праца», «Свіслацкая газета»), Гомельской (0,598 млн с/у по текстам газет «Дняпровец», «Светлае жыццё», «Светлагорскія навіны», «Хойніцкія навіны»), Могилевской (0,573 млн с/у по текстам газет «Прыдняпроўская ніва», «Родная ніва»), Витебской (0,556 млн с/у по текстам газет «Жыццё Прыдзвіння», «Герой працы»), Брестской (0,531 млн с/у по текстам газет «Заря над Бугом», «Навіны Камянеччыны»).

Преобладание доли Гродненской области объясняется включением в РБМ языкового материала иллюстративного лингвистического корпуса СМИ Гродненщины [5] (последний доступен в двух режимах: 1) в автономной версии [2]; 2) через интерфейс корпуса региональной и зарубежной прессы НКРЯ [3] при определении подкорпуса через метапараметр «Регион: Гродненская область».

В финальной версии ресурс будет сбалансирован по объему, в том числе и за счет увеличения объема ресурса по всем областям, кроме Гродненской. Для пополнения ресурса определены потенциальные источники: 1) газеты районов, имеющих выход на границы государств-соседей Беларуси (источники 1-й очереди); 2) газеты «внутренних» районов (источники 2-й очереди). Приведем пример: Витебская область граничит с Литвой, Латвией, Россией; тогда 1-ю очередь пополнения РБМ определяют газеты районов, имеющих выход на эти границы, а 2-ю очередь пополнения – газеты прочих районов Витебской области.

Временной охват ресурса: 2012 год.

**Представление заголовков в РБМ**

*Разметка заголовков по языку и разметка вкраплений*

Регулярным и формально идентифицируемым выражением взаимодействия русского и белорусского языков в РБМ является наличие вкраплений типа «белорусское в русском», которое положено в основу выделения в РБМ трех массивов текстов: русскоязычного, белорусскоязычного, смешанного.

В заголовках (как и в текстах) вкрапления типа «белорусское в русском» размечаются в автоматизированном режиме специально разработанным python-модулем «Вкрапления.BE» с опорой на лингвистическую базу данных графических маркеров, идентифицирующих вкрапление. Графические маркеры выявлены лингвостатистически на тестовых массивах данных (тестовый массив для русского языка определен на основе расширенного словаря А.А. Зализняка от М. Хагена [6], тестовый массив для белорусского языка определен на основе лексико-грамматической базы Белорусского N-корпуса [4]). Графический маркер представляет собой кортеж вида «текстовый фрагмент; коэффициент различительной силы» (например: «дзё; 0,9», «рыф; 0,9», «жы; 1»). Вхождение графического маркера в текст позволяет идентифицировать белорусскоязычное вкрапление. Принципы выявления графических маркеров описаны в совместной с И.И. Бубнович статье «Графічныя маркёры для аўтаматызаванай ідэнтыфікацыі ўваходжання беларускамоўных фрагментаў у змешаны беларуска-рускі тэкст», которая будет доступна в электронной научной библиотеке ГрГУ (www.lib.grsu.by).

Язык заголовков может не совпадать с языком текста; поэтому в РБМ не только каждый текст, но и каждый заголовок получает метку языка: параметр **hlang** («язык заголовка») принимает значение **be** («белорусский»), **ru** («русский») или **mfr** («смешанный»). В заголовках со смешанным языком размечаются вкрапления.

*Лемматизированное представление заголовков*

Каждый заголовок в РБМ имеет параллельное представление в виде набора лемм.

Лемматизация русскоязычных заголовков проводилась парсером MyStem [7].

Лемматизация белорусскоязычных заголовков проводилась декларативно специальным python-модулем «Лемматизатор.BE»; источником данных для лемматизатора стала лексико-грамматическая база Белорусского N-корпуса [4], при этом вручную были лемматизированы и пополнили масив данных для модуля «Лемматизатор.BE» около 700 словоформ, не вошедших в доступную нам редакцию лексико-грамматической базы (преимущественно имена собственные).

Лемматизация смешанных заголовков проводилась в два этапа: на первом этапе – парсером MyStem (по очевидной причине эвристика при лемматизации вкраплений давала ошибки); на втором этапе разбор вкраплений по MyStem автоматически «затирался» разбором вкраплений по python-модулю «Лемматизатор.BE» (позиции вкраплений были заранее проиндексированы python-модулем «Вкрапления.BE»).

В текущей версии РБМ коррекция разбора по леммам (снятие омонимии и вопросов разбора) проведена только для русскоязычных заголовков.

