Тамара & Леннарт Лённгрен

ПРИНЦИПЫ ЛЕММАТИЗАЦИИ СЛОВОФОРМ, НАПИСАННЫХ С ПРОПИСНОЙ БУКВЫ

В Институте славянских языков при Уппсальском университете составлен машинный фонд русских текстов объемом в 1 миллион словоупотреблений. Необходимые экономические средства для проведения работы были предоставлены шведским исследовательским органом "Humanistiska forskningsrådet". Основную работу над проектом осуществляли Леннарт Лённгрен и Людмила Ферм[1]. В последнее время в этой работе также принимала участие соавтор данной статьи, Тамара Лённгрен.

Материал фонда был подвергнут ряду обработок, касающихся прежде всего его словарного состава. В частности, была проведена "лемматизация", то есть приведение каждой словоформы к соответствующей "лемме". Под термином "лемма" в данном случае мы подразумеваем словарную (исходную) форму данной словоформы. Такой анализ всего лексического материала — в фонде содержится свыше 132.000 словоформ — очень трудоемкий, так как он только в ограниченной степени подлежит автоматизации. Однако лемматизация является необходимым условием для получения двух основных результатов настоящего проекта: конкорданса и частотного словаря, представленных совместно на уровне словоформы и леммы.

Лемматизация большого лексического материала является не только трудоемкой механической работой, но и ставит много сложных лингвистических задач, требующих нетривиальных решений. Одной из таких задач является лемматизация словоформ, написанных

[1]См. *Slovo* 36, с. 77-83 и 39, с. 125-141.

с большой буквы. Здесь мы не могли, как во многих других случая обращаться к толковым словарям, в которых, как известно, сло этой категории вообще не приводятся.

В обычном представлении, слова, написанные с большой буквы это прежде всего имена собственные. Тогда возможность возникнов ния проблем при лемматизации имени собственного может показать по меньшей мере странной: лемма словоформы *Москвой* будет толь *Москва*, а *Ольге* – только *Ольга*. На первый взгляд – все доволь понятно, но это только на первый взгляд...

Не углубляясь в проблему имени собственного – объекта иссл дования самостоятельной лингвистической науки ономастики, котором существует огромное количество литературы, – остановимс на предельно лаконичном его определении в "Русской грамматике" "собственные имена в узком смысле этого слова – это географически и астрономические названия и имена людей и животных"[2]. Но ведь прописной буквы пишутся и многие другие названия, не являющиес именами собственными в узком смысле, и для которых существуе довольно пестрая терминология, например, таксоны, номены, хрема тонимы, прагматонимы, документонимы, эргонимы, фалеронимы[3].

В многословных наименованиях с прописной буквы пишется, ка правило, первое слово, например *Тихий океан* (исключение составля ют случаи управления, типа *проспект Мира*, а также единичные слу чаи согласования, ср. *юбилейные Игры*). Написание остальных нарица тельных имен в составе названия регулируется определенными прави лами правописания[4], ср. *Золотой Рог* (бухта), но *Черное море*. Не

---

[2]*Русская грамматика*, М., 1980, с. 480.

[3]А. В. Суперанская, Аппеллятив – онома, *Имя нарицательное и соб ственное* (под ред. А. В. Суперанской), М., 1978, с. 5-33.

[4]Д. Э. Розенталь, *Вопросы русского правописания*, М., 1970, с. 26 и сл.

редко одно название входит в состав другого, ср. *Институт русского языка Академии Наук СССР.*

Что касается имен собственных в узком смысле, то они до настоящего времени в словарной практике лемматизации не подвергались. При составлении имеющегося русского частотного словаря под редакцией Л. Н. Засориной (ЧС)[5] собственные имена (очевидно, авторы отождествляют "имя собственное" и "антропоним") просто не выписывались из текста, не фиксировались клички животных и т. д. (ЧС, с. 14). В Уппсальском машинном фонде принято другое решение – ввести все без исключения словоформы, содержащихся в подобранных текстах. И здесь мы полностью разделяем мнение А. В. Суперанской о том, что "Собственные имена несправедливо игнорируются многими исследователями как маловажные, не занимающие должного места в лексическом составе языка. Однако, как бы ни был беден словарный запас человека, собственные имена неизбежно присутствуют в нем, а некоторые наиболее распространенные имена употребляются едва ли не чаще, чем многие нарицательные."[6]

