СЛОВАРЬ БЫТОВОЙ ТЕРМИНОЛОГИИ: НОВЫЕ ПРОБЛЕМЫ И НОВЫЕ МЕТОДЫ[1]

Иомдин Б. Л. (iomdin@ruslang.ru),
Лопухина А. А. (nastya-merk@yandex.ru)
Институт русского языка им. В. В. Виноградова РАН

Пиперски А. Ч. (apiperski@gmail.com)
Московский государственный университет
им. М. В. Ломоносова

Киселева М. Ф. (mar-fed@yandex.ru),
Носырев Г. В. (grigorij-nosyrev@yandex.ru),
Рикитянский А. М. (nordicdx@gmail.com)
Яндекс

Васильев П. К. (asthmor@gmail.com),
Кадыкова А. Г. (annakadykova@nm.ru),
Матиссен-Рожкова В. И. (heinin@mail.ru)
Российский государственный гуманитарный университет,
Москва

В докладе рассматриваются проблемы, связанные с составлением нового словаря-тезауруса бытовой терминологии русского языка. Привлечение нового материала ставит перед лексикографами интересные задачи, требующие новых подходов. Большое внимание уделяется работе с данными Интернета, в частности для определения частотности употребления лексем и их орфографических вариантов в текстах разных жанров и в логах пользовательских запросов. Описываются инструменты, созданные авторами словаря для получения достоверных численных данных. Важное место среди активно используемых методов анализа лексики занимают опросы информантов и эксперименты. В докладе описываются их условия и анализируются полученные результаты. Результаты работы могут оказаться полезными не только в лексикографии, но и в теоретической семантике, поскольку рассматриваемый материал, ранее не подвергавшийся системному изучению, обнаруживает нетривиальные свойства.

Ключевые слова: семантика, лексикография, предметная лексика, орфография, орфоэпия, частота, статистические методы

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», гранта РГНФ №10-04-00273а и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ.

THESAURUS OF RUSSIAN EVERYDAY LIFE TERMINOLOGY: NEW PROBLEMS AND NEW TECHNIQUES

Iomdin B. L. (iomdin@ruslang.ru),
Lopukhina A. A. (nastya-merk@yandex.ru),
V. V. Vinogradov Russian Language Institute,
Russian Academy of Sciences

Piperski A. Ch. (apiperski@gmail.com),
M. V. Lomonosov Moscow State University

Kiselyova M. F. (mar-fed@yandex.ru),
Nosyrev G. V. (grigorij-nosyrev@yandex.ru),
Rikityanskiy A. M. (nordicdx@gmail.com),
Yandex

Vasilyev P. K. (asthmor@gmail.com),
Kadykova A. G. (annakadykova@nm.ru),
Matissen-Rozhkova V. I. (heinin@mail.ru),
Russian State University for the Humanities (Moscow)

The paper addresses various issues associated with developing a new encyclopedic thesaurus of Russian everyday life terminology (a current project by a group of researchers in V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences). As more new lexical material is involved, lexicographers are faced with interesting tasks, which require new approaches. Dealing with Internet data is one of the most important issues. Calculation of spelling variants and synonym frequencies in texts of various genres, in particular, in blogs and user query logs, is performed using new Yandex-based tools specially designed for the project. The layout and the results of several kinds of speaker surveys and experiments are discussed in detail. The results of the paper may prove useful for various lexicographic projects as well as for theoretical linguistics. The data under discussion, which has never been systematically studied before, appears to possess peculiarities worth further deep semantic research.

Key words: semantics, lexicography, everyday life vocabulary, spelling variants, orthoepy, frequency, statistical techniques

В работах [Иомдин 2009, 2011, Iomdin et. al. 2011] были рассмотрены проблемы, связанные с описанием бытовой предметной лексики, и предложен проект создания нового иллюстрированного словаря-тезауруса бытовой терминологии (СБТ). В 2010 году в Институте русского языка им. В. В. Виноградова РАН был организован и начал работу научный семинар под руководством Б. Л. Иомдина[2], на котором обсуждаются материалы будущего словаря. Словник словаря постоянно пополняется, на текущий момент он составляет более 1600 слов. Предполагаемая структура словарной статьи опубликована в [Иомдин 2011]. Настоящий доклад построен следующим образом: каждая зона словарной статьи комментируется с точки зрения возникающих проблем и предлагаемых методов их решения.

1. Группа слов с близким значением (вход)

1.1. Орфография

При составлении толковых словарей проблема орфографии обычно не является ведущей, поскольку их авторы используют данные специализированных словарей; так, словарь [Апресян 2010] опирается на [Орфографический словарь 2005]. Однако в активно используемой бытовой лексике существенную часть (по нашим подсчетам, не менее четверти) составляют новые слова, еще не зафиксированные словарями. Написание этих слов часто еще не устоялось. В условиях отсутствия признанной нормы приходится ориентироваться на узус, иногда противоречивый (ср. варианты *флэшка/флешка, пихора/пехора, ролеты/роллеты, пусета/пуссета, хайратник/хаератник*). Порой возникает вопрос, предпочесть ли более частотное написание этимологически обоснованному. Так, слово *борсетка*, восходящее к итальянскому *borsetta*, гораздо чаще пишут как *барсетка* (ср. ситуацию со словом *колготки*, восходящим к чешскому *kalhoty*, в [Левонтина 2010: 240–243]).