**Договоренности при снятии омонимии и вопросов разбора для русскоязычных заголовков**

При коррекции лемм заголовка из нескольких доступных режимов («Просмотр по заголовкам», «Просмотр по одной искомой словоформе / лемме», «Просмотр с сортировкой по леммам, подлежащим коррекции») разметчики выбирали последний режим.

В этом случае разметчик корректировал поле «Лемма», работая с таблицей вида:

| **Заголовок** | **Словоформа** | **Лемма** |
|---|---|---|
| Генерал армии посетил мемориальный комплекс в **аг. Копти Витебского** района | аг. | {аг??}. |
| <…> | <…> | <…> |
| Торжественный митинг, посвященный **67-ой** годовщине Великой Победы, собрал ветеранов **Витебского** района у братской могилы в **аг. Новка** | аг. | {аг??}. |
| **Нижне-Силезская** военная операция. Воспоминания **ее** участника **Терещенко** Виктора, жителя **агрогородка Копти** Витебского района | агрогородка | {агрогородок?} |
| <…> | <…> | <…> |
| Семья **Пивоварчик** из **агрогородка Бабиничи Витебского** района участвовала в конкурсе на лучшее молодежное подворье | агрогородка | {агрогородок?} |
| <…> | <…> | <…> |
| 22 декабря – День **энергетика** | энергетика | {энергетик \| энергетика} |

*Комментарий к таблице:*

Число повторов заголовка в таблице заданной структуры равно числу словоформ этого заголовка, леммы к которым требуют коррекции. В примере выше такие словоформы выделены полужирным в поле «Заголовок».

Работа в режиме «Просмотр с сортировкой по леммам, подлежащим коррекции» позволяет существенно увеличить скорость выверки за счет возможности редактирования больших групп однотипных разборов (в таблице выше это группы разборов для сокращения *аг.* (‘агрогородок’) и группа разборов для словоформы *агрогородка*).

Каждый случай выверки разметчик мог прокомментировать, заполнив отведенное для комментариев поле.

Ниже приведем основные договоренности разметчиков.

А. Прописная буква в записи лемм не используется (*«{беларусбанк}», {витебск}, {иванов}, {тэц}* и т.п.).

Б. Леммы инициалов и сокращений не расшифровываются, точки остаются за скобками (например: *{л}.* по фрагменту *сказал **Л.** Бэднарэк*; *{тыс}.* по фрагменту *25 **тыс.** тонн*).

В. Цифры как леммы не рассматриваются и в фигурные скобки не вносятся (например: *{дожинки}-2012*, *70-{летие}*).

Г. Леммы по фрагментам с дефисами пишутся по схеме **{лемма}-{лемма}** во всех случаях, кроме перечисленных ниже:

Г-1) дефис вносится внутрь скобок для лемм к зафиксированным в БТС [1] местоимениям и местоименным наречиям с *-то, -либо, -нибудь, -таки* (ср: *{кто-то}, {что-то}, {как-то},{все-таки}, {так-таки}*, но *{хороший}-{то}, {он}-{то}, {наш}-{то}, {поймать}-{таки}* и т.п.);

Г-2) дефис вносится внутрь скобок для лемм к наречиям по модели *по-...ски, по-...цки, по-...ему, по-...ому* (например: *{по-ивьевски}, {по-немецки}, {по-прежнему}, {по-новому}*);

Г-3) дефис вносится внутрь скобок для лемм к предлогам *из-за, из-под* *({из-за}, {из-под})*.

Такой путь обработки слов с дефисами выбираем для упрощения пользовательского поиска по леммам; любое слово с дефисом, не подпадающее под вышеперечисленное исключение, нужно будет искать как несколько лемм, разделенных дефисом: *{далеко}-{далеко}, {ё}-{моё}, {научный}-{методический}, {премьер}-{министр}, {сон}-{трава}, {тет}-{а}-{тет}, {on}-{лайн},{sos}-{родители}* и т.п.

Д. В записи леммы по словоформе с дефисом как средством экспрессивной пунктуации дефис должен быть удален (например: лемма *{долго}* по фрагменту ***до-о-о-о-лго** искать*).

Е. Леммы для сокращенных записей порядковых числительных пишутся в форме м. р., ед. ч., им. п. по правилам сокращений (например: *67-{й}* как по фрагменту *67-го*, так и по фрагменту *67-ю*).

По этой же договоренности пишутся леммы для сокращенных записей порядковых числительных в названиях периодов (например: *70-{й}* по фрагменту *музыка* ***70-х***).

Ж. Леммы имен, фамилий, отчеств пишутся с учетом пола референта (например: леммы *{мария}, {ивановна}, {иванова}* по фрагменту *поздравляем* ***Марию Ивановну Иванову***).