Таким образом, в процессе обработки большого материала, положенного в основу создаваемого частотного словаря, вполне закономерно встал вопрос о выработке соответствующих принципов лемматизации словоформ, написанных с большой буквы. Здесь важно отметить, что мы не имеем в виду все слова, написанные с большой буквы в самих текстах. Уже при вводе текстового материала различались "синтаксическое" употребление прописной буквы, то есть в начале предложения, в начале цитаты и т. д., и "лексическое" ее употребление, в частности, в именах собственных. Только в последнем случае

---

[5]*Частотный словарь русского языка* (под ред. Л. Н. Засориной), М., 1977 (далее: ЧС).

[6]А. В. Суперанская, *Ударение в собственных именах в современном русском языке.* М., 1966, с. 24.

большая буква является признаком данной словоформы как лекси
ской единицы. При дальнейшей обработке словарного состава фон
было важно, чтобы большая буква "сохранялась" только в последн
случае. Это осуществлялось таким образом: специальным знаком
были отмечены все словоформы последнего – "лексического" – ти
В остальных словоформах информация о прописной букве не сох
нялась.

Однако и эта работа не была полностью лишена проблем. Проп
ные буквы сохранялись в основном только там, где данное назван
обладает хотя бы каким-то постоянством (хотя, может быть, полн
последовательность здесь не была достигнута). Кроме того, сохра
лись прописные буквы только в словоформах знаменательной час
речи (так, например, не была обозначена звездочкой (*) буква "Н"
случае: картина "На реке").

Самым простым решением проблемы было бы оставить прописну
букву в соответствующих леммах. Но это привело бы к нежелател
ным результатам особенно в таких случаях, когда одна и та
словоформа выступает в двух вариантах: со строчной и с прописн
начальной буквой. На уровне леммы мы бы получили, например, д
разные "академии": *академия* и *Академия* (ср. *учиться в академии*
Академия Наук СССР), два разных "тихий": *тихий* и *Тихий* (с
тихий день : *Тихий океан*) и т. д. В результате произошло бы лин
вистически неоправданное расщепление частотности подобных слов
форм по двум леммам, что, в свою очередь, привело бы к получени
неверных статистических данных. Неправомерность такой трактов
словоформ применительно к словосочетаниям, являющимся названи
ми, признавалась, по-видимому, и авторами ЧС, так как, наприме
Черное море они анализируют как одно употребление прилагател
ного *черный* и одно употребление существительного *море* (ЧС, с. 15
Чтобы избежать каких-либо неточностей в статистических подсчета

мы выработали определенные принципы лемматизации тех 11.677 словоформ, в которых сохранилась начальная прописная буква.

Дифференцированная трактовка словоформ, написанных с прописной буквы, обусловлена и такими соображениями: лексическая обработка машинного фонда сводится к ступенчатому обобщению языковых единиц, к их редукции. Первый шаг обобщения – это переход от текстов к списку словоформ (token → type), второй – от словоформы к лексеме. Нам представлялось естественным, чтобы этот шаг включал некоторую редукцию и в области правописания.

Таким образом, "судьба" прописной буквы определяется в нашей системе на трех различных уровнях: на уровне текста, на уровне словоформы и на уровне леммы. Одни прописные буквы существуют только в самих текстах (случай синтаксического употребления), вторые сохраняются как лексическая характеристика словоформы, но потом "сливаются" со строчными (*Тихого* [*океана*] → *тихий*), третьи же сохраняются вплоть до последнего уровня.

Теоретически не исключена возможность "обратной" лемматизации, то есть от словоформы со строчной буквой к лемме с прописной (например, в случае *рождество*, где наблюдается колебание, можно считать написание с прописной буквой ближе к новой складывающейся норме). Но, во избежание сложностей, мы отказались от этого направления лемматизации.