В последнее время для определения частотности написания часто используют данные поисковых систем (количество найденных документов). Однако такой подход ненадежен, особенно для слов, частоты которых не отличаются на порядки. Во-первых, поисковые системы расширяют запрос не только словоформами, но и синонимами, ассоциациями, транслитерацией, переводом и т. п. (что лишь частично решается использованием языка запросов). Во-вторых, количество найденных документов зависит от многих технических факторов, например, размера поисковой базы, загруженности серверов в момент запроса и т. д. Результаты по одному и тому же запросу за один период времени

---

[2] Кроме авторов доклада, в число основных участников семинара входят: Ф. Г. Винокуров, А. Н. Выборнова, С. А. Колмановская, А. С. Панина, А. Д. Стрижевская. Авторы хотели бы выразить признательность всем названным коллегам за активное участие в работе семинара и в подготовке материалов к докладу.

могут существенно различаться (так, запрос *портфель*, отправленный в Яндекс в декабре 2011 года с интервалом в 5 дней, выдал 428 000 и 1 000 000, соответственно). Ср. обсуждение сходных проблем в [Kilgariff 2007; Беликов, Ахметова 2009; Беликов 2011].

Помимо числа найденных документов, о частоте слова можно судить по другим источникам, например, по текстам блогов или по логам запросов пользователя (см. ниже).

Правильное написание часто определяется по поведению поисковой системы, которая может по-разному реагировать на запросы. Например, запрос *кожаная ссумка* будет исправлен на *кожаная сумка* автоматически (автозамена), на запрос *легинсы* система отреагирует подсказкой «быть может, вы искали: *ле**гг**инсы*» (подсказка), а запрос *повербол* признает безошибочным. В случае автозамены вероятность того, что введенный запрос содержит опечатку, больше, чем в случае подсказки. Если исправления не произошло, то, скорее всего, это означает, что единый вариант написания не устоялся. Чтобы увеличить надежность определения опечатки, используется не один однословный запрос, а коллекция неоднословных запросов (*кожаные ссумки*, *женские ссумки*, *дизайнерские ссумки*). Список возможных вариантов написания слова можно получить из пользовательских запросов, выбрав ближайшие по взвешенному расстоянию Левенштейна [Левенштейн 1965]. Этот метод хорошо работает на словах, в которых неоднозначны гласные (*поуэрбол*, *пауэрбол*, *повербол*).

1.2. Орфоэпия

Сходные проблемы возникают и с указанием произношения. Отметим две из них: определение места ударения (*пиала*, *боло, нострил(л)а*, *зиппо*, *унты*, *шарфы*) и произношение твердых/мягких согласных перед звуком, обозначаемым буквой *е* (*постер, дерби, трекки, лабрет(т)а*).

При определении места ударения возникают проблемы трех родов.

1) В недавно заимствованных словах в противоречие вступают две тенденции. С одной стороны, русский язык имеет склонность оформлять заимствования «колонным» ударением на последнем слоге основы, если она оканчивается на согласный, и на предпоследнем слоге, если она оканчивается на гласный (более подробно см. в [Суперанская 1968]). С другой стороны, источником большинства заимствований является английский язык, и многие носители русского языка, используя модное новое слово, произносят его с английским ударением, которое обычно падает на первый слог. Естественно, что в случае неоднословных основ часто возникает конфликт. Примером такого конфликта является слово *ноутбук*, и даже словари здесь расходятся: [Русское словесное ударение 2001] предлагает вариант *ноутбу́к*, а [Орфографический словарь 2007] — *но́утбук*. Таким образом, в этом и других подобных случаях идет борьба двух вариантов тривиального ударения (ударения, которое всегда падает на основу, причем на один и тот же по счету слог слева, см. [Зализняк 1985: 17]), падающих на разный слог основы.

2) С редко используемыми словами возникает иная проблема: конкурируют нетривиальное нормативное и тривиальное ненормативное ударение. Например, обувь *коты́* многие склонны называть *ко́ты*; для слова *доска* в молодежной речи в последнее время отмечается ударение на первом слоге.

3) В свою очередь, хорошо освоенные слова имеют тенденцию переходить от тривиального ударения к нетривиальному (см. [Зализняк 1985: 23]), и не всегда понятно, насколько далеко этот процесс зашел для каждого конкретного слова. Неясно, например, можно ли уже признать нормативным произношение формы мн. ч. от слова *шарф* с конечным ударением (*шарфы́*) или надо по-прежнему указывать вариант *ша́рфы*.

Что же касается твердости/мягкости согласных перед *е*, она, как и ударение, коррелирует с освоенностью слова. Орфоэпические данные такого рода могут быть получены только в результате опросов: нормативные словари, как и в случае с орфографией, часто не успевают отразить новую лексику.

