# База данных, содержащая информацию о частотности различных грамматических характеристик и окончаний у русских существительных[[1]](#footnote-1)

Наталия Слюсарь, НИУ ВШЭ, Москва, и СПбГУ, slioussar@gmail.com

Мария Самойлова, СПбГУ

## 1. Введение

Цель создания этой базы данных — получить важные сведения о грамматических характеристиках существительных русского языка, опираясь на подкорпус Национального корпуса русского языка (НКРЯ) со снятой неоднозначностью (www.ruscorpora.ru). Одной из задач было определить, насколько частотны формы существительных разного рода, в разных числах и падежах, одушевленных и неодушевленных, и как эти характеристики коррелируют друг с другом. Вторая задача заключалась в том, чтобы определить частотность форм с различными окончаниями (в зависимости от падежа, числа, рода и склонения и вне зависимости от них). Очевидно, что, если нужно сравнить, скажем, частотность двух падежей, несложно сделать запрос в НКРЯ, не пользуясь никакой базой. База нужна для того, чтобы получить общую картину (например, частотности интересующих падежей на фоне всех падежей), а также иметь возможность впоследствии включить в сравнение новые факторы, скажем, число или одушевленность.

Сведения такого рода необходимы для целого ряда теоретических и экспериментальных лингвистических исследований, в особенности для всего спектра подходов, ориентированных на употребление, т.н. *usage-based* (Baayen 2003; Bybee 2006; Dressler 1985; Milin et al. 2009 и мн. др.), а также для любых моделей, описывающих ментальный лексикон носителя: к какому бы направлению они ни относились, частотность всегда играет в них ту или иную важную роль. Причем сведения подобного рода могут быть востребованы как сами по себе (например, исследуя, как представлены в ментальном лексиконе грамматические категории рода, числа, падежа, важно знать частотность различных граммем), так и для решения вспомогательных задач, скажем, при подборе стимулов для психолингвистических экспериментов.

## 2. Описание базы

База данных создавалась следующим образом. В подкорпусе НКРЯ со снятой неоднозначностью были собраны сведения о частотности форм существительных с различными окончаниями. При этом учитывались такие параметры, как род, число, падеж, одушевленность и тип склонения, а также тип основы (подробнее об этом рассказано в разделе 2.2). Вся собранная информация и сделанные затем на ее основании расчеты помещались в таблицы в файл «freqdatabase.xlsx». В базу данных не вошли аббревиатуры, неизменяемые существительные и существительные адъективного склонения (их окончания совпадают с окончаниями прилагательных, поэтому становится сложно говорить о частотности окончаний). Некоторые другие исключения (всегда крайне немногочисленные) оговорены в разделе 2.3 и, если речь о совсем частных случаях, на листах с исходными данными.

### 2.1. Несколько слов о парадигмах русских существительных

Зализняк (1978) описывает склонение существительных так. Для каждого рода есть набор основных окончаний. По сути, для м. и ср.р. это парадигма 1 скл., для женского — парадигма 2 скл. [[2]](#footnote-2). Этот набор существует в двух вариантах: первые обычно используются для основ на твердые согласные, вторые — для основ на мягкие (см. Табл. 1-2). У слов с основами на ‑*г/к/х и ‑ш/ж/ч/щ/ц* (заднеязычные, шипящие и *ц*) окончания «твердого» и «мягкого» варианта смешаны, причем основы на ‑*ц* стоят здесь особняком, так как у них почти все окончания из «твердого» набора. Имеют свои особенности и основы на ‑*й*, но окончания у них из того же набора, что используется для других мягких согласных. Кроме того, есть слова м.р., относящиеся ко 2 скл.

Особняком стоят слова ж.р., оканчивающиеся на ‑*ь* (3 скл.), у которых свой набор окончаний (см. Табл. 3), а также т.н. разносклоняемые: слова ср.р., оканчивающиеся на ‑*мя*, и слово *путь*. В файле «freqdatabase.xlsx» они обозначены как *irreg.* Также есть слова адъективного склонения. Мы приняли решение их не учитывать.