Словоформы с прописной буквой разделяются, в первую очередь, на две группы: слова, которые никогда не пишутся со строчной буквы, и слова, для которых мыслимы оба написания. Слова первого типа, конечно, никогда не приводятся к леммам со строчной буквой, независимо от структуры данного названия, ср. *Азия*, *Средняя Азия*. В связи с этим следует обратить внимание на своеобразную омографию типа *Дзержинский* (фамилия) : *Дзержинский район*, где по форме совпадают имя существительное и имя прилагательное; ср. также

Петропавловск-Камчатский (город), но: *Петропавловск-Камчатск училище*. В таких случаях прописная буква сохраняется только леммах имен существительных.

Для лемматизации словоформ второго типа – с двояким возмо ным написанием – очень существенно разделение названий на одн словные (простые) и многословные (составные). Последние, как пр вило, содержат по крайней мере одну словоформу, которая в друго контексте пишется со строчной буквы. В случае однословных назв ний границы слова и названия совпадают. Если данное слово по пр исхождению имя нарицательное или принадлежит к другой час речи, чем существительное, оно неизбежно значительно преобразуе ся. Имя нарицательное превращается в имя собственное и теря способность склоняться по категории числа, например (*река*) *Канав* (*деревня*) *Кулички*. Имя прилагательное, кроме того, переходит другую часть речи, в класс существительных, и в связи с этим теря способность склоняться по родам, например (*поселок*) *Дружно* Исходя из этих соображений мы решили привести все однословн названия к леммам с прописной буквой.

Исключением из названного принципа является более или мен окказиональное употребление прописной буквы в однословных "н именованиях". Известно, что в особом стилистическом употреблени с прописной буквы пишутся слова типа *Родина*, *Отчизна*, *Человек*; ср также следующие контексты: "принадлежность Себе и Делу"; ". полюбить Тень, Звук имени..."; "услыхала Его голос". Все таки случаи приведены к леммам с соответствующей строчной буквой. О окказиональных личных именах можно говорить в случаях тип *Папаня*, *Маманя*, а об окказиональном топониме – в случае "*Гд Лошади Плачут*" (Ф. Искандер: "Широколобый"); ср. также: "... сам Россия на перегоне от станции Вчера до станции Завтра" (Б. Василь ев: "Вы чье, старичье?"). В леммах подобных слов прописная букв

тоже не сохраняется. Такое решение оправдано тем, что "собственность" этих слов сохраняется только в определенном контексте, при особой стилистической обусловленности.

Еще одним исключением являются случаи, где наблюдается колебание. Например, не сразу было найдено однозначное решение лемматизации таких слов как *Запад*, *Восток*, (в смысле "западные, восточные страны"), *Север*, *Юг* (в смысле "территория"). С одной стороны, было бы закономерным оставить в леммах их словоформ прописную букву, но в связи с тем, что в текстовых файлах встречается двоякое написание этих слов в одном и том же значении, они были лемматизированы со строчной буквы (но с большой буквой остаются, конечно, употребления типа *папиросы "Север"*).

Разделение между однословными и многословными названиями было проведено не случайно. В многословных названиях, типа *Тихий океан*, *Средняя Азия*, *Набережные Челны*, входящие в состав данных словосочетаний нарицательные имена сами по себе не переходят в разряд имен собственных. В частности, они сохраняют свои морфологические и синтаксические свойства. Поэтому такие словоформы, написанные с прописной буквы, приводятся к леммам с соответствующей строчной буквой.

В результате вышеупомянутой "фильтрации" уже на уровне текстов из многословных названий остались, в основном, только словосочетания, состоящие из независимого существительного и разного рода определений. Названия другой синтаксической структуры – речь идет, прежде всего, о названиях произведений искусства, типа *картина "На реке"* – не отмечались специально в текстовых файлах и тем самым не сохранялись. Исключения составляют единичные случаи типа *журнал "Знание – сила"*, *журнал "Информатика и образование"*.

Приняв упомянутые принципы лемматизации, мы столкнулись с

особой трудностью: не всегда ясно, является ли данное назва словосочетанием, или только составной частью словосочетан Другими словами, иногда бывает трудно определить линейные (с тагматические) границы названия, что, разумеется, необходимо д правильной его классификации по отношению к признаку односл ности/многословности. Какими критериями руководствоваться случаях типа *курган Садовый* и *Садовая улица*? *проспект Мир улица Матросская Тишина*? Что следует включать в данное назван и что – нет?