Леммы по фрагментам типа *в доме* ***Цыганковых****, династия учителей* ***Отливанчик-Ивановых*** записываются в форме м. р., ед. ч., им. п. *({цыганков}, {отливанчик}-{иванов}*).

З. Леммы топонимов пишутся «как в реестре» (например: леммы *{Боровое}, {Путришки}* по фрагментам *поселились в* ***Боровом****, поселились в* ***Путришках***).

И. Каждый компонент аналитических форм и фразеологизмов, многословных объединений возводится к начальной форме, границы многословных объединений не указываются.

Например: союз *чем ... тем* будет представлен леммами *{что}, {то}*; предлог *по причине* будет представлен леммами *{по}, {причина}*, превосходная степень *самому веселому ... из всех* будет представлена леммами *{самый}, {веселый}, {из}, {весь}*.

К. Леммой к компаративам является нулевая степень (наречия, слова категории состояния либо прилагательного); часть речи определяется по контексту (заголовку).

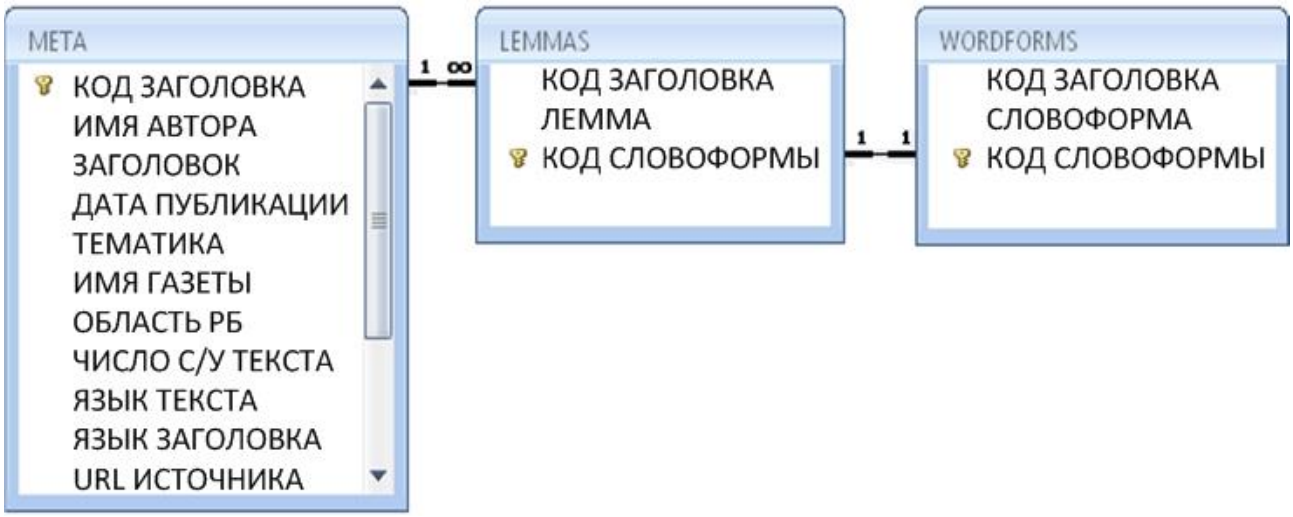
Л. Леммы для субстантиватов должны быть выверены по словарям, предпочтительно по БТС версии портала Грамота.ру [1] (например, слово *окружающие* зафиксировано в БТС, значит, лемма *{окружающие}* по фрагменту *дело рук* ***окружающих*** подтверждена).

Аналогичная выверка выполняется для адъективатов, адвербиатов.

**Массив заголовков в составе РБМ и как самостоятельный ресурс**

Массив заголовков РБМ включен в метатаблицу, каждая запись которой определяет текст как единицу хранения РБМ по основным метапараметрам

«Имя автора», «Заголовок», «Дата публикации», «Тематика», «Имя газеты», «Область РБ», «Число с/у текста», «Язык текста», «Язык заголовка», «URL источника» и служебным метапараметрам («Имя файла» и некоторым другим). Кроме того, массив заголовков используется как самостоятельный ресурс в составе базы данных «Заголовки РБМ» (см. схему данных на рисунке ниже):




Поля таблицы META соответствуют метапараметрам текста как единицы хранения РБМ; таблицы LEMMAS / WORDFORMS содержат соответственно леммы / словоформы, коды их взаимной привязки, коды привязки лемм / словоформ к заголовкам.