Подавляющее большинство остальных особенностей склонения касается изменения основы и иногда нестандартного использования окончаний из основного набора, но не использования каких-то новых окончаний. Самые частотные исключения обозначены ниже (есть еще единичные слова-исключения типа *дитя* и пр.). Огромный пласт очень важной информации, которой мы никак не касаемся, — акцентные парадигмы. К сожалению, учесть их очень сложно.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | М.р. | Ср.р. | Ж.р. |  |
| «тв.»вар. | «мягк.»вар. | «тв.»вар. | «мягк.»вар. | «тв.»вар. | «мягк.»вар. |  |
| Им. |  | *0* | *0 (-ь/-й)* | *о* | *е* | *а* | *я* |  |
| Род. |  | *а* | *я* | *а* | *я* | *ы* | *и* |  |
| Дат. |  | *у* | *ю* | *у* | *ю* | *е* | *е* |  |
| Вин. | одуш. | *а* | *я* | *о* | *е* | *у* | *ю* |  |
|  | неодуш. | *0* | *0 (-ь/-й)* |  |
| Тв. |  | *ом* | *ем* | *ом* | *ем* | *ой* | *ей* | В ж.р. архаичныеварианты ‑*ою/ею*. |
| Предл. |  | *е* | *е* | *е* | *е* | *е* | *е* |  |

Табл. 1. Основной набор окончаний для слов м., ж. и ср.р. в ед.ч.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | М.р. | Ср.р. | Ж.р. | Особенности |
| «тв.»вар. | «мягк.»вар. | «тв.»вар. | «мягк.»вар. | «тв.»вар. | «мягк.»вар. |  |
| Им. |  | *ы* | *и* | *а* | *я* | *ы* | *и* | В м.р. ‑*а/я*, ‑*е[[3]](#footnote-3)*, в ср.р. ‑*и*. |
| Род. |  | *ов* | *ей* | *0* | *0 (-ь/-й)**// ей* | *0* | *0 (-ь/-й)**// ей*  | Много вариации, но доп. окончание одно: ‑*ев*. |
| Дат. |  | *ам* | *ям* | *ам* | *ям* | *ам* | *ям* |  |
| Вин. | одуш. | *ов* | *ей* | *0* | *0 (-ь/-й)**// ей* | *0* | *0 (-ь/-й)**// ей*  |  |
|  | неодуш. | *ы* | *и* | *а* | *я* | *ы* | *и* | В м.р. ‑*а/я*, в ср.р. ‑*и*. |
| Тв. |  | *ами* | *ями* | *ами* | *ями* | *ами* | *ями* |  |
| Предл. |  | *ах* | *ях* | *ах* | *ях* | *ах* | *ях* |  |

Табл. 2. Основной набор окончаний для слов м., ж. и ср.р. в мн.ч.

|  |  |  |
| --- | --- | --- |
|  | Ед.ч. | Мн.ч. |
| Им. | *0 (-ь)* | *и* |
| Род. | *и* | *ей* |
| Дат. | *и* | *ям/ам* |
| Вин. | *0 (-ь)* | *и* |
| Тв. | *-ью* | *ями/ами* |
| Предл. | *и* | *ях/ах* |

Табл. 3. Окончания 3 скл.

### 2.2. Как были собраны исходные данные

В идеале нас интересует частотность тех или иных грамматических характеристик и окончаний в рамках определенной парадигмы. Однако у русских существительных очень много различных особенностей в склонении, и учесть их все, т.е. посчитать всё по отдельности для всех возможных типов, представляется слишком сложным. Мы решили остановиться на полпути, сделав расчеты для основ на твердые, мягкие и шипящие и заднеязычные согласные.

Мы обратились к подкорпусу НКРЯ со снятой грамматической омонимией. Первичный сбор информации осуществлялся в октябре-декабре 2013 г. В выбранный нами подкорпус входят почти 6 миллионов слов из примерно 230 миллионов, т.е. очень много, хотя это и небольшой процент от всего корпуса.