Для решения этой проблемы необходимо остановиться на вопр о приложении, в роли которого выступает название, и о видах синт сической связи в таких конструкциях. В данном случае приложени это определяющий член словосочетания, а в качестве определяемо члена выступает нарицательное слово (редко – словосочетани Между определяемым словом и приложением существует два ви синтаксической связи: согласование или примыкание, причем пост зитивную позицию всегда занимает приложение, например, *река Не* (ср. *на реке Неве*), *река Дон*, (ср. *на реке Дон*). Следует обрати внимание на принципиальное различие между внешне похожи словосочетаниями *река Быстрая* и *улица Детская*. В последнем сл чае *улица* – это составная часть многословного названия. Поряд слов здесь является вторичным, маркированным, и возникает в ко текстах типа *Я живу не на Садовой улице, а на улице Детской*.

В приведенных случаях всегда возможно опускать определяем слово, ср. *на озере Байкал – на Байкале*. Если же данное имя нари тельное присоединяется с последующим зависимым именем собстве ным связью управления, то ни о каком приложении не может бы речи, например *проспект Мира*. Здесь в принципе невозможно опу кать независимое слово (только в разговорной речи можно услыша конструкции типа *Я живу на Мира*), так что границы словосочетан

и названия совпадают.

Изложенное понимание приложения несколько расходится со мнениями советских грамматистов. Так, в "Краткой русской грамматике"[7] приложение характеризуется как вид согласования (с. 352), но, по-видимому, одновременно воспринимается как определяющее слово, так как говорится о "приложениях-топонимах" (с. 354). Отмечается, что случаи, где нет согласования, "по существу выходят за рамки приложения" (с. 354). В. А. Белошапкова[8], не называя подобные случаи приложением, совершенно справедливо указывает на то, что связь примыкания здесь выражается "контактным постпозитивным расположением примыкающего компонента" (с. 54).

Ниже проиллюстрируем указанные принципы лемматизации на примерах, взятых из машинного фонда. Сначала рассмотрим однословные названия, потом – многословные. В обоих случаях данные названия могут быть или "самостоятельными", или выступать в роли приложения, и в соответствии с этим проводится главное разделение. Многословные названия, кроме того, разделяются по синтаксической связи. В случае "вложения" одного названия в другое данный пример классифицируется по связи между наиболее крупными составными.

1.1 Однословные самостоятельные названия

а) антропонимы (имена, отчества, фамилии, прозвища): Ирина, Искра, Зубр, Капа, Машка, Петька; Кузьмич, Петровна, Данилыч; Буренков, Гончаров, Тимофеев-Ресовский, Петров-Воткин; Стамеска, Рыжий, Пипка, Светка-Пипетка;

б) клички животных: Матера, Канитель, Купчиха, Красуля, Лохма-

---

[7] *Краткая русская грамматика* (под ред. Н. Ю. Шведовой & В. В. Лопатина), М., 1989.

[8] В. А. Белошапкова, *Современный русский язык. Синтаксис*, М., 1977.

тый;

в) астрономические названия: Сириус, Венера;

г) топонимы: Байкал, Крым, Серебрянка (река), Швеция, Мин Кислицы, Грязное (два последних – населенные пункты), Невск (проспект /эллипсис/);

д) административные названия, организации, предприятия: Интури Аэрофлот, Госкомиздат, Минстанкопром, КамАЗ, СССР, НИИ;

е) прочие названия: "Известия" (газета), "Мир" (спутник), "Севе (холодильник).

1.2 Однословные названия-приложения

а) антропонимы: старуха Дарья, директор Кутаков;

б) клички животных: корова Поляна, собака Тузик, кот Рыжко;

в) астрономические названия: планета Марс, звезда Солнце;

г) топонимы: река Волга, река Лазурь, курган Садовый, полуостр Рыбачий, озеро Лама, озеро Великое, высота "Плоская", роща "Све лая", канал Ростов, город Бахмач, город Озеры, город Улан-Уд деревня Кулички, деревня Степановская, село Крылос, станция "Ми ный", метро "Сокол";

д) названия предприятий: совхоз "Комсомолец", совхоз "Погорело ский", объединение "Швейтекстильтрикотаж", комплекс "Восто ный", трест "Железобетон";

е) прочие названия: кинотеатр "Чайка", пансионат "Ласточка", общес во "Искусство", журнал "Природа", спутник "Экран" (ср. также, определяемым словом в множественном числе: спутники "Метеор").