1.3. Словообразовательные варианты

Интересную проблему представляют существительные с диминутивными суффиксами. Как известно, уменьшительность часто лексикализуется, что требует включения таких слов в качестве отдельных входов: ср. *театральная сумочка <*сумка>*, *шапочка <*шапка> для душа*, *ремешок <?ремень> для часов*, *половая тряпка <*тряпочка>* vs. *тряпочка <*тряпка> для очков*. В процессе работы над словарем обнаружилось системное различие, особенно характерное для названий предметов одежды: уменьшительное существительное часто применимо только к детским или женским вещам (ср. *трусики — трусы*, *поясок — пояс*, *кофточка — кофта*). Кроме того, диминутивы часто характерны для женской речи (так, по нашим данным, мужчины реже говорят *колечко*, чем *кольцо*). В некоторых случаях неуменьшительное слово используется в официальной номенклатуре, ср. *брошка — брошь*[3], *батарейка — батарея* и т. п.

2. Толкование совпадающей части значения

В лексикографии традиционно признается, что исчерпывающее толкование предметной лексики неосуществимо; ср., например, [Апресян 2010: 89]. Цель нашего словаря — максимально полно представить сведения о значении и употреблении слов с предметным значением в разных стилистических пластах современного русского языка. Одна из трудностей описания значения предметных слов — указание размера. В соответствии с принципами Московской семантической школы мы стремимся к использованию природных “эталонов”. Так, размер *флешки* указывается как ‘не больше пальца’, размер

[3] Наблюдение И. Б. Левонтиной.

карточки — ‘примерно с ладонь’, размер *бумажника* — ‘примерно с кисть руки’. Но иногда найти такой эталон не удается, и более целесообразным кажется использовать артефакты: лист бумаги формата А3 для *дипломата*, паспорт для *ксивника*.

3. Близкие слова и группы слов

В СБТ принят принцип описания слов по группам, подобно синонимическим рядам в [НОСС 2004]. Деление “непрерывного пространства языка”, в любом случае искусственное, но необходимое в лексикографической практике, иногда особенно затруднительно. Одна из типичных ситуаций — наличие в значении слова компонента, конфликтующего с толкованием совпадающей части значения. Так, *чулки* — не **колготки** (являются парным предметом и не покрывают нижнюю часть туловища) и не **носки** (охватывают ногу выше колена, в основном эластичные и тонкие), но против выделения их в отдельную группу говорит тот факт, что и *носки*, и *колготки* могут служить родовым словом для *чулок*. Похожая проблема возникает с составом группы **перчатки**: стоит ли включать туда слово *муфта*? Аргументом “за” является сходство функций и материала; против — признаки ‘парность’ и ‘количество отделений для пальцев’. Таким образом, можно счесть *чулки* и *муфту* маргинальными предметами, которые, примыкая к некоторой группе, отличаются от ее доминанты сильнее других слов группы, а можно описывать эти слова отдельно, тем самым умножая количество групп.

4. Доминанта группы

В [Иомдин 2010] рассматривались различные критерии выделения доминанты в группе слов со сходным значением. Один из способов определить доминанту — эксперименты на представительной выборке носителей. Мы проводим эксперименты нескольких типов. Наиболее простой — прямое обращение к носителям с просьбой назвать предметы, изображенные на картинке (так изучались классы СУМКИ и ГОЛОВНЫЕ УБОРЫ). Более показательно описание картинки, на которой изучаемый предмет находится не в фокусе внимания испытуемого, и выявление таким образом слова, наиболее часто выбираемого для описания сходных предметов (так проводился эксперимент для групп **свитер** и **сандалии**). Еще один письменный эксперимент устроен как модифицированная игра в “банальности”. Испытуемым демонстрируется предмет и дается задание назвать его тем словом, которым его назовет большинство носителей (стимулом является получение числа очков, соответствующее числу игроков, выбравших то же слово). Так изучались слова в классах ОБУВЬ и ГОЛОВНЫЕ УБОРЫ. Это позволяет выявить самое “банальное”, по мнению большинства, название предмета, которое предположительно и является доминантой группы.

При выборе доминанты часто возникает проблема противоречия критериев ее определения, даже если ориентироваться на частотность. Так, в нормативных (и любых сколько-нибудь официальных) документах встречается практически исключительно *бюстгальтер*, а в разговорной речи безусловно лидирует *лифчик*; *сорочка* из ГОСТов и ценников соответствует повседневной *рубашке*.

Еще одна проблема — существование доминант “разных уровней”.

Так, класс ОБУВЬ естественно делить на следующие группы: **сапоги**, **ботинки**, **туфли**, **кроссовки**, **тапочки**, **галоши**, **сандалии**. Однако при таком разделении возникает несколько трудностей, связанных с распределением слов по этим группам.

Внутри группы слов, объединенных одной доминантой, могут встречаться собственные «субдоминанты». Так обстоят дела с группой **сандалии**, куда входят лексемы *сандалии, босоножки, шлепанцы, шлепки, вьетнамки, сланцы*. Носители разделяют *сандалии* и *босоножки* с одной стороны, и *шлепанцы*, *вьетнамки* и *сланцы* — с другой. При этом в первом случае «субдоминантой» будет слово *сандалии*, а во втором — *шлепанцы*. Результаты опроса по картинкам, на которых летняя обувь была не в фокусе внимания, показывают, что обувь с открытой пяткой и без каблука называют *шлепанцами* приблизительно в 40 % случаев (остальные употребления значительно менее регулярны), а обувь с закрытой пяткой называют сандалиями примерно в 60 % случаев.