Для сбора данных использовался лексико-грамматический поиск. В графе «грамматические признаки» мы выбирали, например: сущ., м.р., им.п., мн.ч., одуш. (последнее позволяет сравнить падежи у одушевленных и неодушевленных существительных, не говоря уже о морфологической роли этой категории), затем исключали неизменяемые слова и сокращения. Запрос в результате выглядел так: *S,nom,pl,m,anim ‑abbr ‑0.* В графе «слово» ставили, например: *("\*ки"|"\*ги"|"\*хи"|"\*ши"|"\*жи"|"\*чи"|"\*щи") ‑\*a,* т.е. формы на ‑*ки, ‑ги, ‑хи, ‑ши,- ‑жи, ‑чи, ‑щи*, но не от слов на ‑*а*. Последнее позволяет исключить слова м.р. 2 скл., например, *юноши* или *скряги*. В тех падежах, где у одних и тех же основ есть два варианта окончаний (скажем, *‑ой/-ей* у основ на ‑*ц* в зависимости от ударения), оба варианта были обсчитаны отдельно, чтобы можно было посмотреть, какой частотней.

На основании собранных таким образом данных были произведены все дальнейшие расчеты. Однако все исходные данные, а также информация о запросах включены в файл «freqdatabase.xlsx», чтобы в случае необходимости было проще произвести какие-то дополнительные вычисления или как-то иначе сгруппировать исходные данные (например, разделить основы на шипящие и заднеязычные). Заметим также, что некоторые запросы, в частности, приведенный выше, оказались для системы поиска в НКРЯ слишком длинными. В таких случаях она выдает не ошибку, а ответ «Сервис временно недоступен». Мы разбивали такие запросы надвое или натрое, а потом складывали полученные числа, но в файле «freqdatabase.xlsx» этого не указывали.

### 2.3. Некоторые особенности НКРЯ и того, как мы использовали представленную там информацию

Во-первых, в НКРЯ нет буквы *ё*. Во-вторых, в НКРЯ кроме мужского, женского и среднего выделяется общий род (около 5 тысяч примеров). К нему отнесены одушевленные существительные 2 скл. типа *убийца* (около 3,5 тысяч примеров) и некоторые другие менее интересные для нас слова вроде несклоняемых фамилий (около 1,5 тысяч). Мы его пока не учитывали. В-третьих, кроме основных шести падежей в НКРЯ выделяются еще несколько, перечисленные в Табл. 4.

|  |  |
| --- | --- |
| Звательный | Формы типа *мам, боже.* 659 существительных в ед.ч.: существительные м. и ж.р. 2 скл. (а также шесть имен типа *Паш*, спорно отнесенных к общему роду) и архаичные формы. Пока не учитывали. |
| Родительный 2 | Формы типа *(стакан) чаю.* Сущ. м.р. 1 скл. в ед.ч., а также почему-то форма *пол(у)ночи* (23 шт.).  |
| Винительный 2 | Формы типа (*идти в) солдаты*. 565 существительных м., ж. и общ.р. 1, 2, 3 скл. во мн.ч. Пока не учитывали. |
| Предложный 2 | Формы типа *(в) лесу, (в) сетú*. Существительные м.р. 1 скл. и ж.р. 3 скл. в ед.ч., а также почему-то аббревиатура *гг.* (21 шт.). |
| Счётная форма | Формы типа *(три) стола.* 676 существительных м.р. 1 скл. Пока не учитывали. |

Табл. 4. Падежи, выделяемые в НКРЯ в дополнение к основным шести.

Важно отметить, что в НКРЯ допущено некоторое количество ошибок при разметке материала. Это вносит определенную погрешность в полученные нами результаты. Однако, как мы покажем ниже на одном примере, погрешность эта очень небольшая (кроме того, следует учесть, что наши данные по определению являются приблизительными, так как получены на материале определенной корпусной выборки). Иногда, впрочем, ошибки начинают играть более существенную роль. Например, изначально мы нашли сколько-то форм, определенных как неодушевленные существительные м.р. на *‑а* и *‑я*. При ближайшем рассмотрении оказалось, что все эти случаи — результат различных ошибок, в основном формы типа *методами*, которые привязаны к двум леммам: *метод* (сущ. м.р.) и *метода* (сущ. на *‑а*). Мы надеемся, что нам удалось избавиться от всех проблем такого рода, но, конечно, не можем быть в этом уверенными.