На рисунке ниже приведено окно формы этой базы данных, предназначенное для просмотра лемматизированных заголовков (в текущей версии такой просмотр реализован для русскоязычной части со снятой омонимией и снятыми вопросами разбора):

**Заголовок** Витебск отмечает свой 1038-ой день рождения

**Лемма**

| Лемма заголовка | Словоформа заголовка |
|---|---|
| {витебск} | Витебск |
| {отмечать} | отмечает |
| {свой} | свой |
| 1038-{й} | 1038-ой |
| {день} | день |
| {рождение} | рождения |
| * | |

Структура базы данных позволяет осуществлять поиск заголовков, соответствующих искомым метапараметрам и / или содержащих искомую лемму (леммы), словоформу (словоформы). Некоторые типичные, по нашему мнению, запросы описаны макросами, запускаемыми из формы «Типичные запросы». К примеру, для поиска русскоязычных заголовков, содержащих две леммы, введенные с клавиатуры пользователем, инициируется макрос на выполнение следующей серии запросов:

1) Запрос «L1»:

SELECT LEMMAS.[ЛЕММА ЗАГОЛОВКА], META.ЗАГОЛОВОК, LEMMAS.[КОД ЗАГОЛОВКА], META.[ЯЗЫК ЗАГОЛОВКА]

FROM META INNER JOIN LEMMAS ON META.Код = LEMMAS.[КОД ЗАГОЛОВКА]

WHERE (((LEMMAS.[ЛЕММА ЗАГОЛОВКА]) Like "*{" & [lemma_1] & "}*") AND ((META.[ЯЗЫК ЗАГОЛОВКА])="ru"));

2) запрос «L2», идентичный первому, но с именем параметра [lemma_2] вместо [lemma_1] ;

3) запрос «L1L2»:

SELECT L1.ЗАГОЛОВОК

FROM L1 INNER JOIN L2 ON L1.[КОД ЗАГОЛОВКА] = L2.[КОД ЗАГОЛОВКА];

База данных содержит описание около 13 тыс. заголовков (6,1 тыс. по Гродненской; 2,3 тыс. по Могилевской; 2,0 тыс. по Витебской; 1,4 тыс. по Гомельской; 1,2 тыс. по Брестской областям).

Атрибуты заголовка как лингвистического объекта, заданные в РБМ и в базе данных «Заголовки РБМ» позволяют делать выборки языковых данных, репрезентативных при исследовании заголовков двуязычных региональных СМИ Беларуси. Достаточно прозрачная структура базы данных «Заголовки РБМ» позволяет использовать ее сокращенную версию как учебную базу при изучении темы «Запросы» на занятиях по информационным технологиям для филологов.

**Список литературы:**

1. Большой толковый словарь русского языка / Сост. и гл. ред. С.А. Кузнецов [Электронный ресурс] // Грамота.ру [сайт]. 2014. URL: https://goo.gl/K9Q9sc (дата обращения: 17.09.2018).

2. Иллюстративный лингвистический корпус СМИ Гродненщины / Сост. Л.В. Рычкова [и др.] [Электронный ресурс] // Спец. массив науч. данных (32,3 Мб). Гродно: ГрГУ им. Я. Купалы, 2017. 1 Электрон. опт. диск (DVD-ROM).

3. Корпус региональной и зарубежной прессы [Электронный ресурс] // Национальный корпус русского языка [сайт]. 2018. URL: https://goo.gl/7LVrhq (дата обращения: 17.09.2018).

4. Кошчанка У. Лексіка-граматычная база для Беларускага N-корпуса. Зборка ад 10.08.2016 [Электронный ресурс] // Беларускі N-корпус [сайт]. 2018. URL: https://goo.gl/CS68hN (дата обращения: 17.09.2018).

5. Рычкова Л.В., Станкевич А.Ю. Лингвистический корпус СМИ Гродненщины: технология создания, направления использования: монография / Под науч. ред. Л.В. Рычковой. Гродно: ГрГУ, 2017. 115 с.

6. Хаген М. Полная парадигма. Морфология. Частотный словарь. Совмещенный словарь [Электронный ресурс] // Форум «Говорим по-русски» [сайт]. 2018. URL: https://goo.gl/wzdMxc (дата обращения: 17.09.2018).

7. MyStem / Сост. И. Сегалович, В. Титов [Электронный ресурс] // Технологии Яндекса [сайт]. 2018. URL: https://goo.gl/LDkP7p (дата доступа: 30.03.2018).

**Сведения об авторе:**

Станкевич Алеся Юрьевна – старший преподаватель кафедры русской филологии Гродненского государственного университета им. Янки Купалы (Гродно, Беларусь).

**Data about the author:**

Stankevich Alesya Yur'yevna – Senior Lecturer of Russian Philology Department, Yanka Kupala State University of Grodno (Grodno, Belarus).

**E-mail:** a.stan.lab@gmail.com.