2.1.1 Многословные самостоятельные названия с согласованием

а) антропонимы (имя + отчество + фамилия; прозвища): Римм Львовна Львовская; Сухарь Сухарыч, Нюська Светлячок (с точк зрения современного языка здесь вряд ли можно говорить о словос

четании, так как данная синтаксическая связь скорее сочинительного, чем подчинительного характера); Володька Ржавый, Александр Македонский, Александр Невский (с несомненной подчинительной связью; в последних двух случаях прописная буква словоформ-прилагательных сохраняется и в соответствующих леммах, так как эти определения воспринимаются как имена собственные);
б) астрономические названия: Млечный путь, Большой Пес;
в) топонимы: Лисий Нос, Кривой Рог, Набережные Челны, Северная бухта, Азовское море, Кольский полуостров, Серебряный бор, Кузякин лог, Садовое кольцо, Малая Садовая (улица /эллипсис/);
г) административные названия, предприятия: Таджикская ССР, Ясногорский район, Дзержинский район, Усть-Лабинский район, Краснодарский край, Херсонская область; Октябрьский народный суд, Копейский горком партии, Калевальский райком КПСС, Воронежский облисполком, Селигеровский сельсовет; Светлогорский целлюлозно-бумажный комбинат, Селенгинская фабрика, Чебоксарская АЭС; Советские Вооруженные Силы;
д) прочие названия: Варшавский Договор, Основной Закон; Генеральный Секретарь, Верховный Главнокомандующий, Нобелевские лауреаты; Золотая медаль; "Литературная газета", "Красная Звезда" (газета); юбилейные Игры.

2.1.2 Многословные самостоятельные названия с управлением

а) астрономические названия: созвездие Дракона, альфа Большого Пса;
б) топонимы: море Лаптевых (лемматизируется: море, Лаптев), проспект Мира, площадь Революции, улица Свободы;
в) административные названия: Союз Советских Социалистических Республик, Страна Советов, Совет Министров СССР, Верховный Совет СССР, Верховный Суд СССР, Ревизионная Комиссия КПСС,

Комитет Мира, Институт имени И. В. Курчатова, коммуна имени Ф Дзержинского (в двух последних примерах имеет место "вложен одного словосочетания в другое, причем подчиненное словосочета состоит из названия-приложения и определяемого слова *имя*);
г) прочие названия: Итоговый документ Венской встречи, Заклю тельный акт общеевропейского совещания, Соборное уложение ца Закон Эстонской ССР, Февральский Пленум ЦК КПСС; Председат Совета Министров, Герой Социалистического Труда; Орден Крас Звезды; Игры I Олимпиады, велогонка Мира; могила Неизвестн Солдата.

2.2.1 Многословные названия-приложения с согласованием

а) антропонимы: школьник Сева Гребин, некий старик Евсей Быков;
б) топонимы: улица Матросская Тишина, улица Первый Смоленск Ручей, деревня Белый Бор, поселок Старые Ляды;
в) названия предприятий: колхоз "Ленинский путь", объединен "Краснодарский чай";
г) прочие названия: газета "Советский спорт", журнал "Всемирн следопыт".

2.2.2 Многословные названия-приложения с управлением

а) топонимы: село Годы Турка;
б) названия предприятий, прочие названия: совхоз "Заветы Ильича газета "Огни Ангары", журнал "Вопросы философии".

Своеобразно ведут себя слова типа *имя, название*, которые мог выступать в качестве определяемого слова приложения: *имя Паве название Тверь*, или же просто подчинять себе название, не явля щееся приложением: *имя Сталина, название Солодовой улицы.* первом случае имеет место примыкание или согласование (ср. *назв*

ние "Разбойниково", но: *фамилия Серпилин – фамилию Серпилина*), во втором – управление. Словосочетания этого типа – то же касается словосочетания с определяемым словом "слово" (ср. *слово Тверь, слово "Канны"*) – ввиду их метаязыкового характера не были включены в изложенную классификацию.