Слово *тапочки* в первом значении описывает домашнюю обувь, но у него выделяется и другое значение: ‘легкая обувь на низком каблуке, предназначенная для ношения в теплое время года’. Получается, *тапочки* в этом случае могут служить гиперонимом для таких слов, как *мокасины*, *топ-сайдеры* (**ботинки**), *тенниски* (**кроссовки**), *балетки, балеринки, эспадрильи* (**туфли**), *босоножки*, *шлепанцы* (**сандалии**). Значит, это — своеобразная «супердоминанта», способная заменять слова из разных групп. Ср.: (1) *В этот самый момент в кабинет вошла женщина в форменной куртке, в фуражке, в черной юбке и в тапочках* (М. Булгаков, Мастер и Маргарита); (2) *Я вот на днях съездил в Баку, закупил партию летней обуви, шлепанцев, тапочек всяких* («Бизнес-журнал», 2004.08.17); (3) *Весна, а дворник наш уже в тапочках, в шлепках* («Русская Жизнь», 2008).

В некотором роде «супердоминантой» оказывается и слово *ботинки*, поскольку носители часто называют так мужские туфли. Ср.: (4) *Перед концертом я успел найти в магазине пару ботинок под костюм для первого отделения, а на второе — обуви не нашлось* («Вечерняя Москва», 2002.01.10); (5) *Погуляв некоторое время на улице, я снова пришла, позвонила в дверь, и мне открыл Андрей, одетый в черный костюм с бабочкой и шикарные ботинки* («Комсомольская правда», 2007.08.16). Слово *chaussure* в названии французского фильма «Le grand blond avec une chaussure noire» было переведено как *ботинок* («Высокий блондин в черном ботинке»), хотя на плакате к фильму изображена мужская туфля. Однако объединять **ботинки** и **туфли** в одну группу кажется неправильным, потому что в таком случае в группу **ботинки** войдут, например, *лодочки*, *шпильки*, *сабо*, *танкетки*, а наши опросы показывают, что носители обычно не готовы называть *ботинками* вышеперечисленные предметы обуви.

Определенную сложность представляет и описание класса ГОЛОВНЫЕ УБОРЫ. Прежде всего, само название класса принадлежит к официальному или канцелярскому стилю и редко используется в разговорной речи; вместо него обычно употребляется слово *шапки*. С другой стороны, этот класс включает в себя несколько групп, в том числе **шапки** (например, *ушанка, петушок, кепка*) и **шляпы** (например, *канотье, ток, стетсон*). Таким образом, слово *шапка* может выступать и в роли «супердоминанты», и в роли доминанты своей группы, противопоставленной *шляпе* как доминанте другой группы.

Особый случай — ситуация, когда предполагаемая доминанта является наиболее частотной и наиболее нейтральной в группе слов, называющих предметы со схожей функцией, однако ни одно из слов группы не допускает замены на другое: ср. *зажигалка, огниво, прикуриватель, спички.*

5. Различительные признаки, релевантные для данной группы

Серьезные трудности, возникающие при попытке истолковать слова с предметным значением, подробно обсуждались в [Иомдин 2009, 2011]. Привлечение нового материала выявляет дальнейшие проблемы. Так, при описании слова *сумочка* обнаружилось, что необходимо так или иначе отразить в толковании следующие смыслы: ‘небольшой размер’ (vs. *баул*), ‘элегантность, претензия на стиль’ (vs. *хозяйственная сумка*), ‘использование для транспортировки предметов, которые используются в течение дня’ (vs. *чемодан*), ‘прочный материал’ (vs. *пакет*), ‘пол обладателя’ (vs. *борсетка*). Однако при этом наличие всех названных признаков одновременно не является необходимым: *сумочкой* могут назвать и стильную дамскую сумку с украшениями, куда с легкостью помещается ноутбук (отсутствует признак ‘небольшой размер’), и небольшую дорожную сумку на пояс (отсутствует признак ‘элегантность’).

В случае, если признак действительно релевантен, далеко не всегда удается подобрать точное и общепонятное название для этого признака. Так, термин *тулья* удобен для описания шляп, но почти не известен носителям.

6. Смысловые различия

Основной способ уточнения значения названий предметов быта — опросы информантов, призванные выявить различия между названиями сходных предметов, существующие в их идиолектах.

В одном из наших экспериментов испытуемым были предложены следующие группы слов: *кувшин-графин*; *чашка-кружка*; *колье-ожерелье-бусы*; *джемпер-пуловер*; *варежки-рукавицы*; *шарф-кашне*; *кепка-бейсболка-кепи*; *платок-косынка*; *портфель-рюкзак*; *дипломат-кейс*; *маркер-фломастер*; *шпилька-невидимка*; *кошелек-бумажник-портмоне*; *банкнота-купюра*. Для каждой группы

слов хотя бы несколько испытуемых констатировали различия. Сопоставив все ответы, мы пришли к следующим выводам:

1. Существуют пары слов, для которых не удалось выявить никаких объективных различий: *джемпер-пуловер*; *банкнота-купюра; шарф-кашне*; *дипломат-кейс*. Больше половины опрошенных сходятся на том, что значения этих слов не различаются, остальные приводят различия, не сводимые к одному основанию.