### 2.4. «Контрольный замер»

Мы считаем важным, что все наши расчеты были произведены на одном и том же массиве данных — это позволяет без каких бы то ни было оговорок сравнивать их между собой. Тем не менее, мы решили провести нечто вроде «контрольного замера», посчитав частотность некоторых грамматических характеристик другим способом. В подкорпусе НКРЯ со снятой омонимией мы собрали сведения о количестве форм одушевленных и неодушевленных существительных разного рода и о количестве форм одушевленных и неодушевленных существительных в разных числах и падежах. В расчеты вошли все формы, классифицированные в НКРЯ как существительные, исключая аббревиатуры и неизменяемые, но включая адъективное склонение. Окончания, склонения и типы основ не учитывались.

Сумма всех форм, вошедших в наши основные расчеты, — 1 544 051, а в контрольные — 1 646 295 (данные по роду) и 1 647 107 (данные по числу и падежу). При этом, как можно удостовериться в файле «freqdatabase.xlsx», распределение форм по грамматическим категориям практически совпадает. Различия между выборками связаны с тем, что в основных расчетах мы давали намного более подробную характеристику форм, и в результате было исключено некоторое их количество (как ошибки, так и подходящие формы, у которых по ошибке не проставлены те или иные характеристики). Кроме того, как было сказано в разделе 2.3, мы исключили из поиска некоторые типы форм. В результате получается, например, что, когда мы делали дополнительные расчеты по числу и падежу, мы не брали звательный падеж, а в расчеты по роду эти формы вошли. Именно поэтому нам кажется важным, что все наши основные расчеты сделаны на одной выборке.

Единственная большая группа слов, которая не учтена в основной выборке, — это существительные адъективного склонения. Судя по предварительным запросам, всего их в интересующем нас подкорпусе НКРЯ около 38 000. Решение не включать их в расчеты изначально было принято потому, что трудно рассуждать о частотности присущих этому склонению окончаний в отрыве от частотности соответствующих окончаний у прилагательных, а как именно грамотно совместить эти данные, также не вполне очевидно.

### 2.5. Как устроен файл с базой данных

Листы с исходными данными в файле «freqdatabase.xlsx» помещены в конец и выделены серым. Листы с расчетами выделены голубым (там, где расчеты сделаны с учетом склонения, использован светло-голубой). Зеленым отмечен лист с контрольными расчетами (см. раздел 2.4).

Мы будем стремиться к тому, чтобы вся информация в файле приводилась по-русски и по-английски, окончания и прочее давались кириллицей и в транслитерации. Однако пока английский язык и транслитерацию можно найти только на листах с основными результатами. Там, где это возможно, мы используем общеупотребительные обозначения грамматических категорий латинского происхождения.

## Библиография

Зализняк А.А. Грамматический словарь русского языка: Словоизменение. М., 1978.

Русская грамматика. Под ред. Н.Ю. Шведовой. М., 1980.

Baayen, R. 2003. Probabilistic approaches to morphology. In: *Probability theory in linguistics.* Cambridge, MA: MIT Press, 229–287.

Bybee, J. 2006. *Frequency of use and the organization of language.* Oxford: OUP.

Dressler, W.U. 1985. *Morphonology*. Ann Arbor: Karoma Press.

Milin, P., Filipovic Durdjevic, D., Moscoso del Prado Martín, F. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: evidence from Serbian. *Journal of Memory and Language, 60*, 50–64.

1. Работа была частично поддержана грантом РГНФ № 14-04-12034. [↑](#footnote-ref-1)
2. Нумерация склонений дается по «Русской грамматике» (1980), т.е. слово 1 скл. — это, например, *стол*. [↑](#footnote-ref-2)
3. Только у одушевленных существительных типа *крестьянин-крестьяне*. [↑](#footnote-ref-3)