2. В других случаях различия достаточно хорошо осознаются опрошенными (удалось выявить несколько регулярных различительных признаков). Это пары слов: *варежки-рукавицы; портфель-рюкзак; чашка-кружка; маркер-фломастер*.

3. Для некоторых слов четко выделяется один различительный признак, а другие многочисленные различия и близко не набирают большинства среди ответов: *кувшин-графин* (‘материал’); *кошелек-бумажник-портмоне* (‘назначение’).

4. Для слов, входящих в группы *колье-ожерелье-бусы*; *кепка-бейсболка-кепи*; *платок-косынка*; *шпилька-невидимка*, не удается выявить «лидера» среди различительных признаков. Большинство испытуемых считают, что различия внутри этих групп есть, и описывают их. Однако ни одно из предложенных различий не набирает хотя бы половины голосов. Для одних и тех же слов внутри группы разные опрошенные приводят противоположные признаки, например, *платок-косынка*: ‘квадратная форма’; *колье-ожерелье*: ‘крепится под горлом, плотно охватывает основание шеи’; *шпилька-невидимка*: ‘наличие декоративного элемента’. Для слов *кепка-бейсболка-кепи* различительные признаки многочисленны (‘форма’, ‘козырек’, ‘назначение’, ‘стиль’, ‘пол обладателя’, ‘цвет’, ‘материал’), но ни один из них не выделяется.

Стоит также отметить следующее. В группах *кепка-бейсболка-кепи*, *кошелек-бумажник-портмоне*, *шарф-кашне* слова *кепи*, *портмоне*, *кашне* в описании чаще всего характеризовались эпитетами «изящный, изысканный, для солидных дам» (*кашне*), «гламурное, престижное» (*портмоне*), «что-то английское, экзотическое, приличное» (*кепи*) — в анкетах разных опрошенных. По-видимому, это связано с тем, что они еще не до конца освоены (морфологически это выражено в том, что слова *кепи*, *портмоне* и *кашне* не склоняются, в отличие от хорошо освоенных заимствований *кепка*, *бейсболка*, *шарф*), поэтому предмет, обозначаемый ими, приобретает коннотацию “элитарности” по сравнению с другими словами и предметами из тех же групп.

Результаты эксперимента выявляют интересную проблему: правомерно ли постулировать различия в значениях обычных слов, называющих предметы быта (не терминов!), если они отражены в нормативных источниках (толковых словарях), но не признаются или не ощущаются большинством образованных носителей? Как описывать предметы, различительные признаки которых неочевидны? Задача нашего словаря — по возможности представить весь спектр мнений носителей (соотнесенных с данными об их социальном статусе), чтобы пользователь словаря мог сопоставить эти сведения с данными нормативных

источников, приводимых в соответствующих зонах, и принимать решение в зависимости от своей коммуникативной задачи.

7. Другие значения слов, входящих в группу

Особую сложность при лексикографическом описании предметов быта представляют случаи полисемии. Так, например, лексемы слова *шлепанцы*, по-видимому, входят в группы <1> домашняя обувь с открытой пяткой (**тапочки**), <2> уличная обувь, которую носят в теплую погоду (**сандалии**), обозначая при этом один и тот же предмет. Ср.: (6) *Точно так же тело змеи трется о сухие листья, каблучки отбивают дробь по асфальту, а тапочки старого человека, которому уже трудно поднимать ноги, при каждом его шаге шлепают по полу (неслучайно одна из разновидностей мягких тапочек так и называется — шлепанцы)* (И. Иткин, «Наука и жизнь», 2006); (7) *А потом вдруг обнаружила, что они уже все втроем едут в такси, причем на ногах у Фаины были домашние шлепанцы* (Т. Тронина); (8) *Парень сплюнул, бросил сыпанувший искрами окурок и, пошаркивая пляжными шлепанцами, двинулся обратно на станцию* (А. Иличевский). Получается, что один и тот же предмет, называемый одним и тем же словом, тем не менее попадает в две группы, отличающиеся по функции.

Еще одна трудность — соотнесение в этой зоне лексем разных уровней, как, например, в случае со словом *тапочки* (‘домашняя обувь’ или ‘легкая обувь на низком каблуке, предназначенная для ношения в теплое время года’): здесь приходится описывать “супердоминанту” (см. выше) как другое значение слова, что не позволяет различие “уровней” описания.

8. Другие редкие слова, примыкающие к группе

Среди слов с близким значением, называющих бытовые предметы, неизбежно выделяются ядерная и периферийная части. Основным критерием для отнесения слов в ту или иную часть является их относительная частотность. Информацию о частотах можно получать из разных источников: частотный словарь (например, [Ляшевская, Шаров 2009]), НКРЯ, опросы информантов. Однако все эти способы обладают существенными недостатками: быстрое устаревание данных, недостаточная представленность разговорной, сленговой и т. п. лексики и недостаточность выборки для статистически значимых выводов.

Для подсчета различных словарных статистик можно использовать готовые сервисы, например: Яндекс-статистика (http://wordstat.yandex.ru/), язык

запросов (http://help.yandex.ru/search/?id=1111313), Google Ngram Viewer (http://books.google.com/ngrams), Google Trends (http://www.google.com/trends/). Существуют и иные источники информации о частотности, например, поиск по блогам или текстам запросов пользователей, главные преимущества которых — стабильность и воспроизводимость результатов, замкнутость системы (частоты слов сравнимы), большой объем и актуальность данных, а недостатки — омонимия и высокая специфичность материала.

8.1. Блоги

Поиск и подсчет статистики по Яндекс.Блогам связан с техническими трудностями, общими для любых поисков в Интернете. Отметим еще, что для получения более точного результата необходимо отделять текст поста от комментариев, что, как правило, сложно реализовать.

Разработанный внутри Яндекса специально для нашего словаря инструмент позволяет обойти все описанные проблемы и получить доступ к исходным текстам Яндекс.Блогов. Основной функцией этого инструмента является подсчет точного числа вхождений заданного слова в текстах постов с учетом всех словоформ. Итоговая частота слова складывается из частот всех его словоформ.

8.2. Логи

Пользовательские запросы (логи) являются актуальным источником разносторонней информации для лексикографа. В нашем эксперименте для подсчета частот использовались логи за месяц (декабрь 2011), обрезанные по частоте 10, что составляет порядка 12 миллионов различных запросов.

Словоформы могут иметь разную частоту, и словарная форма не всегда преобладает, поэтому под частотой слова мы понимаем сумму частот всех его словоформ[4]. Частоту слова мы ассоциировали с количеством запросов, в которых оно встретилось.

Основная проблема при определении частоты слова — развитая омонимия (так, по словоформе *гвоздики* могут быть найдены слова *гвоздик* ‘небольшой гвоздь’, *гвоздики* ‘сережки’, *гвоздика* ‘специя’, *гвоздика* ‘цветы’, *Гвоздиков* ‘фамилия’ и т. д.), снятие которой весьма трудоемко. Методы снятия омонимии, рассмотренные, например, в работах [Сокирко, Толдова 2005], [Зеленков и др. 2005], используют большие размеченные корпусы текстов, что неприменимо

[4] Для построения парадигмы слова использовалась технология mystem (http://company.yandex.ru/technologies/mystem).

к нашему материалу: создание таких корпусов по трудоемкости превысило бы ручное снятие омонимии[5].

Рассмотрим предлагаемый нами алгоритм на примере двух слов: *зажим* (для денег) и *ключи* (от дома) (уточнение в скобках описывает целевой смысл). Частота каждого из этих слов порядка 10^5 и $8{\cdot}10^6$ соответственно (за один месяц). Для точного подсчета требуется разделить все пользовательские запросы, содержащие эти слова, на три группы в соответствии с тем, в каком смысле в них употреблено слово (в целевом смысле, не в целевом смысле или затруднительно ответить). Запросы третьей группы удаляются из выборки. Поскольку размечать все запросы часто чересчур долго (скажем, для слова *ключи* их порядка 10^4), можно выбрать заданный процент либо фиксированное число самых частых запросов.

Таблица 1. Пример снятия омонимии

зажим для денег	6611	+	*скачать ключи*	1678566	–
зажим для галстука	4428	–	*горячие ключи*	60068	–
зажим купить	3765	?	*ключ смотреть*	59478	–
зажим анкерный	1947	–	*рабочие ключи*	37938	–
			ключи от квартиры	3175	+
всего запросов со словом *зажим*	94563		всего запросов со словом *ключи*	8668411	
доля целевого смысла ‘для денег’	0,4		доля целевого смысла ‘от дома’	0,002	
скорректированная частота	37058		скорректированная частота	17659	

Такой метод позволяет определить частоту слова в любом заданном значении; можно также повышать точность для отдельных слов, увеличивая количество размеченных запросов. К недостаткам следует отнести необходимость использовать ручной труд, затруднения при разметке запросов (что делать с названиями, например, *фильм «Шпильки»*?). В дальнейшем можно автоматически разметить часть запросов, используя стоп-слова (например, если в запросе со словом *зажим* есть слово *деньги*, то это скорее всего целевой смысл; если в запросе со словом *ключи* есть слово *скачать*, то это скорее всего нецелевой смысл).

Важно отметить, что тексты блогов, логов и иных документов в Интернете могут существенно различаться по сравнительной частотности употребления искомых слов. Поэтому с уверенностью ответить на вопрос, относится ли рассматриваемое слово к ядру или периферии, можно лишь в том случае, если соотношение по всем типам текстов сравнимо; в противном случае можно лишь

---

[5] Отметим также, что в ряде случаев нам необходимо различать и полисемию, ср. *шлепанцы*, *тапочки*, *бахилы*.

говорить о стилистических различиях и снабжать слова соответствующими пометами.

8.3. Результаты

По полученным наблюдениям за период 01.12.2011–31.12.2011 видно, что частоты, посчитанные с помощью поисковых машин, нестабильны и часто дают неверное представление о распределении слов в Интернете (см. таблицу 2: бóльшее значение выделено полужирным шрифтом).

Таблица 2. Статистическая информация по разным интернет-источникам

Слово	Инструмент поиска по Яндекс.Блогам	Открытый поиск на blogs.yandex.ru	Инструмент пологам	Яндекс**	Google**
сапоги	**294 710**	**61 502** 61 250*	**569 271**	**571 000** **2 000 000***	2 230 000 2 180 000*
ботинки	180 019	41 089 41 079*	361 612	403 000 1 000 000*	**2 410 000** **2 380 000***
кружка	**510 153**	58 343 58 287*	**113 792**	**767 000** **2 000 000***	946 000 913 000*
чашка	422 350	**69 090** **69 068***	54 628	653 000 1 000 000*	**1 890 000** **1 840 000***

* — результат по запросу через 5 дней.

** — количество найденных документов

В зоне редких слов также может оказаться целесообразным указывать оценочные слова, чья семантика существенно размыта по сравнению с названиями конкретных предметов одежды, ср. *обдергайка*, *нахлобучка*, *кацавейка, прощай молодость* и т. п.

9. Региональные варианты

Проблемы, связанные с региональными вариантами русского языка, подробно изучаются в рамках проекта [Словарь “Языки городов”], где существенная часть словника относится к области предметной лексики; ясно, однако, что здесь предстоит еще большая работа. В проекте СБТ также планируются исследования региональных вариантов лексики с использованием данных поисковых систем и проведение соответствующих опросов информантов.

10. Данные нормативных документов

Не меньшую проблему вызывает поиск информации в нормативных документах типа ГОСТов. Это обусловлено, в частности, тем, что в таких документах принципиально иной принцип классификации предметов. Так, *галоши*, которые в словаре бытовой терминологии попадают в категорию ОБУВЬ, в ГОСТе следует искать в разделе “резиновые изделия”. Кроме того, для многих описываемых предметов нормативных документов вовсе не существует (например, для маркеров, кроссовок, перстней и пр).

11. Этимология и время появления слова

Здесь проблематична прежде всего вторая часть — время появления слова. Данных существующих корпусов недостаточно для достоверного определения этого параметра с точностью до года — можно получить лишь верхнюю оценку времени появления («Слово вошло в язык не позже … года»). И если для слов, вошедших в язык, например, в XIX веке, мы можем довольствоваться точностью в десятилетие, то для новейших слов, которые стали употребительны уже в XXI веке, такой оценки на нынешней временной дистанции недостаточно. Хорошим инструментом для оценки времени проникновения слова в книги является сервис Google Books, но большинство слов, входящих в словник СБТ, активнее употребляются в разговорной речи, чем в книжной, и поэтому важно время их появления на форумах, в блогах и т. п. Проблемы, связанные с текстами из этих источников, были рассмотрены выше.

Литература

1. *Апресян* и др. 2010 — Апресян В. Ю., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю., Галактионова И. В., Гловинская М. Я., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Птенцова А. В., Санников А. В., Урысон Е. В. Проспект активного словаря русского языка. Отв. ред. акад. Ю. Д. Апресян. М.: «Языки славянских культур», 2010. 784 с.
2. *Беликов, Ахметова* 2009 — Беликов В. И., Ахметова М. В. Статистическая оценка функциональных свойств лексики по материалам интернета // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009, с. 25–130.
3. *Беликов* 2011 — Беликов В. И. Чего не хватает в «оцифрованном мире» лексикографу и социолингвисту // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2011» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М.: РГГУ, 2011.

4. *Зализняк* 1985 — Зализняк А. А. От праславянской акцентуации к русской. М.: Наука, 1985.
5. *Зеленков* и др. 2005 — Зеленков Ю. Г., Сегалович И. В., Титов В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2005» (Звенигород). Вып. 4 (10). М.: Наука, 2005.
6. *Иомдин* 2009 — Иомдин Б. Л. Терминология быта. Поиски нормы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009, с. 127–135.
7. *Иомдин* 2011 — Иомдин Б. Л. Материалы к словарю-тезаурусу бытовой терминологии. СВИТЕР: образец словарной статьи // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. Отв. ред. И. М. Богуславский, Л. Л. Иомдин, Л. П. Крысин. М.: «Языки славянских культур», 2011, с. 392–406.
8. *Левенштейн* 1965 — Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР, 163, №4, 1965: 845–848.
9. *Левонтина* 2010 — Левонтина И. Б. Русский со словарем. М.: Азбуковник, 2010.
10. *Ляшевская, Шаров* 2009 — Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
11. *НОСС* 2004 — Новый объяснительный словарь синонимов русского языка. Изд. 2-е. Под общ. рук. Ю. Д. Апресяна. Москва; Вена: Языки славянской культуры: Венский славистический альманах, 2004.
12. *Орфографический* словарь 2005 — Русский орфографический словарь. Изд. 2-е, исправленное и дополненное. Отв. ред. докт. филол. наук В. В. Лопатин. М., 2005.
13. *Орфографический* словарь 2007 — Орфографический словарь. Под ред. В. В. Лопатина. М.: РАН, 2007.
14. *Русское* словесное ударение 2001 — Русское словесное ударение. Словарь нарицательных имен. Сост. М. В. Зарва. М.: ЭНАС, 2001.
15. *Словарь* «Языки городов». ABBYY Lingvo Клуб, http://www.lingvo.ru/goroda/articles.asp
16. *Сокирко, Толдова* 2005 — Сокирко А. В., Толдова С. Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика 2005. Автоматическая обработка веб-данных [Текст]: материалы о Школе. Ред. И. Сегалович, М. Маслов, Ю. Зеленков. М.: Яндекс, 2005.
17. *Суперанская* 1968 — Суперанская А. В. Ударение в заимствованных словах в современном русском языке. М.: Наука, 1968.

18. *Iomdin* et al. 2011 — Iomdin B., Piperski A., Russo M, Somin A. How different languages categorize everyday items. In: Computational linguistics and intellectual technologies. Papers from the annual international conference “Dialogue” (2011). Moscow: RGGU, 2011, p. 258–268.
19. *Kilgariff* 2007 — Kilgariff A. Googleology is Bad Science. In: Computational Linguistics, MIT Press, 33 (1), pp. 147–151.

References

1. *Apresian V. IU., Apresian IU. D., Babaeva E. E., Boguslavskaia O. IU., Galaktionova I. V., Glovinskaia M. IA., Iomdin B. L., Krylova T. V., Levontina I. B., Ptentsova A. V., Sannikov A. V., Uryson E. V.* (2010) Prospekt aktivnogo slovaria russkogo iazyka [Prospect of the Active dictionary of Russian]. Moscow.
2. *Belikov V. I., Akhmetova M. V.* (2009). WWW statistical estimations of the functional properties of lexical items [Statisticheskaia otsenka funktsionalnykh svojstv leksiki po materialam interneta]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2009” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2009”]. Moscow: RGGU, pp. 25–130.
3. *Belikov V. I.* (2011). What are sociolinguists and lexicographers lacking in a digitized world? [Chego ne khvatajet v “otsyfrovannom mire” leksikografu i sotsiolingvistu]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011”]. Moscow: RGGU, pp. 63–71.
4. *Zalizniak A. A.* (1985) Ot praslavianskoj aktsentuatsii k russkoj [The evolution of stress system from Proto-Slavic to Modern Russian]. Moscow.
5. *Zelenkov IU. G., Segalovich I. V., Titon V. A.* (2005). A probability model of resolving morphological ambiguity based on normalizing substitutions and positions of neighbouring words. [Veroiatnostnaia model sniatiia morfologicheskoj omonimii na osnove normalizuiushchikh podstanovok i pozitsij sosednikh slov]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2005” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2005”]. Zvenigorod.
6. *Iomdin B. L.* (2009). Everyday terminology. In pursuit of standards [Terminologiia byta. Poiski normy]. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2009” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2009”]. Moscow: RGGU, pp. 127–135.
7. *Iomdin B., Piperski A., Russo M, Somin A.* (2011). How different languages categorize everyday items. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” [Computational Linguistics

and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Moscow: RGGU, pp. 258–268.
8. *Iomdin B. L.* (2011). Materials for the thesaurus of Russian everyday life terminology. SWEATER: a sample dictionary entry [Materialy k slovariu-tezaurusu bytovoj terminologii. SVITER: obrazets slovarnoj stat'i]. Slovo i iazyk. Sbornik statej k vos'midesiatiletiiu akademika IU. D. Apresiana [The word and the language. A collection of papers to commemorate Academician Apresjan's 80th annivesrary]. Moscow, pp. 392–406.
9. *Kilgariff A.* (2007). Googleology is Bad Science. In: Computational Linguistics, MIT Press, 33 (1), pp. 147–151.
10. *Levenshtein V. I.* (1965) Binary codes capable of correcting deletions, insertions, and reversals [Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshchenij simvolov]. Doklady Akademii Nauk SSSR [Soviet Physics], No. 4, pp. 845–848.
11. *Levontina I. B.* (2010) Russkij so slovariom [Russian with a Dictionary]. Moscow.
12. *Liashevskaia O. N., Sharov S. A.* (2009) Chastotnyj slovar' sovremennogo russkogo iazyka (na materialakh Natsionalnogo korpusa russkogo iazyka) [Frequency dictionary of modern Russian based on the Russian National Corpus]. Moscow.
13. *Novyj* ob"asnitelnyj slovar' sinonimov russkogo iazyka [New explanatory dictionary of Russian synonyms] (2004). Moscow, Vienna.
14. *Orfograficheskij* slovar [Spelling dictionary] (2007). Moscow.
15. *Russkij* orfograficheskij slovar [Russian spelling dictionary] (2005). Moscow.
16. *Russkoe* slovesnoe udarenie. Slovar' naritsatelnykh imen [Russian lexical stress: A dictionary. Common nouns] (2001). Moscow.
17. *Slovar'* "Iazyki gorodov" [Dictionary "Languages of cities"]. ABBYY Lingvo Club, http://www.lingvo.ru/goroda/articles.asp
18. *Sokirko A. V., Toldova S. IU.* (2005). Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian [Sravnenie effektivnosti dvukh metodik sniatiia leksicheskoj i morfologicheskoj neodnoznachnosti dlia russkogo iazyka (skrytaia model Markova i sintaksicheskij analizator imennykh grupp)]. Internet-matematika 2005. Avtomaticheskaia obrabotka veb-dannykh [Internet mathematics 2005. Automatic processing of web data]. Moscow, Yandex.
19. *Superanskaia A. V.* (1968) Udarenie v zaimstvovannykh slovakh v sovremennom russkom iazyke [The stress patterns of loanwords in present-day Russian]. Moscow.